Lip Based Recognition: A Comparative Analysis

[1]M Haider Mehraj, [2]Ajaz Hussain Mir
[1]Dept. of ECE, COET, BGSB University, Rajouri, J&K, India
[2]Dept. of ECE, NIT Srinagar, J&K, India

Abstract
There is an increasing requirement for robust and reliable person authentication systems in areas of high security or secure access. The majority of current techniques for person recognition focus on either static facial information or speaker recognition via the speech signal. Whilst in clean conditions, the speech signal has proved to be a valuable source of speaker dependent information, problems occur in noisy or channel mismatch conditions. Whilst lip information presents predominantly speech dependent information, valuable speaker dependent information is also contained within the static and dynamic features of the lips. This paper presents a comparative review of various lip based biometric techniques and makes comparisons between them by using Relative Operating Characteristics (ROC) curve and Recognition Time (RT) as performance metric. The ROC curve and recognition time has been obtained at varying frame rates to determine the best lip recognition technique.

Keywords
Lip Biometric; Recognition; ROC; AUROC; RT; Correlation; Motion

I. Introduction

Biometrics parameters like face, lips, iris and other various parameters are being used for individual identification and verification, recent works in field has enabled similar recognition for various states or expressions automatically. A major technical advancement in past few years has propelled face recognition technology into spotlight [1]. Face recognition is natural and passive and hence has clear advantage over other biometrics like fingerprints, iris, palm print and hand geometry.

Lip is one of the most important benchmark and advance parameter used for individual recognition of its varying postures [2]. Lips can be represented as an image of size p x q pixels and further can be represented by a vector in p.q dimension space. In practical applications lip as parameters for identifying individuals outperform other biometrics [3]. Images may be acquired from the distance without the knowledge of examined person. Lip biometrics is anatomical i.e. better results are expected than other behavioural biometrics.

The use of visual information specifically a person's mouth region as a feature to recognition system has been reported. Wark and Sridharan investigated the use of shape and intensity information from a person's lips in speaker identification systems [4]. M.I Faraj et al.[5], developed a robust audio visual digit and speaker-recognition system using lip motion and speech biometrics. Moreover, a liveness verification barrier based on a person's lip movement was added to the system to guard against advanced spoofing attempts such as replayed videos. S A Dzati et al.[6], proposed a method based on lip motion sequence and the Unconstrained Minimum Average Correlation Energy (UMACE) filter as a classifier for person identification. H.ECetingul et al.[7], addressed the selection of robust lip-motion features for audio-visual open-set speaker identification problem. Lip based recognition can be classified along three different techniques: motion based, shape based and pixel or correlation based. In this paper an overview and comparison of these three different techniques is provided. Comparisons are made on basis of Relative Operating Characteristics (ROC) and Recognition Time (RT) at varying frame rates in training set.

II. Lip Based Recognition Techniques

A. Motion Based

It can be further divided in two scenarios[8]. In the first one, the rectangular mouth region is first to be detected and then the mouth movement is represented by the motion vectors computed on a predefined rectangular grid within this region. In the second scenario, the lip boundary has to be tracked over time and only motion of lip boundary pixels are taken into account. Whatever the scenario used the various steps involved are lip contour tracking, motion estimation and motion representation.

There are a number of approaches such as splines, active contours, and parametric models in the literature in order to represent and extract the lip contour. Classical active contours and splines suffer from complex parameter tuning [8] and they are mostly unable to perfectly fit to the characteristic lip parts such as Cupidon's bow because of the erroneous gradient information due to illumination differences. Eveno et al. [9] proposed to fit cubic polynomials on the outer lip contour using the color information of the lip image. In this technique a pre-processing stage to find predefined six key points on the lip is followed by an optimization stage in which four cubic polynomials and two lines are fitted to the outer lip contour. The pre-processing stage includes computing the pseudo hue component h(x, y) at pixel (x, y) for lip segmentation. The technique proposed in [9] works well only under some assumptions on the acquisition environment and illumination conditions. However in many practical conditions, these assumptions do not hold. The lack of discriminative color information, especially on the lower lip boundary, becomes occasionally so severe that even a human eye can hardly make a distinction. H.E Centigual [8] proposed to use a quasi-automatic strategy that needs user interaction. In cases where the algorithm fails, the tracking task is assisted with some hand labeled points on the lip boundary.

The first scenario uses a uniform grid of size N×M on the intensity lip image. This grid definition allows one to analyze the information content of both the lip area and the non-lip area inside the rectangular mouth region. The motion vectors computed on this grid are computed using block matching approach. The idea behind block matching is to divide the current frame into a matrix of 'macro blocks' that are then compared with corresponding block and its adjacent neighbors in the previous frame to create a vector that stipulates the movement of a macro block from one location to another in the previous frame. This movement calculated for all the macro blocks comprising a frame, constitutes the motion estimated in the current frame. The search area for a good macro block match is constrained up to p pixels on all fours sides of the corresponding macro block in previous frame. This 'p' is called as the search parameter. Larger motions require a larger p, and the larger the search parameter the more computationally expensive the process of motion estimation becomes. Usually the macro block is taken as a square of side 16 pixels, and the search

parameter p is 7 pixels. The idea is represented in fig. 1.

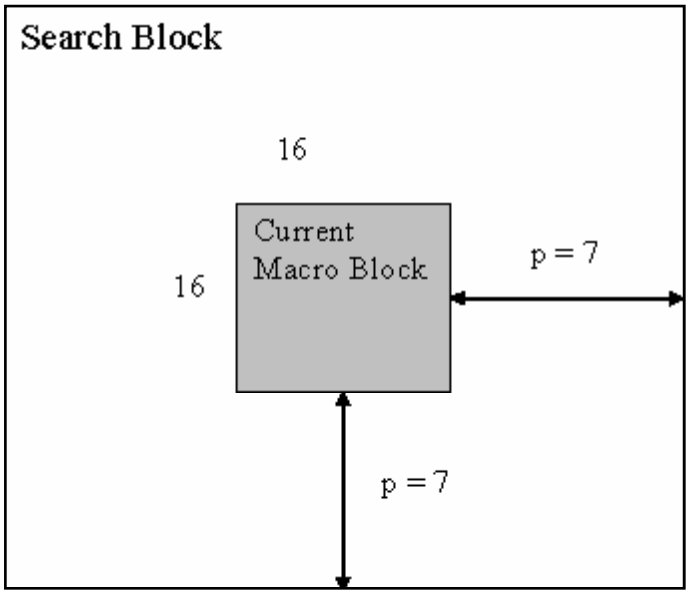


Fig. 1: Block Matching a Macro Block of Size 16 Pixels and a Search Parameter p of Size 7 Pixels

The matching of one macro block with another is based on the output of a cost function. The macro block that results in the least cost is the one that matches the closest to current block. There are various cost functions, of which the most popular and less computationally expensive is Mean Absolute Difference (MAD) [10] given by equation below:

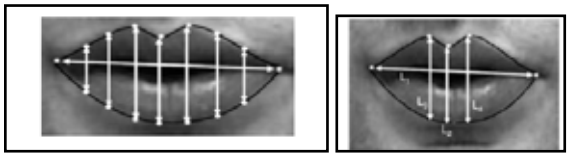
$$MAD = \frac{1}{N^2}\sum_{i=0}^{N-1}\sum_{j=0}^{N-1}|C_{ij} - R_{ij}| \quad (1)$$

Following the motion estimation, the x and y components of the motion vectors are separately transformed via 2-DCT [8]. The DCT is similar to the discrete Fourier transform, it transforms a signal or image from the spatial domain to the frequency domain. The Discrete Cosine Transform (DCT) represents an image as a sum of sinusoids of varying magnitudes and frequencies. The DCT has the property that, for a typical image, most of the visually significant information about the image is concentrated in just a few coefficients of the DCT. The extracted parametric lip contour is in fact a rough sketch of the real lip and does not contain sufficiently detailed information to characterize discriminative biometrics of different speakers. Discriminative information can be provided by incorporating the motion vectors computed along the parametric lip contour. Thus in the second scenario, only the motion vectors estimated on the pixels of the extracted lip contour are taken into account and the rest is discarded. Since there is no 2D grid, the two sequences of x and y motion components on the contour pixels are separately transformed using 1D-DCT.

If the DCT coefficients obtained for training images match with those with those obtained for the testing images, then person is positively identified. This is done by using HMM toolkit in which log likelihood ratios are obtained between world HMM (which is trained over entire database) and HMM trained over the frames corresponding to the person to be identified.

B. Shape Based

In this approach geometric features such as distances between different lip points are taken into consideration. Generally eight parameters [11] are taken where each parameter is the distance between key points on lip contour representing the horizontal and the vertical opening of lip. Lip contour and key points need to be extracted and is done from lip contour tracking approach. One can reduce the number of parameters to four by taking only three vertical distances from three points in cupidon's bow to lower lip and one maximum horizontal distance. Further reducing the parameters will result in erroneous results as discarding any of three vertical distances will then in reduced feature vector set will not cover entire lip region distances.



(a) (b)

Fig. 2:(a) Eight Lip Shape Parameters and (b). Four Lip Shape Parameters [11].

C. Pixel Based

In pixel based group, the entire region containing the speaker's mouth (region of interest – ROI) is considered as informative for lip reading [12]. These usually work on the principle of differentiating skin color from non-skin color i.e. a face detection algorithm is used. The various steps involved [6,13] are face detection and lip localization. For the purpose of face detection color based and template matching technique is used. Summarily, the Gaussian model N(μ,C), with mean vector μ = E[x] and covariance matrix $C=[E(x-\mu)(x-\mu)^T]$ which are obtained from different skin color images are used to find out the skin likelihood for any pixel of an image. The skin likelihood is calculated as

$$p(r,b) = \exp[-0.5(x-m)^T C^{-1}(x-m)] \quad (2)$$

where $x=(r,b)^T$ i.e. chromatic pair value of red and blue.

The skin-likelihood image is transformed to the skin-segmented image (binary image). Since the skin regions are brighter than the other parts of the images, the skin regions can be segmented from the rest of the image through a thresholding process. To process different images of different people with different skin, a fixed threshold value is not possible to be found. Since people with different skins have different likelihood, an adaptive thresholding process is required to achieve the optimal threshold value. The threshold value at which the minimum increase in region size is observed is the optimal threshold After optimal threshold has been set, all pixel values which have likelihood values higher than threshold are set to 1 and the rest of pixels are set to 0. Thus, resulting in binary image. Finally, the face region is then detected as depicted in fig. 3.



Fig. 3: Face Detection Process: Original Image, Skin-Likelihood Image, Skin-Segmented Image, Detected Face Region [6]

For the lip localization task, hue/saturation color thresholding is implemented to differentiate the lips from the skin. The detection of the lip in hue/saturation color is much easier owing to its robustness under wide range of lip colors and varying illumination conditions [14]. From the hue-saturation image, a binary image is then obtained by setting the threshold values, i.e. setting some specific threshold for hue and saturation. Lastly by employing

morphological image processing, the lip region can be localized by finding the largest blob in the binary image.

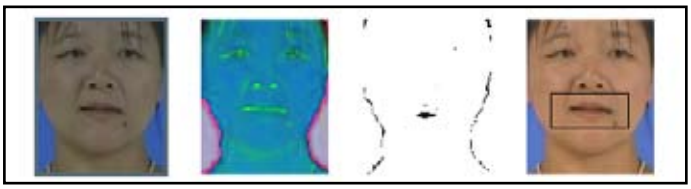


Fig. 4: Lip Localization Process: Face Detection Region, Hue-Saturation Image, Binary Image, Localized Lip Region [6].

Finally features of interest are extracted using correlation filter which performs two fold task of obtaining features of interest and classification. Since correlation filter is used, it is also known as correlation filter or simply correlation approach.

III. Performance Metric

In a Biometric system for person recognition, the possible outputs are either positive, p (verified as the person in the system database) or negative, n (identified as someone who is not in the database). If the output is p and this person is the person in the database, the result is a True Positive (TP); however, if this person is not really the person in the database, the outcome is a False Positive (FP). Conversely, a True Negative (TN) results when the output and the actual identity are n, and a False Negative (TF) when the result is n but the actual identity is p.

The detection power (also called the true positive rate) is the fraction of all positives that are correctly classified as positive. The False Rejection Rate (FRR) is the fraction of all positives that are incorrectly classified negative. The false accept rate (FAR) is the proportion of all negatives that are incorrectly classified positive. The ROC is obtained by plotting True Positive rate against False Accept rate or against false reject rate. It can be also be obtained by plotting FAR against FRR. However it is not preferred because the Area under ROC (AUC) which determines the accuracy can effectively be determined by plot of True Positive Rate against FAR. The AUC is equal to the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one For successful recognition AUC should be greater than 50% and for good recognition it should be between (80-90)%. ROC can also be used determine the optimum threshold at which the system should operate by taking a point on true positive rate axis corresponding to ideal classifier. Another Parameter that we have used is recognition time i.e. the time taken for the completion of recognition process. It may be important in certain application where time is an important factor such as access to schools by means of the biometrics.

IV. Database and Experimental Setup

The audio-visual Digit database developed by Sanderson titled VIDTIMIT was used for the experiments described in [15]. The database consists of video and corresponding audio reciting digits zero to nine. The video of each person is stored as a sequence of JPEG images with a resolution of 512 x 384 pixels while the corresponding audio provided as a monophonic, 16 bit, 32 kHz WAV file.

In our implementation we have taken fourteen people from the VIDTIMIT database. We have used the two sequences per person each consisting of 20 frames for the purpose of testing the lip based system. For training purposes we have taken 5 and 10 frames per sequence and tried to choose only those frames that have maximum illumination changes and distortions. In the experimental setup, for pixel based approach skin likelihood values in range of "0 to 1" and a threshold of 0.4 for these values gives us a segmented image and followed by obtaining face region from segmented image by using 0.8 height to width ratio as well as taking only that region having at least one hole in it. Lips are localised by taking HSV image of RGB image and setting hue= 0.04 and saturation=0.1 to obtain binary image and then take largest blob in binary image as lip.

In the motion based approach first an initial lip region of size 128* 80 is segmented from each video frame, following the registration of successive face regions by global motion compensation. For grid-based motion analysis, a rectangular grid of size Gx*Gy = 64*40, or 32*20 or 16*10 is considered for each lip segment. Following motion estimation and 2D-DCT, a feature vector of size "70*13 or 140*13 or 210*13 or 280*13" is obtained by eliminating some high indexed DCT coefficients. For contour-based motion analysis, a similar procedure is used. First, the lip contour is extracted in each frame. Following motion estimation and 1D-DCT on the lip contour pixel locations, a feature vector of size 6*2 is obtained. Identification is carried out using HMM structures trained for each person as well as entire database. The HMM structures are implemented using the HTK tool version 3.3.

The third candidate for the lip feature representation is obtained by using four lip shape parameters that is one horizontal distance and three vertical distances. The identification is carried out by simple matching of lip shape parameters.

V. Results

We have obtained the ROC Curves for 5 and 10 frames per sequence per person and made the evaluations based on those plots.

The ROC curve obtained for the 5 frame per sequence case is shown in fig. 5.

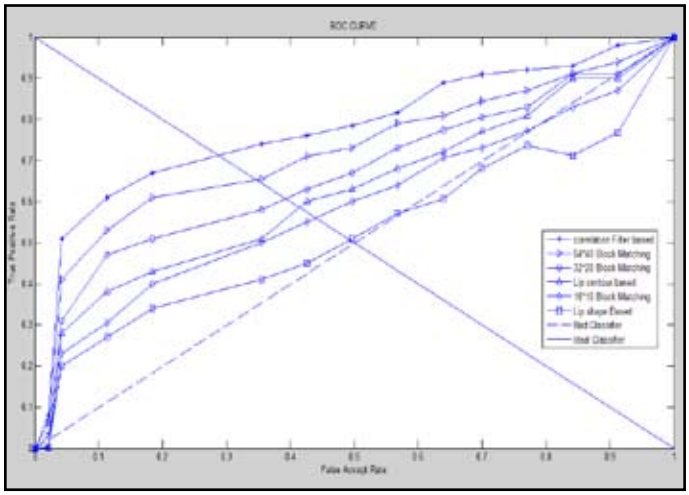


Fig. 5: ROC Curve at 5 Frames Per Sequence for Various Approaches

Although the 64*40 blocking matching approach is best using 5 frame per sequence but the AUROC in that case is 54% which is just marginally above the one required for the successful test. In case of correlation(pixel based), 32*20 block matching, Lip contour, Lip shape and 16*10 Block matching the AUROC's obtained are 51%, 41%, 38%, 36.3% and 32% respectively. Thus last four cannot be used at all for the recognition purposes. Now if we select the cutoff point i.e. the point at which system will operate as .59 on true positive rate axis. This is chosen because at this point 64*40 block matching curve intersects the ideal classifier.

Now the system using 64*40 block matching, which is best at this frame rate will have recognition rate of around 60% but FAR will be around 40% which is too high.
The ROC plot for 10 frames per sequence is shown in fig. 6.

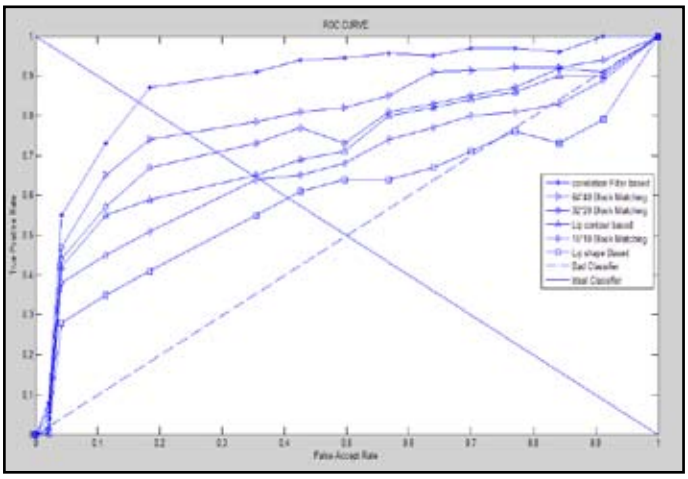


Fig. 6: ROC Curve at 10 Frames Per Sequence for Various Approaches

The correlation filter based approach is best with AUROC of 81 while in case of 64*40 block matching, 32*20, lip contour, 16*10 block matching and lip shape based, it is 71%,63%, 56% and 47% respectively. Only lip shape based approach fails.In case of correlation(pixel based), the feature vectors obtained are quite discriminative resulting in better recognition rate. In case of 64*40 Block matching approach the motion vectors between different blocks tend to be distinct from one another as compared to 32*20 and 16*10 block matching approaches. In case of lip shape based approaches the various distances obtained between cupidon's bow and lower lip and one horizontal distance are not quite distinct from one person to another. However AUROC has increased from 36.3% to 47% meaning as the number of frames increase we have more number of distances to work with and hence get better results.
The time in seconds is plotted against person number as shown in fig. 7 and 8 for 5 frames and 10 frames per sequence respectively

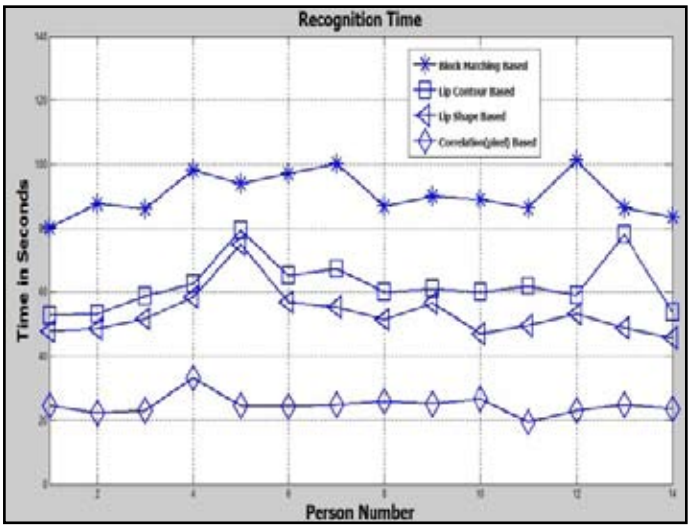


Fig. 7: Recognition Time for 5 Frames Per Sequence in Training Set for Fourteen People

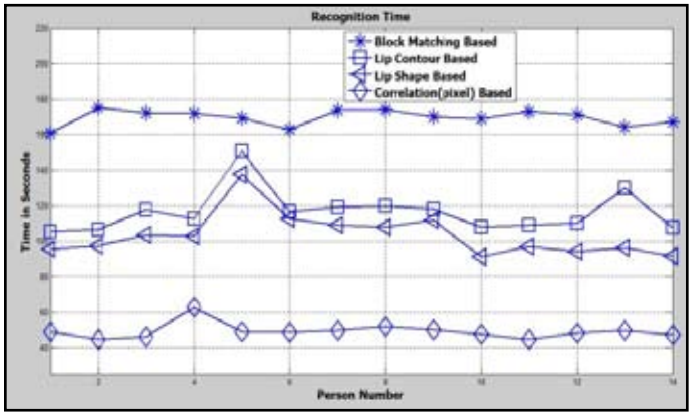


Fig. 8: Recognition Time for 10 Frames Per Sequence in Training Set for Fourteen People

The Recognition time for block based approach is highest as the computation of the motion vectors is a complex process and takes about half of time followed by lip contour based approach which takes less time owing to fact that motion vectors are computed only for lip contour. Lip shape takes almost similar time as that of lip contour as contour tracking step is same for both cases. Correlation (pixel) based takes least time as the steps involved are computationally simple. The recognition time almost doubles when number of frames per sequence is increased from 5 to 10.

VI. Conclusion

This work discussed the various lip based recognition approaches and compared their performance on the basis of relative operating characteristics and recognition time. The correlation (pixel based) approach performed best at all the four different frame rate considered because of its ability to trade-off between discrimination and distortion e.g. variations in pose, illumination and facial expressions. The Lip motion on a rectangular grid of size 64*40 approach performed reasonably well as compared to other grid sizes as each block had only those pixels which corresponded to only one apparent motion. The 32*20 and 16*10 based performed just above average because of inability to have pixel in each block having single apparent motion. The lip shape based approach failed owing to very small difference in mouth opening and closings of each person.
The recognition time required was least in case of the correlation (pixel) approach owing to less computations required as compared to other cases.

References

[1] Szewczyk, R.,"New Features Extraction Method for People Recognition on the Basis of the Iris Pattern", IEEE International Conference on Mixed Design of Integrated Circuits and Systems 2007 (MIXDES-2007), 21-23 June. pp. 645-650, 2007.
[2] Mok, L.L. Lau, W.H. Leung, S.H. Wang, S.L.Yan, H. "Person authentication using ASM based lip shape and intensity information", IEEE International Conference on Image Processing 2004 (ICIP-2004), 24-27 Oct. pp. 561- 564.
[3] Lievin, M. Luthon,"A hierarchical segmentation algorithm for face analysis: Application to lipreading", IEEE International Conference on Multimedia and Expo 2000 (ICME-2000), 30 Jul -02 Aug. pp. 1085-1088.
[4] T Wark, S Sridharan,"A Syntactic Approach to Automatic Lip Feature Extraction for Speaker Identification", Proceedings IEEE International Conference on Acoustics Speech and

Signal Processing(1998), 12th-15th May, pp. 3693–3696.
[5] Maycel-Isaac Faraj, Josef Bigun,"Synergy of Lip-Motion and Acoustic Features in Biometric Speech and Speaker Recognition", IEEE Transactions on Computers, Vol. 56, No.9, 2007, pp. 1169-1175.
[6] Salina Abdul Samad, Dzati Athiar Ramli, Aini Hussain. "Knowledge-Based Intelligent Information and Engineering Systems: KES 2007-WIRN 2007", Springer, Berlin / Heidelberg, 2007.
[7] H.E.Cetingul, Y.Yemez, E.Erzin, A.M.Tekalp,"Robust Lip-Motion Features for Speaker Identification", Proceedings IEEE International Conference on Acoustics, Speech and Signal Processing 2005(ICASSP-2005), 18th -23rd March, Philadelphia PA, USA, pp. 509-512.
[8] H.E. Cetingul, Y. Yemez, E. Erzin, A.M. Tekalp. "Discriminative Lip-Motion Features for Biometric Speaker Identification", IEEE International Conference on Image Processing 2004 (ICIP-2004), 24th-27th Oct, Singapore, pp. 2023-2026.
[9] N. Eveno, A. Caplier, P. Y. Coulon,"A Parametric Model for Realistic Lip Segmentation", Proceedings International Conference on Control, Automation, Robotics and Vision 2002 (ICARV-2002), 2nd-5th Dec. pp. 1426-1431.
[10] Aroh Barjatya,"Block Matching Algorithms For Motion Estimation", DIP 6620 Spring 2004 Final Project Paper, Utah state University [Online] Available: http://www.cc.usu.edu/~arohb .
[11] M. E. Sargin, E. Erzin, Y. Yemez, A. M. Tekalp,"Lip feature extraction based on audio-visual correlation", Proceedings European Signal Processing, 2005.
[12] Potamianos, Gerasimos, Chalapathy Neti,"Improved ROI and within frame discriminant features for lipreading", IEEE International Conference on.Image Processing, 2004 (ICIP-2004), 7th –10th Oct. pp. 250-253.
[13] M.H. Mehraj, A.H.Mir.,"Performance Evaluation of Correlation Filter Based Lip Biometric", International Journal of Engineering and Research Technology (IJERT), Vol. 1, No. 8, 2012.
[14] Matthews I, Cootes J, Bangham J, Cox S, Harvey R. "Extraction of visual features for Lip reading", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 24, No. 2, 2002, pp. 198–213.
[15] Sanderson C., Paliwal K.K.,"Noise Compensation in a Multi-Modal Verification System", Proceedings. International Conference on Acoustics, Speech and Signal Processing 2001 (ICASSP-2001), 07th-11th May 2001, pp. 157–160.



M Haider Mehraj received the B.TECH. degree in Electronics and Communication from the Guru Nanak Dev University, Amritsar, India in 2009, the M.Tech. degree in Communication and Information Technology from the National Institute of Technology, Srinagar, India in 2011. He joined Baba Ghulam Shah Badshah University, Rajouri, India as Assistant Professor in September 2011. His research interests are in the field of Digital Image Processing, Biometrics and Computer and Network Security.