Silent Sound Technology for Mandarin

Pradeep B.S*
Technology Director, R&D department,
Linyi Top Network Co. Ltd., China
Email:pradeepbs78@yahoo.com

Zhang Jingang
CEO, Linyi Top Network Co. Ltd. China
Email:zjg@iamtop.net

Abstract- Silent Sound Technology for mandarin is a non-speech interaction embedded engine with a variety of lip movement and facial expression interactive service platform based on cloud computing. This technology helps people to communicate in noisy places and to reduce noise pollution to some extent. If we don't communicate..!! As a result of being unable to communicate is a withdrawal from many of the social activities. An attempt is made to develop a mobile terminal oriented integration of a variety of interactive way emotional speech interaction system just to put an end for embracing situation like when a person need to convey that 'I can't talk to you right now' or apologetically rushing out of the place along with the device to answer and so on.. A high resolution camera is used to capture the video of human silent speech and is morphologically segmented for noise. From the binary image frames the lip montage is built to generate a threshold value for every word in sequence and a database is created for all known templates. After the threshold test pass, other facial parts like eyes with eye brows and nose are mapped with the data base contents to obtain emotional expression and finally, user can receive text messages in sequence as well as audio output. Initial results obtained are discussed in this paper.

Keywords- Frame montage, Morphology, Silent sound, Threshold test, Video and image processing.

I. INTRODUCTION

SST is a technology for devices that helps for communication purpose in the nasty environment. The uses of this technology are immense for people who are vocally challenged or have been rendered mute due to some accidents or others. Lip detection is a complex problem because of high variability range of lip shapes & colour [1]. Lip-reading is an inference and inspired guesswork because of fast speech, poor pronunciation, bad lighting, faces turning away, hands over mouths, moustaches and beards etc. Lip Tracking is one of the biometric systems based on which a genuine system can be developed. With multiple levels of video processing, it's possible to obtain lip contour and location of key points in the subsequent frames is usually referred to as lip tracking [14]. A large category of techniques referred to as model-based, build a model of the lips and its configurations are described by a set of model parameters [1]. Most of these techniques include tracking of the lip in sound speech may be with different accent & other facial parts consideration. Our effort is to work on silent speech which means no sound is incurred; a device oriented package to design and implement for the purpose of lip reading that can recognize mandarin words, single sentence or even continuous sentences of the people of different regions in China country considering their non-speech accent and pronunciation by observing their every movement of the lip and facial expression.

A. *Objectives*

Aim of this research work is to analyze and understand -every movement of the lips and facial expressions then transform them into text and audio output: Capturing the video using an integrated camera and process it based on histogram equalization for gray scale mode selection or normal mode. Skin segmentation and morphological operation helps to locate facial features in the interior of the face and colour coded perimeter with fitting points on the contour of the lip. Obtained multi image montage of lip will be converted to an average threshold value that helps to set the matching parameters to a very close value of other known templates in the database so as to test the feedback provided by the system will be a match or a miss-match to obtain text and sound output.

B. *Scope*

Scope of this technology is immense for communication purposes especially it;

- Helps people who migrate from one place to another place where they cannot communicate with the public in their regional/local language due to language disorders.
- Helps to Analyse and understand the people who have lost voice to speak or stuttering problem or patients on the bed to convey their needs to others to understand.
- Helps people to make silent calls or send busy tone to callers during conversation/meetings/ standing in the mass/crowded place without bothering others.
- User can tell PIN no., credit card no., password and other personals without bothering some eavesdrops. Customization of video stream history will be facilitated with Help.
- Software can be installed in wrist watch, wrist tag or display/Mobile/Pc and etc. The frustration that comes with misunderstanding or losing the thread of what is going on can be avoided and many more.

II. PROCESS MODEL AND WORKING METHODOLOGY

To proceed with this research work, the Process Model assumed is Iterative Process Model since it is more adaptable for this work. Once the face detection and mouth region detection is achieved, speech analysis can be performed with the use of lip motion features strategies and emotional expression with the use of other facial parts. If efficiency with identification technique is not proper then the threshold value falls out of the defined unique index value and retrial has to be made. Those are one of the main reasons to choose the Iterative Process [11]. Fig.1 shows the overall architecture of the process model and its working methodology.

As the live video is captured by a high resolution camera, the video can be processed as normal or grayscale colour mode /or saved as Mpeg, Avi, Flv etc. for customization. Region of Interest (ROI) video is segmented from which Facial features like Mouth, nose & Eyes are detected.

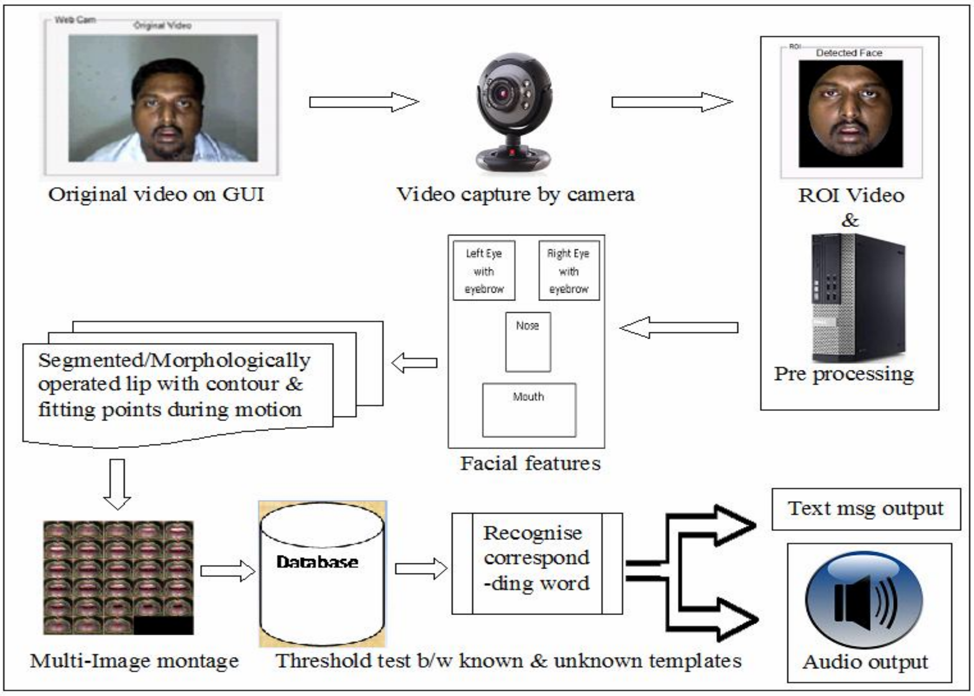


Fig. 1 Process model architecture and its working methodology

As the lip contour is initiated accurately, the extracted lip contour is morphologically processed and corners are fitted by the key points like left/right and upper/lower corners with centroid. Obtaining the lip contour in subsequent frames and the lip movement is completely tracked. A multi frame montage in a single object image montage is built and a database is created for it and obtained feature like eyes and nose vector in the database. The unknown templates of a user are then compared with the existing templates in the database. If the frames pass the threshold test with the known & unknown templates, based on trained and tested index value/trials the user can receive a text output & later an audio output [12].

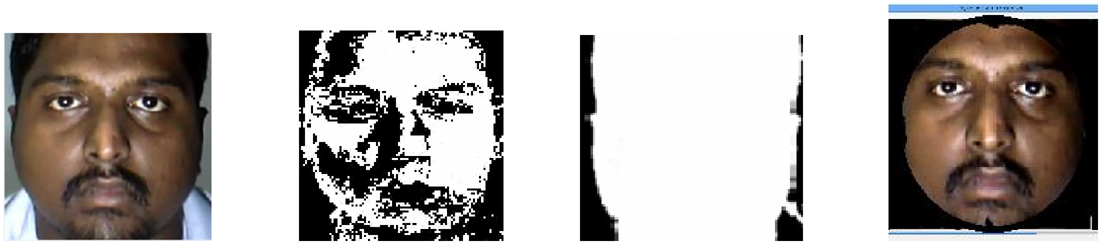
III. SKIN SEGMENTATION AND MORPHOLOGY

One the important step in face feature extraction process is skin segmentation. As we know human face varies from person to person, so the race, gender, age and other physical characteristics of an individual have to be considered. The partition is based on the color difference between the skin and non-skin regions [1]. The efficiency of the colour segmentation of a human face depends on the colour space that is selected. We used the finding by Yang and Waibel who noted that the skin colours of different people are very closely grouped in the normalized R-G colour plane. So some of the seed pixels are taken from the face area obtained the normalized R-G spread and classified the pixels that were close enough to this spread as belonging to the face. [9, 11].

In this phase we detect face region from input video, extracting it into frames either in grayscale or colour mode.
To extract face region we perform lighting compensation on image, then extract skin region and remove all the noisy data from image region. Finding skin colour blocks from the image and then check face criterions of the image. In lighting compensation we normalize the intensity of the image, when extracting skin region then apply threshold for the chrominance and then select the pixels that are satisfying the threshold to find the colour blocks.[1,13] The skin colour blocks are identified based on the measure properties of image regions in image. Height and width ratio is computed and minimal face dimension constraint is implemented. Crop the current region, existence and localization of face then compute vertical histogram [2, 10]. As a result of skin segmentation the interior holes that are created on the face region are morphologically processed.

Morphological image processing is a collection of techniques for digital image processing based on mathematical morphology. Since these techniques rely only on the relative ordering of pixel values, not on their numerical values, they are especially suited for processing of binary images and grayscale images based on two functions [2]:

Erosion- is an operation that thins or shrinks the objects in the binary image. Erosion are performed by IPT function imerode().
Dilation- is an operation that grows or thickens the objects in the binary image. Dilation is performed by IPT function imdialate().
For successful facial feature extraction the accurate computation of the contour points is very important. This helps in locating searching regions. Both neck and ears have the same colour as that of the face. Hence they are connected to the face region. Therefore we need to separate them so as to better locate the facial features. The Fig. 2 (a,b,c,d) demonstrates our results.



a. Original image b. Skin segmented image c. Binary image d. Active shape of face

Fig. 2 Morphological image and face

The face detection from the captured video is based on the Active Shape Model (ASMs) [12]. The shape model is learned from a set of manually annotated shapes of faces (unobstructed frontal views) as follows: the first step is to align all shapes of the learning data to an arbitrary reference by a geometric transformation (rotation, translation and scaling). The second step is to calculate the average shape. These two steps are repeated until convergence by minimizing the average Euclidean distance between shape points. At the end of the process, the facial shape model is obtained by active contour analysis (ACA) of the average of the aligned shapes using computer vision toolbox.

```
faceDetector=vision.CascadeObjectDetector('FrontalFaceLBP');
videoFileReader = vision.VideoFileReader('path');
bwface = activecontour(grayface,mask,40,'Chan-Vese');
bw = activecontour(A,mask,method)
```

The above command helps to detect the face and specifies the active contour method used for segmentation, either 'Chan-Vese' or ‘edge'. Thus, we obtain the principal modes of shape variations.

To extract the facial features from a new image, first, face is detected using the MAT lab command based on Viola & Jones face detection algorithm [6, 7]. After that, the shape model is positioned on the face, and iteratively deformed until it sticks to the face of the image in the respective bounding boxes. The facial features can be located with high reliability in standard lighting conditions, frontal face position and classical facial expression [4]. Both the methods are feasible.

A. Contour Fitting Points Location

The points of interest also referred to as key points or contour fitting points widely used for lip reading and other applications like relationship with a particular structure. A corner can be defined as points for which there are two dominant and different edge directions in a local neighbourhood of the point. We know that lip detection is a complex problem because of the high variability range of lip shapes and colour.[1] So the main advantages of a corner detector is its ability to detect the same corner in multiple similar images, under conditions of different lighting, translation, rotation and other transforms. We detect the centroid point of the bounding box then centroid Column and centroid Row using the Mat lab command:

```
centroidColumn = int32(centroid(1)); % "X" value
centroidRow = int32(centroid(2)); "Y" value.
middleRow = binaryImage(centroidRow, :);
middleColumn = binaryImage(:, centroidColumn);
centroidColumn, centroidRow – centroid point
```

The upper and lower key points are found as the intersection points between the minor axis line and the upper and lower lip boundary, respectively.

```
topRowY = find(middleColumn, 1, 'first');
centroidColumn, topRowY       -this gives top
```

bottomRowY = find(middleColumn, 1, 'last');
centroidColumn, bottomRowY -this gives bottom

Unlike the upper and lower keypoints, which are precisely detected from the color segmented lip object, the mouth corners (left and right keypoints) are more difficult to detect because of their location in dark areas, where chromatic information is not visible. In order to detect them, we use the extreme left and right points of the lip object as starting points and search for corners in the proximity area using the below MAT lab command:

leftColumnX = find(middleRow, 1, 'first');
leftColumnX, centroidRow -this gives left
rightColumnX= find(middleRow, 1, 'last');
rightColumnX, centroidRow -this gives right

The proximity areas, where corners are searched out are explained above. From the corners detected in each side, we choose the one with the smallest Euclidean distance from the corresponding extreme object points as the left and right keypoints, respectively. If no corners are detected in either side, the corner strength threshold is automatically reduced until at least one corner is detected. In the case where only on the one side a keypoint is found, the corresponding keypoint on the other side is assumed using symmetry towards the minor axis [1]. Once the image has got the lip shape, we select lip as the biggest object inside the image. The contour of these detected local maxima pixels will be defined as the left/right corners of the mouth and top/bottom corners. Get the perimeter of the lip region and find the (x, y) of that perimeter. Colour coding is applied on the perimeter for ease of understanding the proper lip contour [11].

IV. IMAGE MONTAGE AND SIMULATION PROCEDURE

As the video file reader divides the live video into individual image frames, each frame is resized and saved in the MAT lab directory after checking for its existence. All disordered frames are sorted and each one is stitched in the ascending order to form a multi frame montage in a single object image montage. These montages are considered to be known templates provided a threshold value is calculated based on the coordinates x and y v/s time (t) for every phoneme or a word from which an average or mean value is obtained [3]. Thus stored features vector in the data base and threshold value comparison test is conducted between these known and unknown templates. If the frames pass the threshold test of average trained tested index value/trials the user can receive a text output & later an audio output. Fig. 3 shows the simulation procedure of the model.

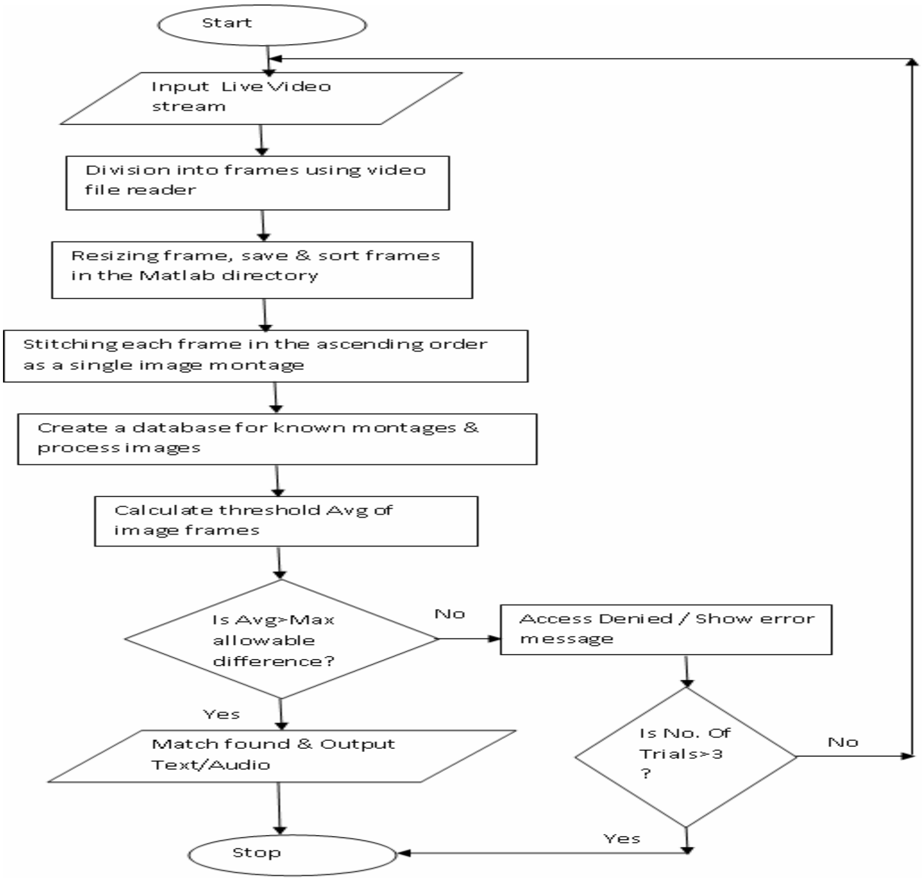

Start
Input Live Video stream
Division into frames using video file reader
Resizing frame, save & sort frames in the Matlab directory
Stitching each frame in the ascending order as a single image montage
Create a database for known montages & process images
Calculate threshold Avg of image frames
Is Avg>Max allowable difference?
No
Access Denied / Show error message
Yes
Match found & Output Text/Audio
Is No. Of Trials>3 ?
No
Yes
Stop

Fig. 3 Simulation procedure

V. Results

In this section, some of the initial results obtained in each phase are shown.

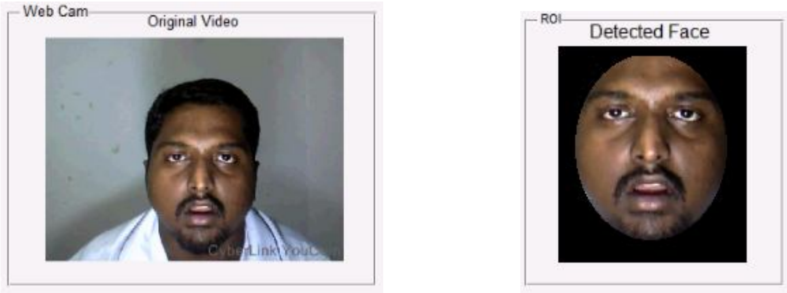


Fig 4 (a) Live video captured by Webcam (b) Region of Interest live video for Silent 'Nihao'

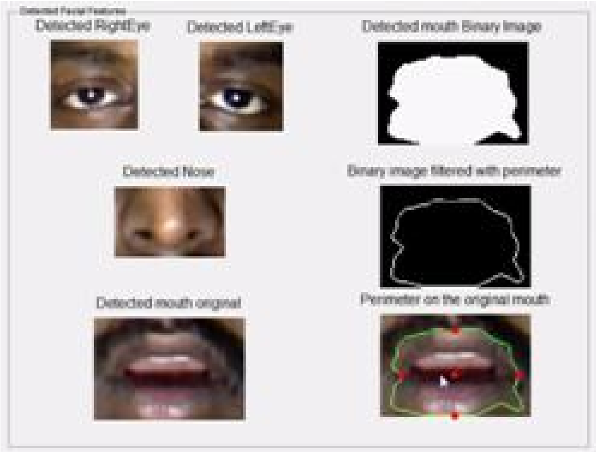

Detected Facial Features
Detected RightEye
Detected LeftEye
Detected mouth Binary Image
Detected Nose
Binary image filtered with perimeter
Detected mouth original
Perimeter on the original mouth

Fig 5 Facial features detected live video for Silent 'Nihao'

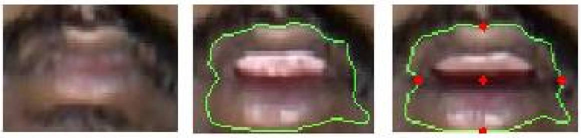


Fig 6 Lip during motion with perimeter contour& key points

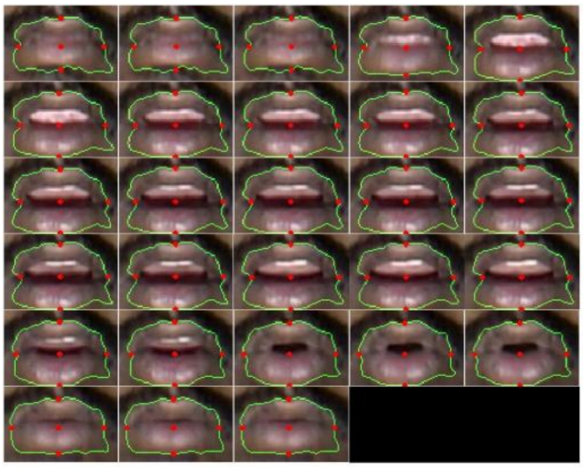


Fig 7 Multi frame image (28 frames) montage in a single object image montage for Silent 'Nihao'

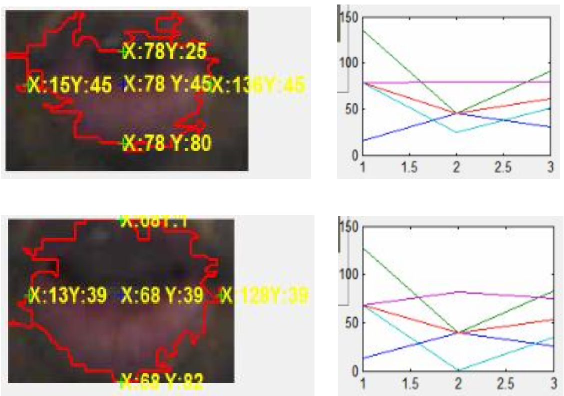


Fig 8 Threshold analysis based on Coordinates X, Y v/s time

VI. Conclusion

Pronunciation of a silent word varies from person to person especially this difference makes the lip-read for Chinese language mandarin highly personalized. The software is being trained based on the lip structure, complexion and features of the lip area. The maximum allowable difference in the threshold value depends on environment and lighting factor. Minimizing the threshold values further enhances the level of security. Inter-disciplinary applications of this lip-reading technique helps for communication where language disorder raises, providing easier mode of communication for people with speech disabilities by converting the identified lip movements directly to speech. Emotional expressions can be correlated and synchronized to gain accuracy and adaptable dynamic. Since many of the systems are still preliminary, it would not make sense to justify the system comparing with speech recognition score or synthesis quality at this stage and hence initial results obtained are shown in the previous section.

When this software integrated onto mobile oriented or hand-held devices and public system machines, silent sound technology could prove to be much more secure, convenient and user-friendly technique for user authentication. Engineers claim that devices work on SST is not still accurate, so this would be an attempt to improve the accuracy

ACKNOWLEDGEMENT

Authors express their deep gratitude and thanks to Mr. Wang, Chairman of the Company Linyi Top Network Co. Ltd for his constant support & encouragement throughout this research work and bring it to technical article.

REFERENCES

[1] Evangelos Skodras and Nikolaos Fakotakis, *An Unconstrained method for lip detection in color images* by, IEEE ICASSP, ISSN : 1520-6149, pp: 1013 – 1016, 2011.

[2] Jian-Gang Wang, Eric Sung, *Frontal-view face detection and facial feature extraction using color and morphological operations*, Journal Pattern Recognition Letters archive, Volume 20 Issue 10, Oct. 1999, Pages 1053-1068.

[3] Kamil S. TALHA1, Khairunizam WAN, S.K.Za'ba, Zuradzman Mohamad Razlan and Shahriman A.B, *Speech Analysis Based on Image Information from Lip Movement*, 5th International Conference on Mechatronics (ICOM'13) IOP Publishing IOP Conf. Series: Materials Science and Engineering 53 (2013) 012016 doi:10.1088/1757-899X/53/1/012016.

[4] Meriem Bendris, Delphine Charlet, and Gerard Chollet, *Lip activity detection for talking faces classification in TV- content 2010*, The 3rd International Conference on Machine Vision (ICMV 2010).

[5] Nicolas Eveno, Alice Caplier and Pierre-Yves Coulon, *A Parametric Model for Realistic Lip Segmentation‖*, 7th International Conference on Design and Implementation of Face Recognition System in Matlab Using the Features of Lips Control, Robotics and Vision, December 2002, pp. 1426 – 1431.

[6] P. Viola and M. Jones. *Rapid object detection using a boosted cascade of simple features. Computer Vision and Pattern Recognition*, IEEE Computer Society Conference on, April 2001.

[7] P. Viola, and M. Jones *Rapid Object Detection Using a Boosted Cascade of Simple Features*, CVPR, Kauai, 2001.

[8] Prof. Samir K. Bandyopadhyay, Lip contour detection techniques based on front view of face, Journal of Global Research in Computer Science, Volume 2, No. 5, May 2011.

[9] Rainer Stiefelhagen, Jie Yang and Alex Waibel, *A Model-Based Gaze-Tracking System Int. Journal of Artificial Intelligence Tools*, Vol. 6, No. 2 (1997), pp. 193-209.

[10] Robert Krutsch and David Tenorio Microcontroller Solutions Group Guadalajara, Histogram Equalization, *Freescale Semiconductor Application* Note Document Number: AN4318, Rev. 0, June 2011.

[11] Sasikumar Gurumurthy, B.K.Tripathy, *Design and Implementation of Face Recognition System in Matlab Using the Features of Lips*, I.J. Intelligent Systems and Applications, 2012, 8, 30-36 July 2012 in MECS.

[12] Sharmila Sengupta, Arpita Bhattacharya, Pranita Desai, Aarti Gupta, *Automated Lip Reading Technique for Password Authentication*, International Journal of Applied Information Systems (IJAIS) – ISSN : 2249-0868 Foundation of Computer Science FCS, New York, USA Volume 4– No.3, September 2012.

[13] S. Gundimada, Li Tao, and V. Asari, *Face detection technique based on intensity and skin color distribution‖*, 2004 International Conference on Image Processing, October 2004, vol. 2, pp. 1413–1416.

[14] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham, *Active shape models - their training and application*. Comput. Vis. Image Underst., 61:38–59, 1995.

AUTHORS BIOGRAPHY

Dr. B.S. Pradeep, B.E(CSE), M.Tech(N/w), Ph.D(CSE), D.Sc(CSE). He is working as Technology Director of Linyi Top network Co. Ltd, China. Overall 13 years experienced in teaching and R&D activities. His areas of interest are networking, mobile computing, Computer organization, system software, Image and video processing.

Mr. Zhang Jingang, B.E in Electrical Engineering. He is working as CEO of Linyi Top network Co. Ltd, China. Overall 14 years experienced in software industry, R&D and Administration. His areas of interest are Software development & management, IOT Image recognition.