International Journal of Applied Information Systems (IJAIS) – ISSN : 2249-0868
Foundation of Computer Science FCS, New York, USA
Volume 4– No.3, September 2012 – www.ijais.org

Automated Lip Reading Technique for Password Authentication

Sharmila Sengupta
B.E., M.E. E.X.T.C.
Department of Computer Engineering, VESIT, Mumbai

Arpita Bhattacharya
B.E. Computer
Department of Computer Engineering, VESIT, Mumbai

Pranita Desai
B.E. Computer
Department of Computer Engineering, VESIT, Mumbai

Aarti Gupta
B.E. Computer
Department of Computer Engineering, VESIT, Mumbai

ABSTRACT

As technology is starting to conquer every strata of the society, the war for protecting confidential data from being intercepted is growing intense by the hour. Biometric- security stands out as the most secure form of authentication in high security zones such as defense, space missions and research head-quarters. Today, forms of password-protection range from face-recognition to retina -scan. Here, we develop a system for recognizing and converting lip movement of an individual into a recognized pattern which is set as a password for the system using image-processing. This system is also a break-through for providing people with motor-disabilities a robust and easy way of protecting their data. By capturing and tracing the successive movement of lips during speech, the corresponding word can be detected. The captured images are represented as points on a two-dimensional flat manifold that enables us to efficiently define the pronunciation of each word and thereby analyze or synthesize the motion of the lips. The motion of lips helps us track the word syllable-by-syllable. With multiple levels of image processing, it becomes possible to set the matching parameters to a very close value, hence not allowing any brute-force or other infamous hacking techniques to break into the user's system. This lip reading technique also serves applications in areas where communication via direct speech is not possible.

General Terms

Image acquisition, pattern recognition, text to speech conversion, human computer interaction, histogram, enhanced blowfish algorithm.

Keywords

Lip-reading, lip-contour, syllable tracking, silent-password, threshold analysis, image encryption.

1. INTRODUCTION

There is a need to provide a large number of people with access to applications in varied contexts. New solutions in basic and applied research in the field of human computer interaction play a major role in integration of information systems. The research work done in the area of automated lip-reading is mainly based on color analysis, mouth parameters and dynamic contour mapping. There exist methods which require special lighting conditions or make-up for distinction of lip movement which are not easily achievable. We have developed a system for recognizing and converting lip movement of an individual into a pattern which will be used as the password for authentication of the individual. This paper examines a secure and voice-less method for recognition of speech-based commands using video without evaluating sound signals. The basic idea underlying silent password involves lip-reading and syllable-identification techniques. Image encryption techniques are applied to scramble the password pattern, thereby increasing the security of the system.

2. BACKGROUND REVIEW

Lip reading has been treated as a stand-alone process by some researchers, while others use it to improve voice recognition systems. Some techniques handled the first ten English letters, considering each of them as a short word. For training, images of a person saying the letters several times were used [9]. All images were aligned using maximum correlation; a new sequence was aligned and matched with each of the possible letters. The amount of preserved energy is tested. The letter which preserves most of the new letter's energy is chosen as the pronounced letter in the new sequence.

Another interesting trial for lip reading was introduced by Bregler et al [11] called 'the bartender problem'. The customer is asked to choose between four different drinks, and due to background noise, the bartender's decision of the customer's request is based only on lip reading. Then, a Hidden Markov Model (HMM) system was trained for each of the four options. With a test set of 22 utterances, the system was reported to make only one error (4.5%).

Duchnowski et al. [8] developed a similar framework for an easy interaction between human and machine. A person, sitting in front of a computer, was recorded and videotaped while pronouncing letters tracking the subject's head and mouth using a neural network based system. The acoustic and visual data was processed by a multi-state time delay neural network system with three layers, and 15 units in the hidden layer. By combining the audio and visual information, they achieved a 20-50% error rate reduction over the acoustic processing alone, for various signal/noise conditions.

Kumar et al. [5] has reported on a speech recognition method based on surface Electromyography (EMG) signals of the speaker's facial muscles that shows the movement of these muscles during speech. However, such techniques require mounting of electrodes on the face of the user. Speech recognition based on visual data is the least intrusive [6] and thus the most attractive. A system called "image-input microphone" [7], analyzes the lip features such as mouth width and height, and computes the corresponding vocal-tract transfer function for speech synthesis.

The lip contour is extracted and tracked in figure 1, the specific labels associated to each tracked lip border is approximated to an ellipse. Internal lip contour has vertical height B and horizontal width A. External mouth borders have vertical height B' and horizontal width A'. The area of the mouth-opening is as calculated in Table 1.

Area of lip contour=$\frac{1}{4}$ [A'B'-AB] π

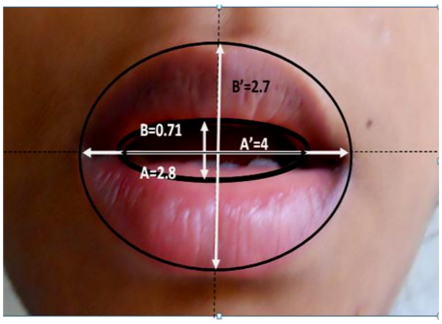


Figure 1: Mouth Parameters approximated as ellipse taken as subject says "he" in 'hello'.

Table 1: Calculation of red area of Lip-contour

Area	**Major axis**	**Minor axis**	**Area** sq inches
Inner Contour	A=2.8"	B=0.71"	1.561
Outer Contour	A'=4"	B'=2.7"	8.482"

Area of red contour= 6.921 sq inches.

We review some of the recent results in the field of automatic analysis of lip movements and information security. Large files, images can be transmitted over computer networks. So data encryption is widely used for safe transmission. Most of the encryption algorithms available are used for textual data. As digital image data is large in size and different from textual data so methods used for textual encryption may not be good for real time, multimedia data.

A novel image encryption method called BRIE (Bit Rate circulation Image Encryption) is a pixel transformation cipher [1]. Ozturk I. and Sogukpinar I, [2] proposed schemes which add compression capability to the mirror like image encryption MIE and visual Cryptography VC algorithms to improve these algorithms. Droogenbroeck M.V. and Benedett R. [3] have proposed two methods for the encryption of an image; selective and multiple selective encryption. Maniccam S.S., Nikolaos G. and Bourbakis [4] have presented a new methodology, which performs both lossless compression and encryption of binary and gray-scale images. The compression and encryption schemes are based on SCAN patterns. The SCAN is a formal language based two dimensional spatial accessing methodologies which can efficiently specify and generate a wide range of scanning paths or space filling curves [4].

Other popular encryption techniques like chaotic algorithms, differential evolution, Rubik's cube or any other customized algorithm can be used to distort the image matrix. Hence it results into an image which is almost un-identifiable and impossible to replicate.

2.1 Syllable tracking

Subjects were videotaped while pronouncing a word, each pronounced six times, each time as a different vowel such as A, E, I, O, U, and 'hello', a consonant that carries an ultra-short vowel or no vowel sound as shown in figure 2**.** Different people pronounce the same vowels differently .Even the pronunciation of the same person in different scenarios may change. Each vowel is pronounced differently when said in different parts of a word. For example, the vowel 'A' in 'America' is different from the vowel 'A' in 'Los Angeles'[9]. This difference occurs due to the location of the syllable 'A' in the word, and the syllables that appear before and after it. The main reason for different pronunciation of the same vowel is the formation of the mouth just before and after this syllable is said.

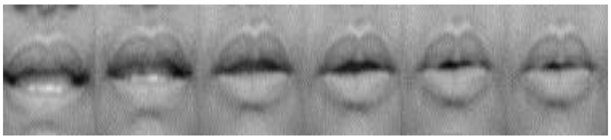


Figure 2: Sequence of lip-movement while pronouncing hello

Each word is isolated into parts, each containing a consonant and a vowel, or a consonant alone, e.g. 'he', 'al', 'di', 'lo' etc. Each of these sounds is considered as a syllable. Each syllable has its own sequence of mouth motions that must occur in so that the sound can be vocalized. Then the lip contour is extracted and tracked in a bounding box. The specific syllables associated to each tracked lip border depending on the illumination conditions.

2.2 Personalization

Instead of handling a generalized case we personalize the software for a single user. First stage of the system is learning process, where the user enunciates the words and the threshold variations are normalized and set. The sequence of pronounced phonemes is equal, when saying the same word. Therefore, after a detailed matching of the lip's configurations for the model and the new person, contours of the same word look similar. Comparing mouth area images of two different people might be deceptive because of different facial features such as the thickness of the lip, skin texture or teeth structure. Moreover, different people pronounce the same phonemes differently, and gray level of mouth's contour comparison between the images might not reveal the true similarity between phonemes.

3. METHODOLOGY

3.1 Block Diagram

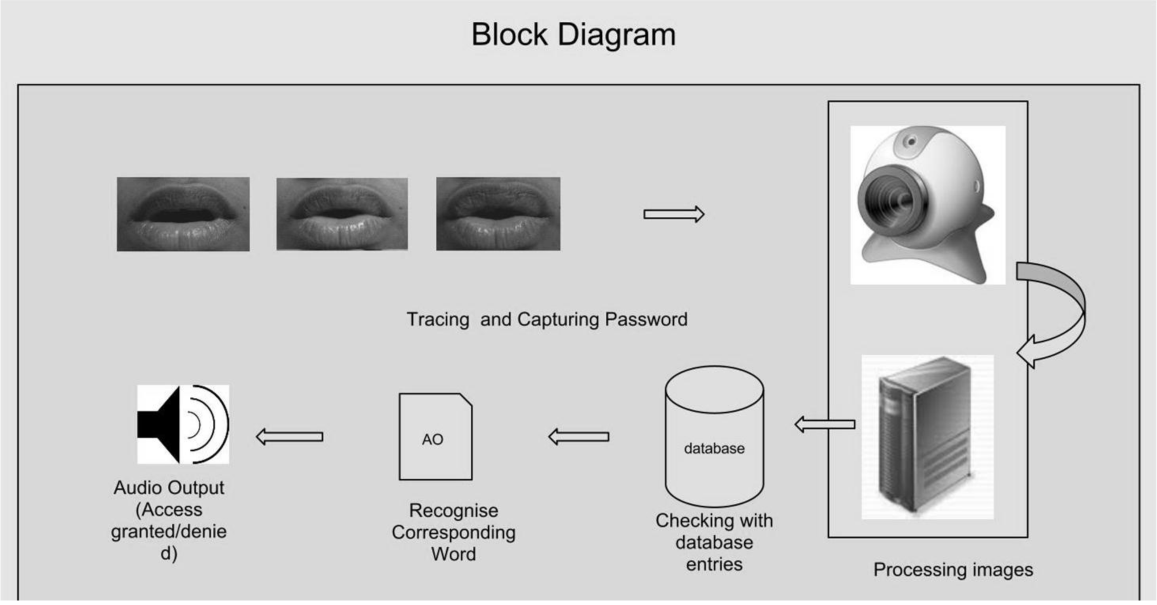

Block Diagram
Tracing and Capturing Password
Audio Output (Access granted/denied)
AO
Recognise Corresponding Word
database
Checking with database entries
Processing images

Figure 3: Block diagram

The elementary blocks in automated lip reading are as shown in Figure 3. A sequence of Images of the user is acquired using web camera. The training set acquires and stores a password spoken by the user. During authentication, the web cam acquires real-time video for verification which is given frame by frame to an image processing unit in MATLAB.

The frames are converted into binary image based on the hue of red color. The sequence of processed images is compared with a pre-existing database of analyzed images obtained as password corresponding to the test subjects' user-name. The word spoken by the user is recognized syllable by syllable. The last block of the system is an additional feature for audio-authentication. If the frames pass the threshold test, the user is granted access into the system with an audio output of "WELCOME" or any welcome text. The maximum number of allowable trials can be limited to three for instance.

3.2 Simulation Procedure

A flow chart of the steps involved in our simulation technique is shown in figure 4. First, the web camera acquires a fixed number of frames of the lip as the user enunciates the password. It captures a sequence of 16 or more frames using image acquisition <imaq> tool in matlab and stores it as a video file (.avi). The frames of the avi file are then extracted as jpg images per frame. The 16 by 16 frames obtained by enunciation of the word "hello" is represented in figure 5. This is a reverse of the process that the animation industry uses to animate pictures in sequence. Each frame of this is video is now ready for processing.

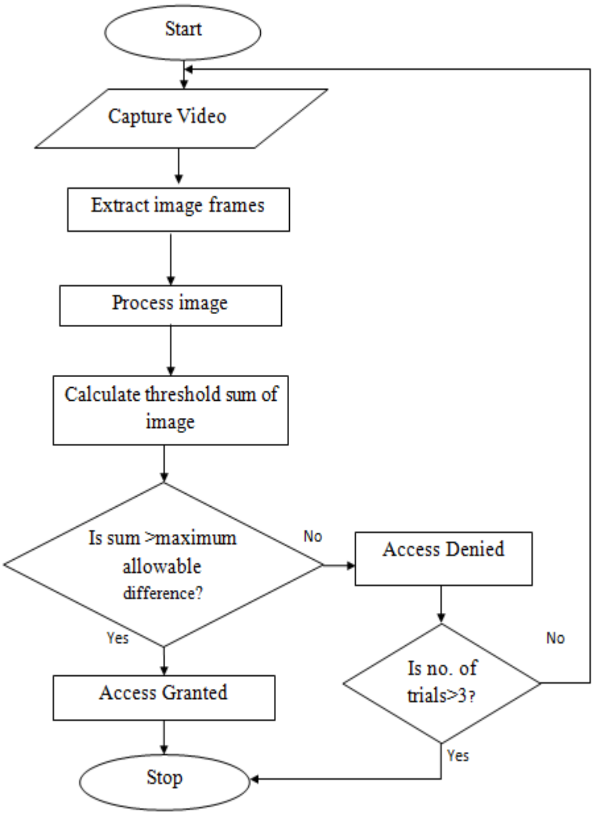


Figure 4: Flow chart for authenticating the lip-read password

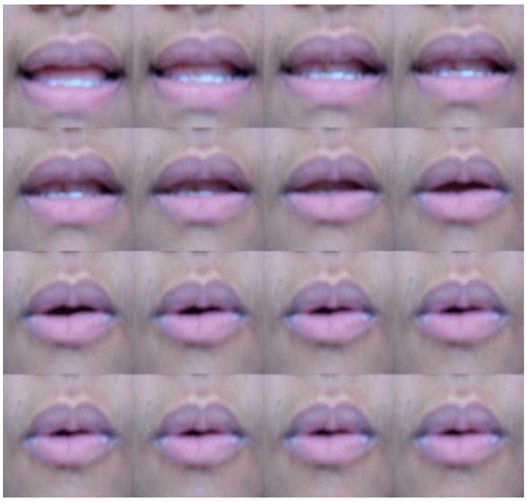


Figure 5: Frames of the captured video for 'Hello' said by a subject.

The images are converted to black and white binary format as shown in figure 6, to avoid the dissimilarities in color hues and intensities. A threshold is determined for the red component of the image from the lip region of the face. Each frame is converted into binary using the threshold value. However the images should be having a uniform lighting. Variations in lighting conditions will lead to variation in threshold factor which is a potential cause for the mismatch of images.

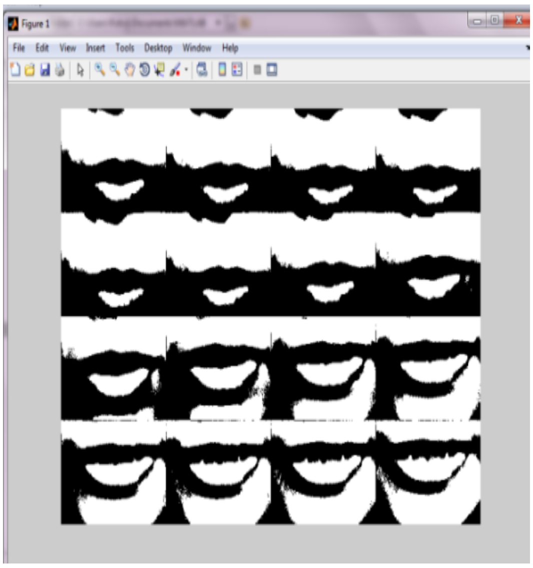


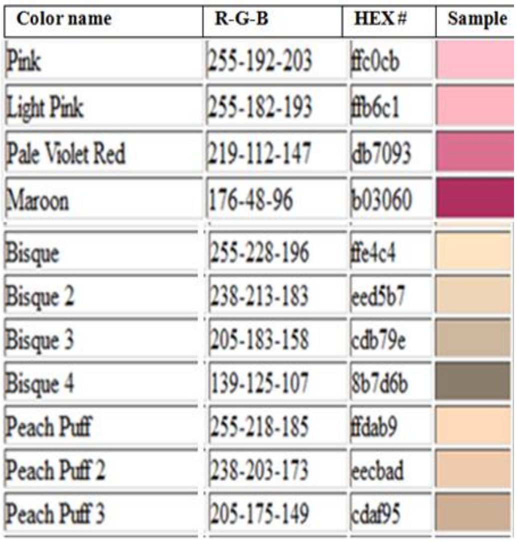
Figure 6: Red Threshold Analysis of figure 5.

3.2.1 RGB Filter

The RGB color model is used for specifying colors. This model specifies the intensity of red, green, and blue on a scale of 0 to 255 are as shown in figure 7, with 0 (zero) indicating the minimum intensity. The values are indicated for shades of skin for the face and pink for the lip. The RGB Filter used focuses the RGB values towards the primary red hue in the RGB colors. Based on the threshold generated, this filter will diminish the pixels that are not of the selected colors and normalize the image. R is normalized with respect to the maximum red value. Due to normalization really dark pixels can be elevated in intensity and generate too much noise in the resulting image.

For a true RGB image, the color is stored in a [m*n*3] dimensional matrix which gives the RGB components of each pixel. These component values are set in order to filter the desired color. Each pixel in the two dimensional image is identified by its coordinates (i ,j) where;

i: x-coordinate of the pixel
j: y-coordinate of the pixel



Color name	R-G-B	HEX#	Sample
Pink	255-192-203	ffc0cb	■
Light Pink	255-182-193	ffb6c1	■
Pale Violet Red	219-112-147	db7093	■
Maroon	176-48-96	b03060	■
Bisque	255-228-196	ffe4c4	■
Bisque 2	238-213-183	eed5b7	■
Bisque 3	205-183-158	cdb79e	■
Bisque 4	139-125-107	8b7d6b	■
Peach Puff	255-218-185	ffdab9	■
Peach Puff 2	238-203-173	eecbad	■
Peach Puff 3	205-175-149	cdaf95	■

Figure 7: RGB color model for pink and skin related colors.

The values of separate Red, Green and Blue components of a pixel (i,j) of an image 'Img' are obtained in a range of 0-255 as follows:

```
Red_value = Img (i, j,1);
Green-value = Img (i,j,2);
Blue_value = Img (i,j,3)
```

The interior region of the lip appears blue in the colormap plot of the pattern in figure 8. The contour of the lip is traced in green lines in figure 9. These techniques provide high accuracy in pattern recognition for desired password.

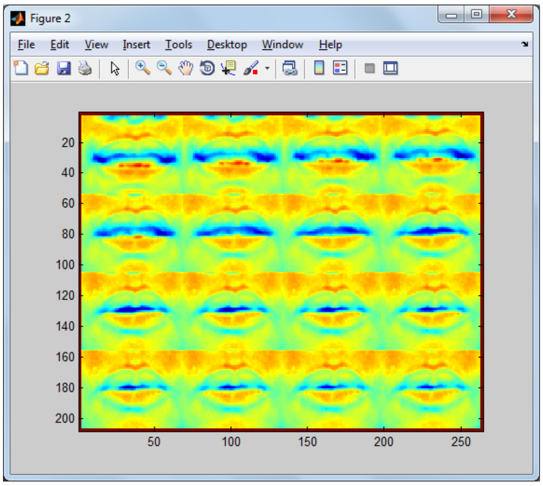


Figure 8: Color map analysis for img(:,:,1) for figure 5.

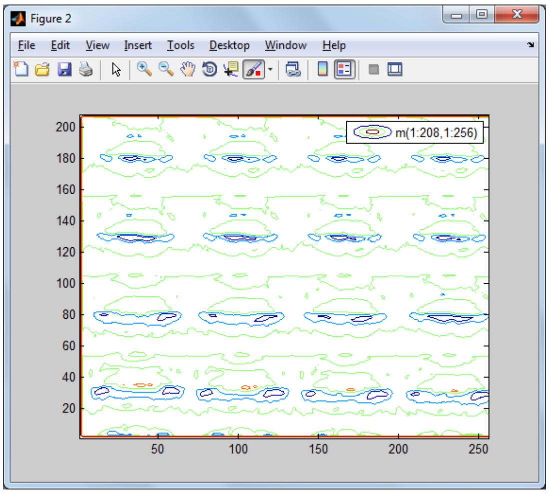


Figure 9: Contour plot for figure 5.

3.2.2 Syllable tracking

The processed images are then compared by finding the difference between the corresponding frames of the obtained image and the reference image for the defined password. The difference image is again normalized over a threshold. The threshold value for various set of images taken from different subjects was calculated. Minimum threshold values that gave correct results ranging from 300000-600000 were obtained. This value was set as the minimum allowable difference parameter for each syllable in the password.

```
if sumOfAllPixels<minAllowableDifference
    % Accept the syllable.
    % Append to password.
else
    % Deny access.
```

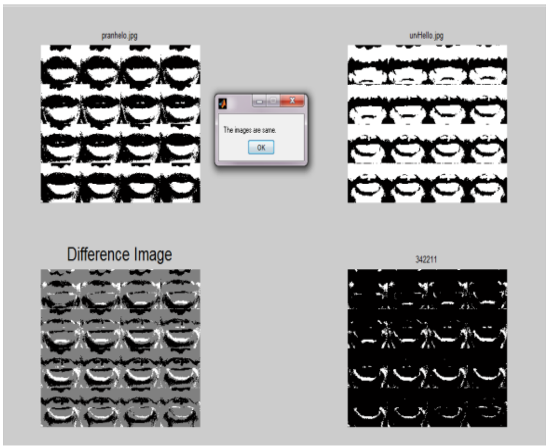
If the first syllable is accepted, the next set of frames is verified. Each syllable is appended to the next till the final password is authenticated. If the syllables do not match the user is allowed to try again until the system blocks after a fixed number of trials.

4. RESULT ANALYSIS

The snapshot of a comparative analysis of a selected case is shown in figure 10. In this case 'hello' was selected as the password of the user. Access is granted only when the same user says 'hello'. Figure 11 and figure 12 give the histogram plots of the image frames acquired on each trial. Access is denied when the user says 'no', figure 13. Other words like 'yes', 'no' and similar sounding words like 'yellow' were also verified. Similar analysis has been done for other test cases and their threshold difference is recorded as shown in Table 2.

4.1 Test cases

In testing the simulation, following were the test cases verified:
Case 1: Same words said by different users.
Case 2: Similar sounding words said by the same user.
Case 3: Different words said by the same user.
Case 4: Different words said by different users.



The images are same.

Figure 10: 'Hello' said by the same subject as compared to stored password 'Hello' (Access granted).

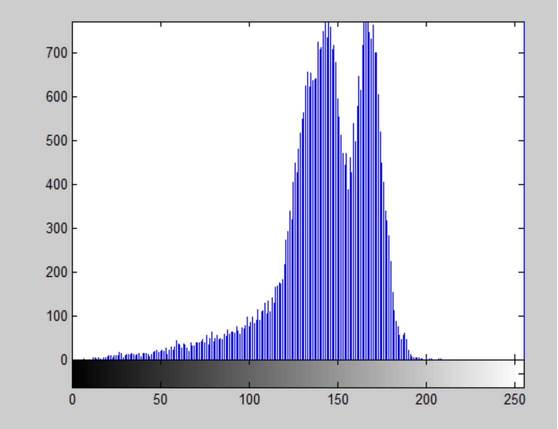


Figure 11: Histogram plot of 'Hello' stored as password.

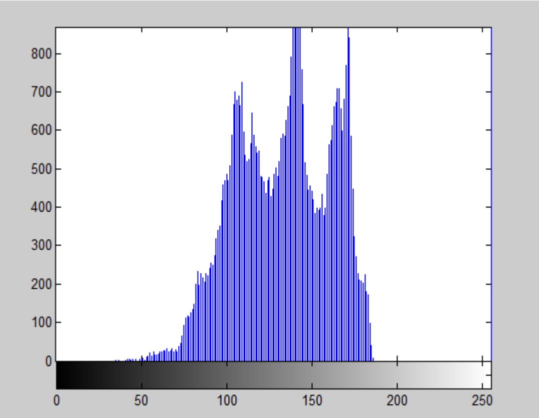
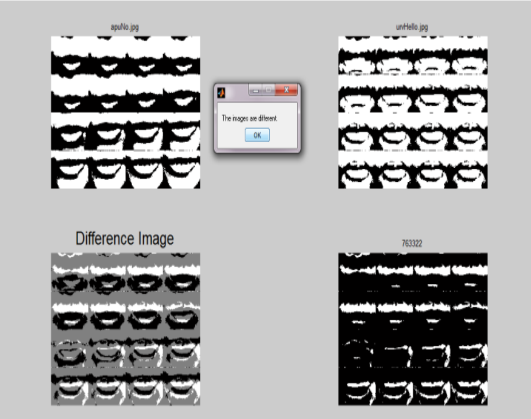


Figure 12: Histogram plot of 'Hello' taken for password comparison.

Set minimum allowable threshold difference=350000.

Table 2: Analysis of test cases for different set of subjects.

	Test cases	Threshold difference	Result
1.	Hello said by same subject.	342211	Access Granted.
2.	No and hello said by two different subjects.	763322	Access denied.
3.	No said by different subjects.	599485	Access denied.
4.	Hello said by two different subjects	425850	Access denied.
5.	Yes said by same subject.	325821	Access Granted .
6.	Hello and Yellow said by different subjects.	375125	Access denied.
7.	Yes said by two different subjects.	548538	Access denied.
8.	Hello and yellow said by user under variant lighting condition.	1.00131e+006 (aberrant value)	Access denied.



The images are different.

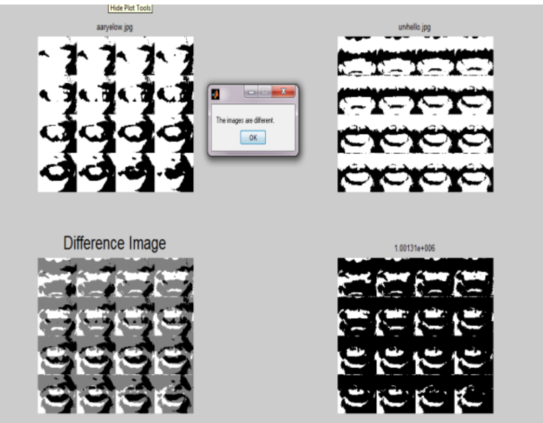
Figure 13: 'No' said by the subject doesn't match with password 'Hello' set by same subject. (Access denied).

Test case 8 gives a very high aberrant value in spite of very little difference in enunciation of 'hello' and 'yellow' due to high difference in lighting conditions. The distorted threshold pattern for 'yellow' under high illumination can be observed in Figure 14. A normal closer value is obtained in test case 6 for stable lighting condition. Table 3 gives a summary of the result obtained for single user trying out different words. This can be used to train the software to predict and identify user patterns.

Table 3: Analysis table for words said by same user.

	Input			
Password	Hello	Yes	No	Yellow
Hello	1	0	0	0
Yes	0	1	0	0
No	0	0	1	0
Yellow	0	0	0	1

0: access denied 1: access granted



The images are different.

Figure 14: 'Hello' and 'yellow' said by same user under variant lighting conditions gives aberrant values (Access denied).

5. IMAGE ENCRYPTION

In order to increase the security of the system, the password-image is stored in an encrypted form. Derivatives of strong algorithms like DES which are applied for text encryption can be used to encrypt and decrypt the images. Here the image matrix is treated as the plain text. For example, the Enhanced Blowfish algorithm [12] has a block size is 64 bits, and the key can be any length up to 448 bits. By using this algorithm, 64 bit data can be encrypted into RGB values and plotted as a pixel in a bitmap image. Thus the security of blowfish algorithm is enhanced by using water marking technology. Here, the original image is the binary black & white image. The binary matrix of the chosen image is then encrypted the image matrix using keys generated over a series of 16 rounds. When the user speaks the password same algorithm is applied on the newly generated frames and a stronger comparison is obtained [10]. Since only the encrypted image is stored in the database, even if a hacker is able to intercept into the system, he cannot recover the password since it is scrambled.

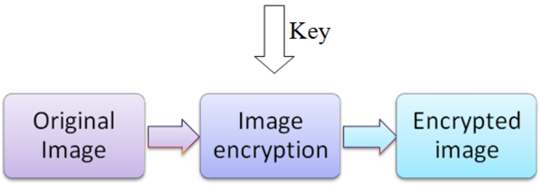


Figure 15: Image Encryption

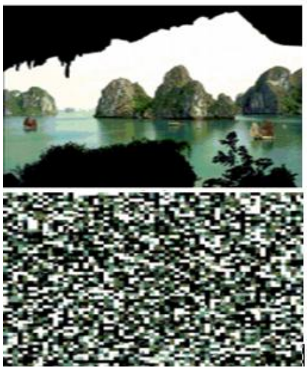


Figure 16: Example of an encrypted image. [10]

6. CONCLUSION

The difference in pronunciation of different people makes the lip-read password highly personalized. The software is trained based on the lip structure, complexion and features of the lip area. Environment and lighting are a limiting factor which greatly varies the maximum allowable difference in the threshold value. Minimizing the threshold values further enhances the level of security. Inter-disciplinary applications of this lip-reading technique include communicating in outer-space where medium for sound to travel is not available, providing easier mode of communication for people with speech disabilities by converting the identified syllables directly to speech. The future of security based on human computer interaction lies in artificial intelligence and neural networks which can train the system to be highly adaptable and dynamic. When integrated onto hand-held devices and public system machines, lip read passwords could prove to be much more secure, convenient and user-friendly technique for user authentication.

7. ACKNOWLEDGMENT

We thank Professor Parmeshwar Birajdar from Department of Electronics Engineering, VESIT, who is pursuing his Ph.D. in the field of image-processing from IIT Bombay, and has guided us with his expertise.

8. REFERENCES

[1] Jui-Cheng Yen and Jiun-In Guo, "A new image encryption algorithm and its VLSI architecture," in Proc. IEEE workshop Signal Processing Systems, 1999,pp. 430-437.

[2] I.Ozturk, I.Sogukpinar, "Analysis and comparison of image encryption algorithm," Journal of transactions on engineering, computing and technology December, vol. 3,2004,p.38.

[3] M. V. Droogenbroech, R. Benedett, "Techniques for a selective encryption of uncompressed and compressed images," In ACIVS'02, Ghent, Belgium. Proceedings of Advanced Concepts for Intelligent Vison System, 2002.

[4] S.S. Manicaam., G. Nikolaos, and Bourbakis, "Lossless image compression and encryption using SCAN," Journal of Patterns Recognition, vol. 34, no. 6, 2001, pp.1229-1245.

[5] S. Kumar, D. K. kumar, M. Alemu, and M. Burry, "EMG based voice recognition," In Intelligent Sensors, Sensor Networks and Information Processing Conference, 2004.

[6] G. Gravier A. Garg G.Potamianos, C. Neti and A. W. Senior, "Recent Advances in the automatic recognition of audio-visual speech," Proc. Of the IEEE, vol. 91,2003.

[7] K. Otani, and T. Hasegawa, "The image input microphone-A new nonacoustic speech communication system by media conversion from oral motion images to speech," IEEE journal on Slected Areas in Communications, vol. 13,no. 1, pp.42-48, January1995.

[8] Duchnowski, P., Hunke, M., Bsching, D., Meier, U., and Waibel,A. 1995.Toward movement-invariant automatic lipreading and speech recognition. In Proc. ICASSP'95, pp. 109–112.

[9] Michal Aharon, Ron Kimmel, Representation Analysis and Synthesis of Lip Images Using Dimensionality Reduction, International Journal of Computer Vision 67(3), 297–312, 2006.

[10] Mohammad Ali Bani Younes and Aman Jantan,Image Encryption Using Block-Base Transformation Algorithm, IAENG International Journal of Computer Science, 35:1, IJCS_35_1_03

[11] Bregler, C., Hild, H., Manke, S., and Waible, A 1993. Improving connected letter recognition by lip reading. In proc. IEEE Int. Conf. on ASSP, pp. 557-560.

[12] Dinesh Kumar Jain N. ; Palaniswamy N, Raaja Sarabhoje, G. Enhanced Blowfish algorithm using bitmap image pixel plotting for security improvisation, 2nd International Conference on Education Technology and Computer (ICETC), 2010.