Facial Expression and Visual Speech based Person Authentication

S. Saravanan
Research Scholar
Dept. of Computer Science & Engg.
Annamalai University, India

S. Palanivel
Professor
Dept. of Computer Science & Engg.
Annamalai University, India

M. Balasubramanian
Assistant Professor
Dept. of Computer Science & Engg.
Annamalai University, India

ABSTRACT

Most of the person authentication system lacks perfection due to face poses and illumination variation. One more problem in person authentication is the selection of source for feature generation. In this work, videos have been recorded with variations in poses. The videos have been taken in normal office lighting condition. Videos of persons are taken in three situations. First when faces are kept normal, then with smile facial expression and third during speech. Second session of video recording is done similar to the first session with a time gap. This work employs a powerful method to identify the video frames which have single face without pose and excerpt necessary number of frames from the video. Methods are used to automatically identify the mouth area. Features are generated from mouth area in such a way to overcome the issues due to illumination variation. The features created from the first and second session are used to train and test respectively a neural network for person authentication. Among several neural network models, auto associative neural network is used due to its features distribution capturing ability. Person authentication capacity is compared while features created from normal face, features created from smile expression, and visual speech. Equal error rate is used as a tool to compare the capacity of person authentication. The outcome of this project is that while intensity based feature vectors like this is used for person authentication, the visual speech is more efficient than normal face, and face with smile expression performs the lowest.

General Terms

Authentication, Face detection, Feed forward neural networks, Image and video processing

Keywords

Auto associative neural network, Automatic pose free face detector, Facial expression, Person authentication, Visual speech.

1. INTRODUCTION

Person authentication or verification is the process of deciding whether the particular person claimed identity is correct or not. Person authentication is a one to one comparison in contrast to person identification, which is one to many comparison of the stranger with the database of persons. Practically person authentication can be employed for access control in so many places like automated teller machines and entrances of an institute or organization. Lot of researches has dealt with person authentication using face and its components. But still it needs more research to achieve total perfection. Main impediments in achieving quality in a person authentication using face are the face poses and illumination variations. Poses may be tilt, roll and yaw. Tilt is also called pitch, which is moving the head up and down. Roll is slanting the head sideways towards the shoulders while the nose is stationary. Turning the face to left and right side is called yaw. Artificial neural network works just like the central nervous system of animals and has the capability to recognize patterns. The performance of a biometric verification system like facial expression and visual speech based person authentication is usually measured using equal error rate. It is the rate at which the rate of false acceptance and rate of false rejection are equal. Person verification system with lower equal error rate is regarded as performing better. This proposed work consists of two phases. Phase one is deciding of mouth area. Section 3 explains phase one. Phase two is person authentication. Section 4 explains phase two. Phase one, deciding of mouth area includes face detection which is explained in section 3.1, nose detection, which is explained in section 3.2, detecting video frames with single face, which is explained in section 3.3, landmarks detection in the face, which is explained in section 3.4, identifying boundaries of mouth area, which is explained in section 3.5, filtering video frame based on poses, which is explained in section 3.6. Phase two, person authentication includes feature excerption, which is explained in section 4.1, auto associative neural network based person authentication, which is explained in section 4.2. Section 5 explains the potential of auto associative neural network. Section 6 is analysis of results. Section 7 is conclusions and future work. Figure 1 exhibits the schema of facial expression and visual speech based person authentication system. Figure 2 exhibits the user interface of this person authentication system.

2. RELATED WORK

In authentication system of single person, if two or more face or nose is observed in a video frame, that frame can be rejected to boost performance [1]. For pose recognition eyes and mouth are regarded as significant landmarks. To recognize positions in a face, color details can be used. Performance of a person authentication system is usually measured utilizing equal error rate [2]. Center of the mouth can be computed from eyes position [3]. A real life authentication system which uses face has to handle the variations in poses without human interaction [4]. As per many studies, an authentication system which utilizes more than one method shows improvement in results [5]. Variations in pose subdue the efficiency of authentication systems which uses image. Video can yield large quantity of data than still images [6]. Finding the position of the mouth area accurately is a crucial issue for person authentication system [7]. In auto associative neural network, the number of epochs required hinge upon the training error [8]. In place of finding the features relevant to recognition in a mouth image and then excerpting the features, the full image can be utilized. It is advantageous as it has teeth, tongue and skin texture in addition to lips [9]. In speaker recognition job, extremely dependable single modality system may sometime perform well than multimodal approach, if a less dependable modality in the multimodal system reduces its performance [10]. The only lip tracking system that is publicly available is the one which is in Intel audio visual continuous speech recognition toolkit [11]. Compared with face or voice based biometric systems, lip based biometric systems are barely dealt with in scientific literature [12]. Features for visual speech can be based on intensity, shape or mouth movement [13]. Works on speaker recognition using lips are very less, mainly because, to model the biometrics, the non-principal components of the lip are also important [14].

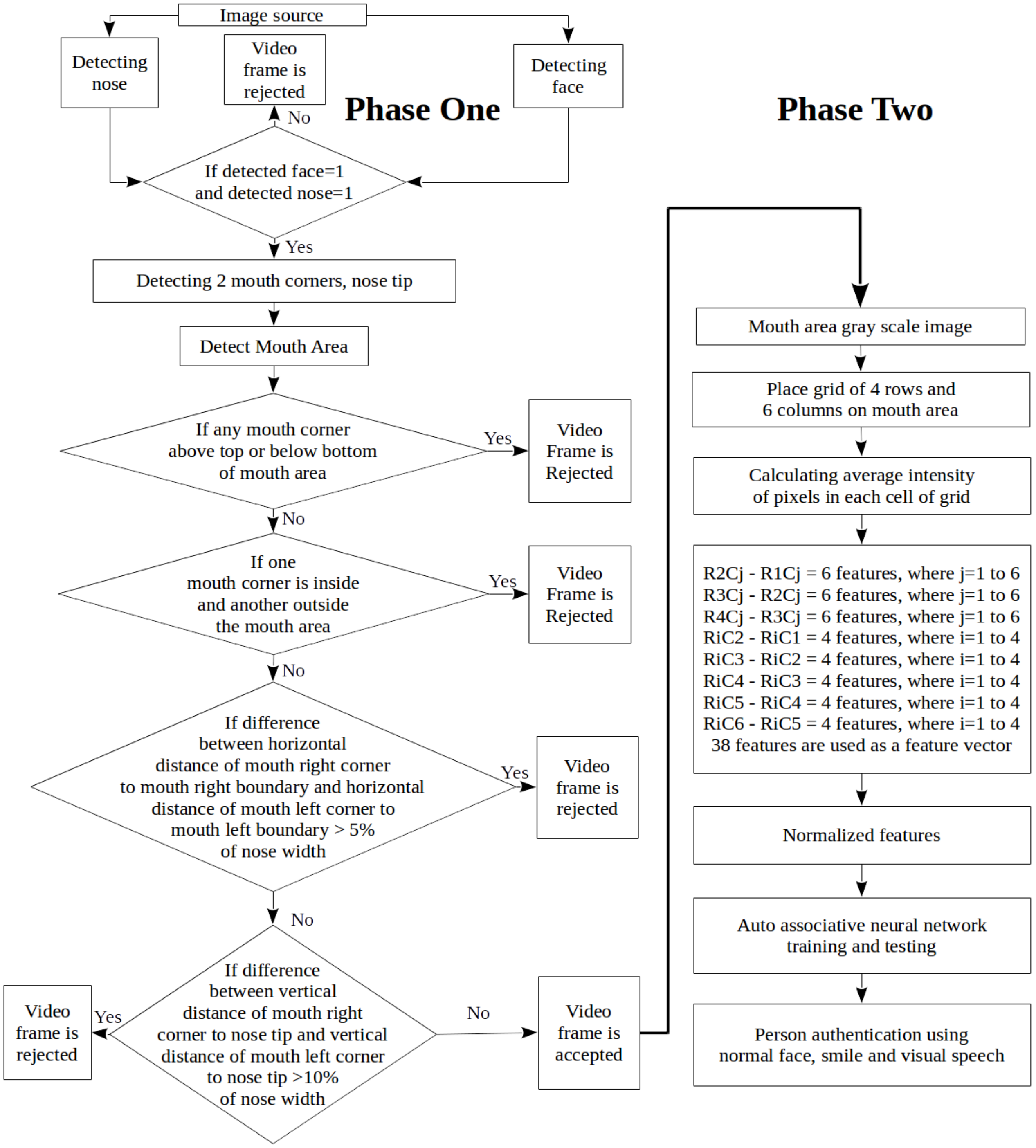
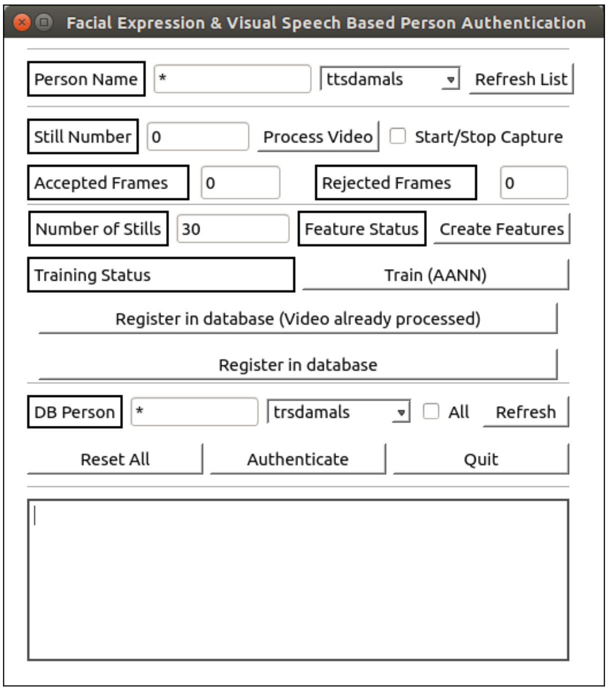

Phase One
Image source
Detecting nose
Video frame is rejected
Detecting face
No
If detected face=1 and detected nose=1
Yes
Detecting 2 mouth corners, nose tip
Detect Mouth Area
If any mouth corner above top or below bottom of mouth area
Yes
Video Frame is Rejected
No
If one mouth corner is inside and another outside the mouth area
Yes
Video Frame is Rejected
No
If difference between horizontal distance of mouth right corner to mouth right boundary and horizontal distance of mouth left corner to mouth left boundary > 5% of nose width
Yes
Video frame is rejected
No
If difference between vertical distance of mouth right corner to nose tip and vertical distance of mouth left corner to nose tip >10% of nose width
Yes
Video frame is rejected
No
Video frame is accepted
Phase Two
Mouth area gray scale image
Place grid of 4 rows and 6 columns on mouth area
Calculating average intensity of pixels in each cell of grid
R2Cj - R1Cj = 6 features, where j=1 to 6
R3Cj - R2Cj = 6 features, where j=1 to 6
R4Cj - R3Cj = 6 features, where j=1 to 6
RiC2 - RiC1 = 4 features, where i=1 to 4
RiC3 - RiC2 = 4 features, where i=1 to 4
RiC4 - RiC3 = 4 features, where i=1 to 4
RiC5 - RiC4 = 4 features, where i=1 to 4
RiC6 - RiC5 = 4 features, where i=1 to 4
38 features are used as a feature vector
Normalized features
Auto associative neural network training and testing
Person authentication using normal face, smile and visual speech

Fig 1: Schema of facial expression and visual speech based person authentication system


Facial Expression & Visual Speech Based Person Authentication

Person Name: * | ttsdamals

Still Number: 0

Accepted Frames: 0

Rejected Frames: 0

Number of Stills: 30

Feature Status

Training Status

DB Person: * | trsdamals

Fig 2: User interface of this person authentication system

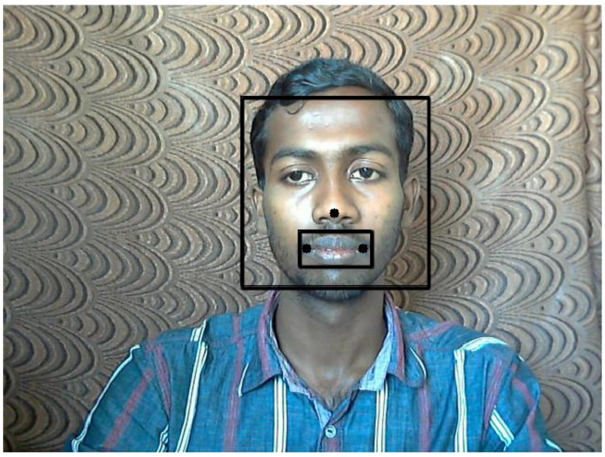


Fig 3: An admitted video frame

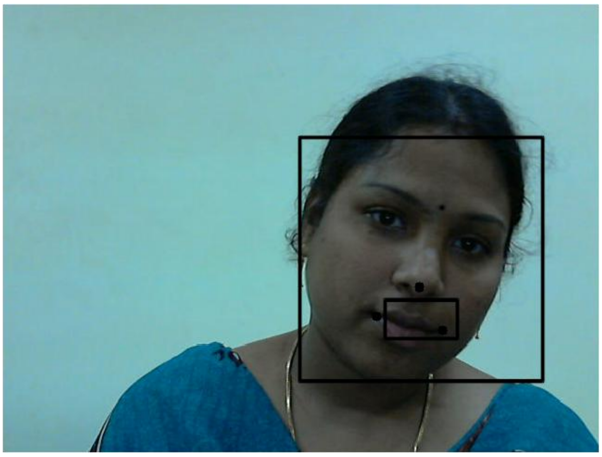


Fig 4: A removed video frame

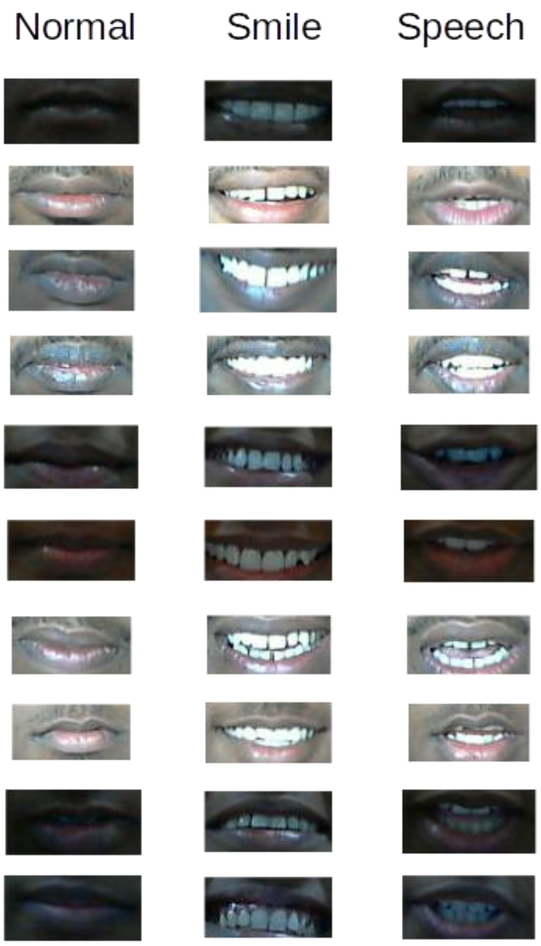


Fig. 5: Samples of excerpted mouth area

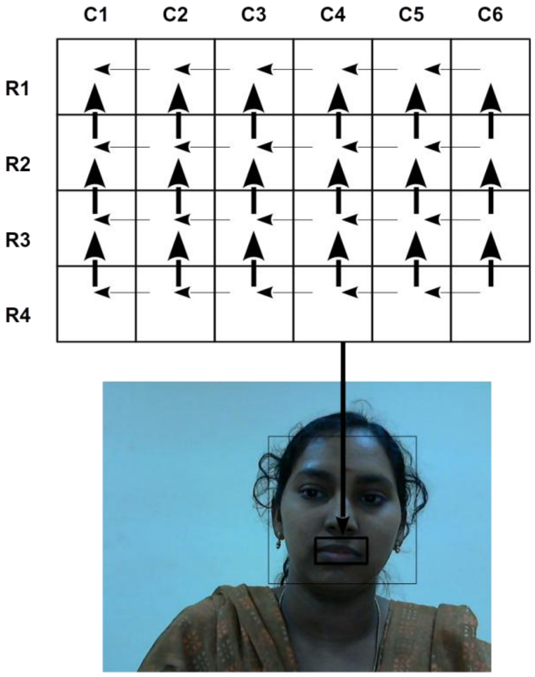


Fig. 6: Feature excerption process

3. DECIDING OF MOUTH AREA

3.1 Face Detection

To rapidly detect frontal face from an image with less computation time, in [15] a harr like simple feature along with cascade classifier was used. Cascade classifier was like a degenerated decision tree with multiple stages. In each stage the classifier rejects certain portion of image as not a frontal face. Adaboost algorithm was used to train each stage of the cascade classifier. In [16], the performance was further improved with rotated haar like features, called extended set of haar like features. Even at different scales, the time taken to detect frontal face will be same. This was used in our proposed work for face detection, due to its improved performance.

3.2 Nose Detection

In [17], greater performance and speed was achieved in detection by combining elements of technologies from two families of detection namely implicit or pattern based and explicit or knowledge based. For example, geometry and color can be considered as explicit and cascade classifiers can be considered as implicit. Temporal coherence is also integrated to accelerate the processing of the succeeding frame. This was used to detect nose separately apart from face as this method showed better performance than the method in [16] for nose detection on empirical studies. From each video frame in which the nose was detected, the width of the nose, w_{nose} was calculated.

3.3 Detecting video frames with single face

The proposed person authentication system was designated to handle one person at a time. In some situations more than one person face may appear in the video frame. In some video frames detectors may detect face or nose in more than one location falsely but there may be only one. To remove video frames of this type, every video frame was tested by both the systems and removed if multiple faces or multiple noses are found by any one of the system. Due to this multi-modality approach the perfection of accepting video frames with only one person increases. Also video frames having additional face like objects get filtered in this process, which boosts the total quality of the system.

3.4 Landmarks Detection in the Face

For landmarks detection in the face, the detected face had to be given as input to a facial landmark detector. In [18], it was shown that the facial landmark detector based on the Deformable Part Models using structured output Support Vector Machine for supervised learning of the parameters of the landmark detector from examples perform better. Hence this is used in our proposed work as facial landmarks detector. Other landmark detectors used for comparison are the independently trained Support Vector Machine landmark detector, the Deformable Part Model based landmark detector, and Active Appearance Model based landmark detector. This facial landmark detector learned by the structured output Support Vector Machine assumes that the number of face in the input image is always one. Local Binary Patterns pyramid feature descriptor is used as feature descriptor for the local appearance model as it outperforms normalized intensity values, derivatives of image intensity values, and histograms of Local Binary Patterns. Landmarks detected by this detector are center of the face, both corners of both eyes where the upper and lower eyelids meet, both mouth corners and tip of the nose. For our proposed work, among these landmarks only nose tip and both mouth corners are used.

3.5 Identifying Boundaries of Mouth Area

The main ideas which are used in this are that mouth width, w_{mouth} is assigned the value of detected nose width, w_{nose}, for a particular face with width w_{face}, the ratio of width of face by width of nose, w_{face} / w_{nose} is constant, and the height of the mouth, h_{mouth} is calculated as half of the width of the mouth, w_{mouth}. So $h_{mouth} = w_{mouth} / 2$. All calculated sizes like width of mouth, w_{mouth} and height of mouth, h_{mouth} are relative quantity of output values of systems like width of nose, w_{nose}. This ensures that even if scaling of faces occur due to changes in the distance between the face and the camera, the mouth width and mouth height always covers the same location in the face for a particular person. The top edge of the mouth area is calculated as the middle location of the vertical distance between the right corner of the mouth and the tip of the nose. The horizontal location of the tip of the nose is fixed as the horizontal center of the mouth area. The bottom edge of the mouth area is calculated by adding the height of the mouth with the top edge of the mouth area. The left edge of the mouth is calculated by adding half the mouth width to the middle location of the mouth. Similarly the right edge of the mouth is also calculated. Thus all the four boundaries of the mouth area are defined.

3.6 Filtering Video Frame Based on Poses

In this the video frames which have poses are identified and removed. The main idea used for filtering is that the nose detection and facial landmarks detection is done independently. Mouth area is calculated using both of them. Mouth area and mouth corners in detected facial landmarks align well if the face is free from poses. A video frame is removed when any one of the corner of the mouth is above or below the mouth area. This filters video frames of faces with roll. A video frame is not accepted when one corner of mouth is inside the mouth area and another corner of mouth is outside the mouth area horizontally. This filters video frames of faces with yaw. For each frame, the absolute horizontal distance between the right corner of mouth and right edge of mouth area was calculated. Similarly, the absolute horizontal distance between the left mouth corner and left mouth area is calculated. A video frame is removed, when the absolute difference between the above two differences is greater than 5% of the width of the nose. For each frame, the absolute vertical distance between the right mouth corner and tip of the nose is calculated. Similarly, the absolute vertical distance between the left mouth corner and tip of the nose is calculated. A video frame is removed, when the absolute difference between the above two differences is greater than 10% of the width of the nose. This further filters video frames of faces with yaw. As threshold values are based on relative size of nose, no issues due to scaling arises. After automatically removing all the video frames with poses, pose free video frames are given as input to phase two. An admitted video frame is shown in Figure 3. A removed video frame is shown in Figure 4.

4. PERSON AUTHENTICATION

4.1 Feature Excerption

The pose free face video frames got as input from phase one is converted to grey scale and the mouth area is excerpted. Samples of excerpted mouth area are shown in Figure 5. A grid of four rows and six columns is virtually placed over the image to create features. An intersection of a row and column creates a cell. The number of rows and columns decide the size of the cell. The size of the cell is proportional to the tolerance of alignment in the mouth area between the training and testing images, which in turn is proportional to the accuracy of person authentication. Number of cells is proportional to the number of features, which in turn is proportional to the accuracy of the person authentication system. Hence an optimum number of rows and columns are used for the grid based on experimental studies. Average intensity of pixels is calculated for each cell. The initial six features are created by deducting the average intensity of each cell in row one from the average intensity of matching cell of row two. Features seven to twelve are created by deducting the average intensity of each cell in row two from the average intensity of matching cell of row three. Features thirteen to eighteen are created by deducting the average intensity of each cell in row three from the average intensity of matching cell of row four. Features nineteen to twenty two are created by deducting the average intensity of each cell in column one from the average intensity of matching cell of column two. Features twenty three to twenty six are created by deducting the average intensity of each cell in column two from the average intensity of matching cell of column three. Features twenty seven to thirty are created by deducting the average intensity of each cell in column three from the average intensity of matching cell of column four. Features thirty one to thirty four are created by deducting the average intensity of each cell in column four from the average intensity of matching cell of column five. Features thirty five to thirty eight are created by deducting the average intensity of each cell in column five from the average intensity of matching cell of column six. These thirty eight features actually give the horizontal intensity variations and vertical intensity variations in the mouth area. The feature excerption process is shown diagrammatically in Figure 6. The pseudo code for this feature excerption algorithm is shown in Algorithm 1. As features are created by deduction of average intensity values it will boost tolerance to intensity changes. To have more tolerance to changes in illumination the resultant features are normalized to a fresh minimum 0 and maximum 1 using the formula

$$fn = (fk - fmin)\ x\ \frac{(nmax - nmin)}{(fmax - fmin)} + nmin$$

where
fn= Intensity after normalization
fk = Current intensity
fmin = Current minimum value
fmax = Current maximum value
nmin = 0, which is the new minimum value
nmax = 1, which is the new maximum value

```
Function Feature_Excerption
Data: m, n, a[i][j]
Result: f[k]
/* Process Starts */
for k = 1 to (m-1)*n do
    for i = 1 to [m-1) do
        for j = 1 to n do
            f[k] = a[i+1][j] – a[i][j]
        end for
    end for
end for
for k = [(m-1)*n+1] to [(m-1)*n+m*(n-1)] do
    for j = 1 to [n-1] do
        for i = 1 to m do
            f[k] = a[i][j+1] – a[i][j]
        end for
    end for
end for
end
```

Algorithm 1: Pseudo code of feature excerption algorithm

In this algorithm,
f[k] is the feature vector, where
f[k] = f[1], f[2], f[3], . . . , f[(m-1)*n+m*(n-1)]
a[i][j] = average intensity of ith row and jth column cell
i = 1 to m
j = 1 to n
m = number of rows
n = number of columns
In this proposed work, m = 4, n = 6
[(m-1)*n+m*(n-1)] = 38 and
f[k] = f[1], f[2], f[3], ..., f[38]

4.2 Auto Associative Neural Network Based Person Authentication

Videos are recorded for ten persons in the first session, out of which five are males and five are females. The videos are recorded in normal room conditions with normal face, smile expression and speech. Throughout the three cases, people are asked to move their head enough so that to include the three types of poses, that is tilt, roll and yaw. Pitch or tilt is moving the head up and down. Tilting the head obliquely on both sides by facing the camera when the nose is stationary is called roll. Yaw is left and right side turning of head. From the thirty recorded videos, pose free face images are automatically sensed and images of mouth areas are excerpted in phase one as explained in section 3. Features are generated as explicated in section 4.1 from the mouth area of the video frames and utilized to train an auto associative neural network. Similar to first session, in the second session also thirty videos are generated from the same set of ten persons. Session one and session two has a time gap of twenty days to ensure a real life situation. No prior information is given to persons about the recordings during both the sessions for more real life situations. From the thirty recorded videos of session two, pose free images are automatically sensed and images of mouth areas are excerpted automatically and used for testing in the auto associative neural network for person authentication. Equal error rate is utilized to compare the efficiency of the person verification system when using normal face, smile expression and speech. The performance is considered better if the equal error rate is lower. The rate at which false acceptance rate and false rejection rate are equal is known as equal error rate, which is also called as crossover error rate. The percent of persons who are allowed wrongly as veritable persons at a specific threshold is called false match rate or false acceptance rate. The percent of veritable persons who are rejected falsely at a particular threshold is called false rejection rate or false non match rate. While using normal face, equal error rate for person authentication is 0.32% at threshold value 0.73. The corresponding graph is shown in Figure 7. While using smile facial expression, equal error rate for person authentication is 0.4% at threshold value 0.83. The corresponding graph is shown in Figure 8. While using visual speech, equal error rate for person authentication is 0.29% at threshold value 0.82. The corresponding graph is shown in Figure 9.

5. POTENTIAL OF AUTO ASSOCIATIVE NEURAL NETWORK

Auto associative neural network models are feed forward artificial neural networks. A feed forward artificial neural network has three types of layers, namely input layer, output layer and hidden layers. Layers are made up of processing units which are like artificial neurons. The connections between the units have associated weights with them. The information passes from input nodes to hidden nodes and then to output nodes and the interconnections do not form a directed cycle. The hidden layers may be one or more in number. By identity mapping the input space, artificial neural network models can be used to capture the distribution of the entered data [2]. The potential of the auto associative neural network model is explained in this section. Fig. 10(a) shows the auto associative neural network model employed in our work. This auto associative neural network model has five layers, out of which three are hidden layers. There are equal number of units in input layer and output layer. Among the three hidden layers, the middle layer is the compression layer which has less number of processing units than the number of units in the input layer or output layer. Hence input vectors get compressed to lower dimensions. The number of units in the other two hidden layers are more in number than the units in the input layer or output layer. The processing units in the hidden layers other than the compressed layers should be nonlinear. The middle hidden compressed layer processing units can be either linear or nonlinear. Error is minimized between the input vector and the expected output vector. To learn the dispersion of the feature vector of the mouth area, the structure of the neural network used in our work is 38L 76N 19N 76N 38L. In this structure, linear units are indicated with L. Nonlinear units are indicated with N. Number in the structure denotes the number of processing units in each layer. The cluster of points in the input space decides the form of the hyper surface got by the conversion onto a space of lower dimension. The area occupied by the single dimensional compression layer for the structure of network 2L 10N 1N 10N 2L is shown in Figure 10(c). In this network structure, same as above, linear units are indicated with L. Nonlinear units are indicated with N. Number in the structure denotes the number of processing units in each layer. Figure 10(b) shows the corresponding data. The auto associative neural network is trained using back propagation algorithm. The strong lines in Figure 10(c) denote the mapping of the input data due to the single dimensional compression layer. This shows that based on the conditions enforced by the network structure, auto associative neural network is able to capture the distribution of the data entered. To envision the distribution better, each input data point error have to be plotted as probability surface. For each data point i, the error e_i is plotted as $p_i = exp^{-(e_i/\alpha)}$, where α = constant. In spite of p_i not being a perfect probability density function, we consider the created surface as a probability surface. The probability surface exhibit more amplitude for smaller error e_i, which signals better fit of the network for that data point. The probability surface helps to learn the features of the input data distribution acquired by the auto associative neural network. The error surface form exhibits the constraints enforced by the neural network. The network is tuned to accomplish the probability surface with lower average error. The distribution capturing potential of this neural network is used in capturing the distribution of the mouth area feature vectors. For each mouth area feature vector, to lower the mean square error, back propagation algorithm is utilized to adapt the network weights.

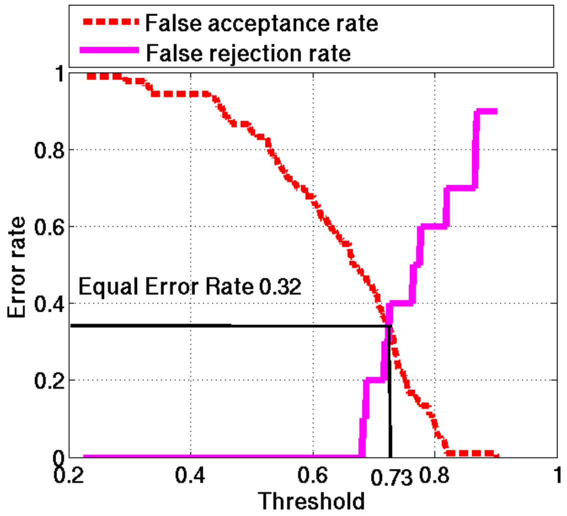


Fig 7: Equal error rate when using normal face

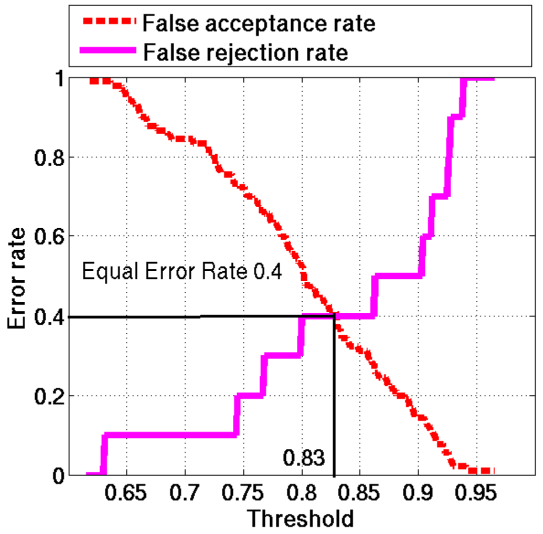


Fig 8: Equal error rate when using smile

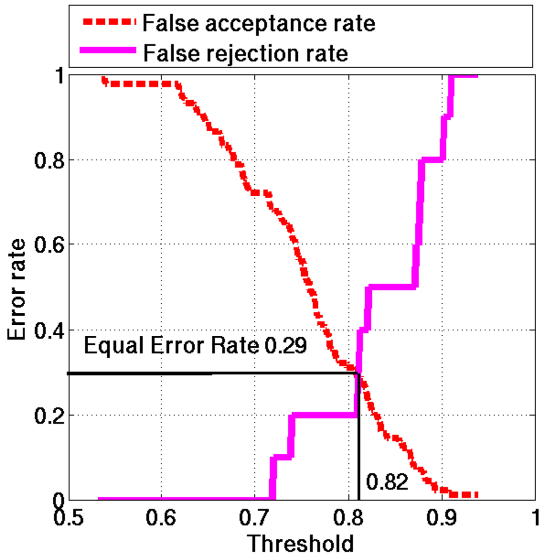


Fig 9: Equal error rate when using speech

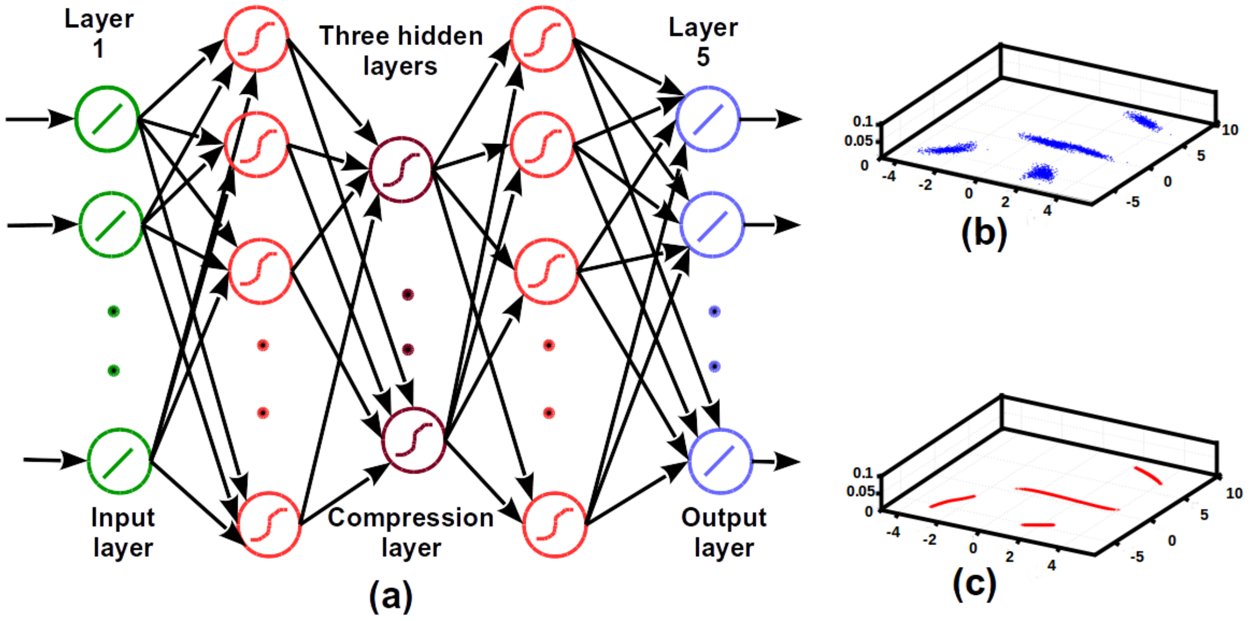


Fig. 10: Auto associative neural network

6. ANALYSIS OF RESULTS

The performance of the person authentication system is tested in sixty videos of resolution 640 x 480. The videos consist of ten persons at fifteen frames per second. They are recorded under normal room environment in two sittings. Care has been taken that the videos include poses like tilt, row and yaw. The lengths of the videos vary from one recording to another because recording is stopped based on the number of accepted video frames and not based on timing. In this work, the number of accepted video frames is assigned as thirty frames. The person authentication system uses for operation only the video frames which are accepted as images with single face without pose. But among the video frames removed as images with pose, there are some images which can be accepted as pose free images. Such wrong rejection may affect the fastness of the system a little, but it will boost the accuracy of the system which is highly essential. The videos are recorded without audio which helps reduction in size of the files. Background need not be bothered about, as the mouth area alone will be sensed automatically and used for feature excerption. The size of the face in the image depends on the fluctuation in the space between the camera and the face. But due to improved haar feature based cascade classifier, this fluctuation does not affect the performance in face detection. Mouth area features are excerpted from the collected pose free face images. Thirty video frames with a face in each without poses are selected automatically as explained in phase one and thirty eight mouth area feature vectors are excerpted from each video frame to train. Feature vectors are normalized for better performance under illumination variation. Normalized feature vectors are used in auto associative neural network to capture its distribution by training. The structure of this auto associative neural network is 38L 76N 19N 76N 38L. In auto associative neural network, when all the vectors, used for training are presented once, it is known as an epoch. The network is trained for 3000 epochs. The computer system used for this work has a Pentium Dual Core 2.3GHz processor. OpenCV 2.3 on Ubuntu 12.04 is used for coding. The time to train in this computer system takes less than a minute. Through experimental observation, the network structure perfect for the data is decided, when the confidence score is near one, when identical data is used for training and testing. For computing the normalized squared error (e), the output of the model is compared with the input. The normalized squared error for the feature vector y is given by, $e = \frac{||y-o||^2}{||y||^2}$, where o is the output vector given by the model. The confidence score (c) is calculated from error (e) using the formula c = exp(-e). The average confidence score is computed from the features of all the selected video frames. For authenticating a particular person the confidence score of testing data should be above a particular value, otherwise the person is considered fake. Instead of fixing the same threshold value for all the persons, each person can be fixed a unique threshold value as it is person verification. The performance is decided by using equal error rate. The equal error rate for person authentication is 0.32%, when using normal face, at threshold value 0.73, 0.4%, when using smile expression, at threshold value 0.83, and 0.29%, when using visual speech, at threshold value 0.82. The chart in Figure 11 helps to visualize the comparative effectiveness of the person authentication system when using normal face, smile, and speech in terms of equal error rate.

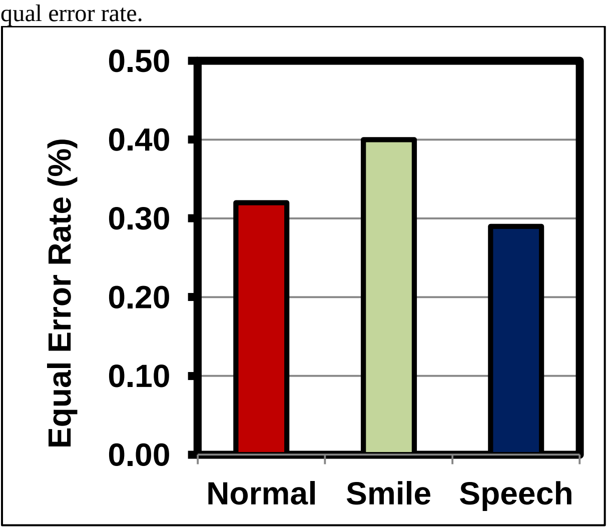


Fig. 11: Performance comparison

7. CONCLUSIONS

This work created procedure to excerpt from videos, video frames that has only one face and without poses without any need for manual intervention. Working without much deviation in the result under normal room conditions shows the tolerance of this system against illumination variation. This tolerance is mainly due to the creation of features based on differences in intensity. Features are created when the faces are not having any expression, face with smile and visual speech and compared using equal error rate as a measure of verification efficiency of the system. The equal error rate for person authentication is 0.32%, when using normal face, at threshold value 0.73. Equal error rate stands as 0.4%, when using smile expression, at a threshold of 0.83. It is 0.29%, when using visual speech, at threshold value 0.82. The results clearly pave way to conclude that efficiency is high when normal face is used instead of using smile facial expression during feature creation. Results also makes it clear that efficiency is still higher when visual speech is used instead of using normal face during feature creation. The reason may be because of similarity in the mouth area in between the speeches of the same person in different sessions. In other words, visual speech was more similar between sessions for a particular person. At the same time, compared with normal face, during visual speech, the mouth area of a particular person in a session shows enough variation within the mouth area to be supported well by neural networks. And this may be mainly as the features are intensity based. The future work is to compare the performance of visual speech based person authentication using Support Vector Machines and Radial Basis Function Neural Networks and also by varying the feature excerption methods. This work leads to establish a better person authentication system.

8. REFERENCES

[1] S. Saravanan, S. Palanivel, and M. Balasubramanian, "Facial Expression based Person Authentication", International Journal of Computer Applications, vol. 94, no. 13, pp. 1-8, May 2014.

[2] S. Palanivel, and B. Yegnanarayana, “Multimodal person authentication using speech, face and visual speech”, Computer Vision and Image Understanding, vol. 109, no. 1, pp. 44–55, Jan. 2008.

[3] M. Balasubramanian, S. Palanivel, and V. Ramalingam, “Real time face and mouth recognition using radial basis function neural networks”, Expert Systems with Applications, vol. 36, no. 3, pp. 6879-6888, Apr. 2009.

[4] Xudong Xie, and Kin-Man Lam, “Face recognition using elastic local reconstruction based on a single face image”, Pattern Recognition, vol. 41, no. 1, pp. 406-417, Jan. 2008.

[5] Roland Hu, and R.I. Damper, “Optimal weighting of bimodal biometric information with specific application to audio-visual person identification”, Information Fusion, vol. 10, no. 2, pp. 172-182, Apr. 2009.

[6] Federico Matta, and Jean-Luc Dugelay, “Person recognition using facial video information: A state of the art”, Journal of Visual Languages and Computing, vol. 20, no. 3, pp. 180-187, Jun. 2009.

[7] Meng Li, and Yiu-ming Cheung, “Automatic lip localization under face illumination with shadow consideration”, Signal Processing, vol. 89, no. 12, pp. 2425-2434, Dec. 2009.

[8] N. J. Nalini, S. Palanivel, and M. Balasubramanian, "Speech Emotion Recognition Using Residual Phase and MFCC Features", International Journal of Engineering and Technology, vol. 5, no. 6, pp. 4515-4527, Dec. 2013-Jan. 2014.

[9] N. Michael Brooke, “Using the visual component in automatic speech recognition”, Proceeding of Fourth International Conference on Spoken Language Processing ICSLP 1996, Philadelphia, PA, IEEE, Oct. 3-6, 1996, vol. 3, pp. 1656-1659.

[10] Engin Erzin, Yücel Yemez, and A. Murat Tekalp, “Multimodal Speaker Identification Using an Adaptive Classifier Cascade Based on Modality Reliability”, IEEE Transactions on Multimedia, vol. 7, no. 5, pp. 840-852, Oct. 2005.

[11] Kate Saenko, Karen Livescu, Michael Siracusa, Kevin Wilson, James Glass, and Trevor Darrell, “Visual Speech Recognition with Loosely Synchronized Feature Streams”, Tenth IEEE International Conference on Computer Vision ICCV 2005, Beijing, IEEE, Oct. 17-21, 2005, vol. 1, pp. 1424-1431.

[12] B. Goswami, C. Chan, J. Kittler, and W. Christmas, “Speaker authentication using video-based lip information”, IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP, Prague, IEEE, May 22-27, 2011, pp. 1908-1911.

[13] Wai Chee Yau, Hans Weghorn, and Dinesh Kant Kumar, “Visual Speech Recognition and Utterance Segmentation Based on Mouth Movement”, 9th Biennial Conference of the Australian Pattern Recognition Society on Digital Image Computing Techniques and Applications DICTA 2007, Glenelg, Australia, Dec. 3-5, 2007, pp. 7-14.

[14] H. E. Çetingül, E. Erzin, Y. Yemez, and A. M. Tekalp, “Multimodal speaker/speech recognition using lip motion, lip texture and audio”, Signal Processing, vol. 86, no. 12, pp. 3549–3558, Dec. 2006.

[15] Paul Viola, and Michael Jones, “Rapid object detection using a boosted cascade of simple features”, IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Hawaii, IEEE, Dec. 08-14, 2001, vol.1, pp. I-511-I-518.

[16] Rainer Lienhart, and Jochen Maydt, “An Extended Set of Haar-like Features for Rapid Object Detection”, Proceedings of the 2002 International Conference on Image Processing, USA, Sep. 22-25, 2002, vol. 1, pp. 900-903.

[17] M. Castrillon, O. Déniz, C. Guerra, and M. Hernández, “ENCARA2: Real-time detection of multiple faces at different resolutions in video streams”, Journal of Visual Communication and Image Representation, vol. 18, no. 2, pp. 130-140, Apr. 2007.

[18] Michal Uřičář, Vojtěch Franc, and Václav Hlaváč, “Facial Landmarks Detector Learned by the Structured Output SVM”, Proceedings of the 7th International Joint Conference on Computer Vision Theory and Applications, on Computer Graphics Theory and Applications and on Information Visualization Theory and Applications, Springer Berlin Heidelberg, Italy, Feb. 24-26, 2012, pp. 383-398