This report describes the demonstration of the Audio-Visual based person authentication system. The system records the audio-visual data and splits it into acoustic and visual streams. The spectral features are derived from the acoustic stream and are represented by Weighted Linear Prediction Cepstral Coefficients (WLPC). The system uses motion information to detect the face region in the visual stream, and the region is processed in YCrCb color space to determine the location of the eyes. The system extracts the gray level features relative to the location of the eyes. Autoassociative Neural Network (AANN) models are used to capture the distribution of the extracted acoustic and visual features. The proposed system can be used to recognize the identity of a test subject in addition to authentication. The performance of the system is invariant to size, and tilt of the face and is also not sensitive to variations in natural lighting conditions.

1. INTRODUCTION

Automatic authentication of a human subject is one of the challenging problems in human-computer interaction. The automatic person authentication system has many applications in the area of biometrics, information security, smart cards, access control and video surveillance. The objective of an audio-visual based person authentication system is to accept or reject the identity claim of a subject using acoustic and visual features. The proposed person authentication system consists of three main modules, namely acoustic feature extraction, visual feature extraction and Autoassociative Neural Network (AANN) model for authentication. The acoustic and visual feature extraction techniques are described in Section 2 and 3, respectively. Section 4 describes the AANN model for authentication. The enrollment and authentication procedure are described in Section 5.

2. ACOUSTIC FEATURE EXTRACTION

The acoustic feature extraction module extracts the vocal tract system features from the acoustic signal. The differentiated acoustic signal is analyzed by segmenting it into frames of 20 msec, using a Hamming window with a shift of 10 msec. The silence frames are removed using an amplitude threshold. A $10^k$ order Linear Prediction (LP) analysis is used to capture the properties of the signal spectrum. The recursive relation between the predictor coefficients and cepstral coefficients is used to convert the 10 LP coefficients into 19 cepstral coefficients. The LP cepstral coefficients for each frame is linearly weighted to form the Weighted Linear Prediction Cepstral Coefficients (WLPC). The distribution of the 19 dimension WLPC feature vectors in the feature space for a given subject is captured using an AANN model. Separate AANN model is used to capture the distribution of feature vectors of each subject.

3. VISUAL FEATURE EXTRACTION

The accumulated gray level difference image is used to estimate the face region in the image. The estimated face region is processed in YCrCb color space to determine the location of the eyes for each frame. The visual feature extraction module extracts 73 dimension gray level features relative to the location of the eyes, and hence the features are invariant to size and tilt of the face. The detailed description of the visual feature extraction is given in [1]. The 73 dimension gray level features are normalized to the range [-1,1]. The normalized feature vector is invariant to the image brightness. The distribution of the normalized feature vectors is captured using an AANN model.

4. AUTOASSOCIATIVE NEURAL NETWORK MODEL FOR AUTHENTICATION

Autoassociative neural network models are feedforward neural networks performing an identity mapping of the input space, and are used to capture the distribution of the input data [2]. The five layer Autoassociative neural network model as shown in Fig.1 is used to capture the distribution of the feature vectors [3].
The standard backpropagation equation is used to adjust the weights of the network to minimize the mean square error for each feature activation value of the unit. The nonlinear units used in that layer. The structures of the AANN models used in our system are 19L 38N 4N 38N 19L and 73L 90N 30N 90N 73L for capturing acoustic and visual features, respectively, where L denotes a linear unit, and N denotes a nonlinear unit. The integer value indicates the number of units used in that layer. The nonlinear units use $\tanh(s)$ as the activation function, where $s$ is the activation value of the unit. The standard backpropagation learning algorithm is used to adjust the weights of the network to minimize the mean square error for each feature vector.

5. Enrollment and Authentication

For enrolling a subject, an AVI file of 60 sec duration at 15 fps is recorded using a camera with a resolution of $160 \times 120$. The system records the audio signal at 11025 samples per second. The system splits the AVI stream into acoustic and visual streams. The acoustic and visual features are extracted from the streams as described in Section 2 and 3 respectively. The extracted acoustic feature vectors are given as input to the AANN model 19L 38N 4N 38N 19L, and the network is trained for 60 epochs as described in Section 4. Similarly, the extracted normalized visual feature vectors are given as input to the AANN model 73L 90N 30N 90N 73L, and the network is trained for 200 epochs.

For authenticating the identity claim of a subject, an AVI file of 15 sec duration at 15 fps is recorded, and the streams are separated. The 73 dimension normalized feature vector is extracted from each video frame, and is given as input to the corresponding model. The output of the model is compared with the input to compute the square error. The error ($e$) is transformed into a confidence value ($c$) by using the equation $c = \exp(-e)$. Similarly, the acoustic feature vectors corresponding to each video frame are extracted from the acoustic signal, and the average confidence value is estimated. The confidence values from the models are combined using a weighting rule and the process is repeated for each video frame. The average weighted confidence value is used to decide the identity claim of the test subject.

The system is implemented in VC++ and the enrollment and authentication of a subject requires approximately 3 minutes and 30 sec, respectively, on a Pentium machine at 2.53 GHz.

The proposed Audio-Visual based person authentication system has the following salient features:

1. Invariant to size and tilt of the face
2. Not sensitive to variations in natural lighting conditions
3. Able to authenticate a subject within a reasonable time
4. New subject can be added to the system without using the features of other subjects
5. No prior information of the subject (such as skin patches) is required.

The proposed person authentication system can be used as a recognition system. The objective of the recognition system is to determine the identity of a test subject from the set of reference subjects. For recognizing the identity of a test subject, the acoustic and visual feature vectors are extracted from the AVI file and is given as input to all the acoustic and visual models respectively. The weighted confidence value is computed for all the enrolled subjects. The identity of the test subject is determined from the largest weighted confidence value.

6. References

