Audio-Visual Authentication System over the Internet Protocol

Yee Wan Wong, Kah Phooi Seng, and Li-Minn Ang

Abstract—**In this paper, an audio-visual (AV) authentication system is developed with the objective to increase the robustness of the system to face illumination variation. A multiband feature fusion approach is proposed to search for the mid- and high-frequency subbands that are insensitive to variation in illumination based on wavelet packet decomposition to solve the illumination problem. Simulation results show that the multiband feature fusion approach achieved higher recognition accuracy as compared to a previous study. To further improve the robustness of multiband feature fusion approach to not only invariant to face illumination but also invariant to facial expression variation, principle component analysis is employed to work in conjunction with the multiband feature fusion approach. Then the AV authentication system is implemented over internet protocol (IP) to enable long distance access. In this system, we are concerned about video and audio streaming. Hence, the effects of speech and face compression on recognition performance of AV authentication system over the internet protocol are investigated. The experiment results show that the AV authentication system over IP with smaller data size achieved the same recognition rate as in the standalone system.**

Index Terms— **Audio-visual authentication system, wavelet packet transform, internet protocol, real-time video.**

I. Introduction

Audio-visual (AV) authentication system is an automatic system that verifies a person identity using acoustic and visual biometric signals such as the person's voice and face. The AV authentication system is easier to be socially accepted because of its unobtrusive and user-friendly procedures and low-cost sensors [1]. However, the audio and visual data are easily influenced by uncontrolled factors especially visual data are sensitive to variation in illumination. The variation in illumination of the visual data degrades the recognition accuracy of the AV authentication system.

Some of the well-known methods in solving illumination problems were proposed [2], [3]. In Ekenel and Sankur paper [4], they searched for the subbands that are insensitive to the variation in illumination by using wavelet transform to decompose the original image to three-level decomposition. They found that the mid-range frequency subband containing horizontal details is successful against variations in illumination. However, the high-frequency subbands that are less affected by illumination [5], [6] are abandoned.

In this paper, we proposed a multiband feature fusion approach to search for the mid- and high-frequency subbands that are insensitive to variation in illumination based on wavelet packet transform (WPT) [7] to solve the illumination problem. In our approach, high frequency subbands are also being further decomposed. From these subbands, subbands that do not contain illumination-based frequency components are chosen to better represent the face image. We use a statistical method and recognition accuracy to test the recognition performance of the subbands. Simulation results show that the multiband feature fusion approach achieves higher recognition rate as compared to the recognition rate of Ekenel and Sankur paper. To further improve the robustness of the multiband feature fusion approach to not only invariant to face illumination but also to facial expression variation, principle component analysis (PCA) [2] is employed to work in conjunction with the proposed approach.

After developing the AV authentication system that is robust to illumination variation and facial expression variation, the AV authentication system is implemented over internet protocol (IP) to enable long distance access. The architecture of the AV authentication system will be a client-server based system where a distant person recognition server is remotely accessed by the many network-connected clients. In this system, we are concerned about video and audio streaming. Due to its real-time nature, video and audio streaming typically has bandwidth, delay and loss requirements. However, there is no quality of service (QoS) guarantee for video and audio transmission over the current internet [8]. Thus, transmitting video and audio files that are usually big in size over the internet poses many challenges to researchers [8]. To address these challenges, some researchers conducted investigation on compression factors to the recognition accuracy to face recognition system [9] and speaker recognition system [10] over the internet. These recognition systems referred to the implementation of face and speaker recognition system over the IP. In this paper, an AV authentication system over IP is presented. The effects of speech and face compression on recognition performance on AV authentication system over the internet are investigated. Although data compression is essential for application over internet, the recognition accuracy of the AV authentication system is also very important. As a result, the video and audio data sizes are effectively selected to fulfill the trade-off between the size and recognition accuracy of the AV authentication system. The experiment results show that the AV authentication system over IP with smaller data size achieved the same recognition rate as in the standalone system.

[1] Yee Wan Wong, Kah-Phooi Seng and Li-Minn Ang are with the University of Nottingham Malaysia Campus, Jalan Broga, 43500 Semenyih, Selangor, Malaysia (corresponding author to provide phone: +60389248358; fax: +60389248017; e-mail: yeewan.wong@Nottingham.edu.my).

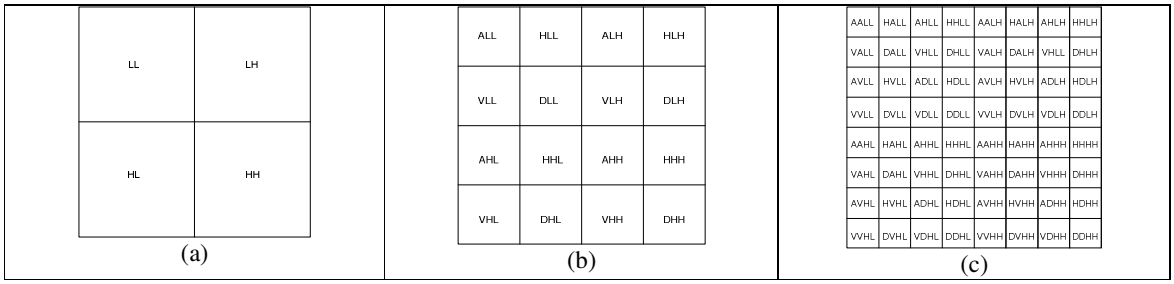

Figure 1 Frequency subbands of (a) 1-level (b) 2-level (c) 3-level

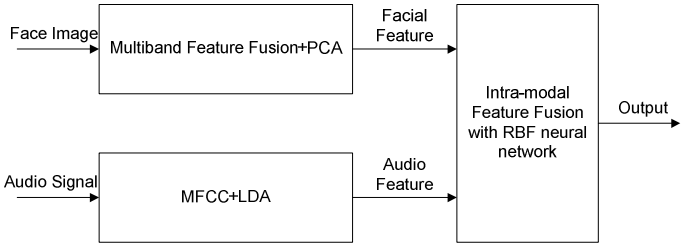

Figure 2 Block diagram of the proposed AV authentication system

II. Proposed AV Authentication System with Multiband Feature Fusion approach

Recognition accuracy of the AV authentication system is easily influenced by uncontrolled factors such as the variation in illumination in face image. In this section, we proposed a multiband feature approach based on wavelet packet transform (WPT) [7] to search for the mid- and high-frequency subbands that are insensitive to face illumination. In our approach, we used wavelet packet decomposition because it provides a completely evenly spaced frequency resolution and it allows a finer and adjustable resolution of frequencies at high frequencies.

In the proposed method, images are decomposed into subbands by wavelet packet decomposition until level-3 decomposition as shown in fig. 1 (c). Each of the subbands is named as shown in fig. 1 and their recognition accuracies are tested. Recognition accuracy and class separation are used to evaluate the recognition performance of the subbands. Recognition accuracy shows how well the system can match images from the same people and class separation shows how well the system can distinguish images from different people [11]. To test the class separation, N face images from a database, one image per person are used. The face images are chosen randomly from the training and testing sets. The similarity matrix $\rho(i, j)$ with the size NxN records the similarity between image i and image j. For a good representation, $\rho(i, j)$ should be closed to one if $i = j$ and $\rho(i, j)$ should be close to zero if $i \neq j$. The Average Unmatched Similarity Value, AUMSV [11] that is defined as below,

$$AUMSV = \frac{1}{(N^2 - N)} \sum_{i=1}^{N} \sum_{j=1}^{N} \rho(i, j) \qquad i \neq j \qquad (1)$$

AUMSV is used to give a single numerical value to the similarity performance of the subband. This term shows how well the subband representation distinguishes the images from different people, and it ranges from 0 to 1, which means the higher the discriminatory power, the smaller the AUMSV value.

Below are the steps proposed to select the optimal subbands:

Step 1: Compute the AUMSV and recognition accuracy in each subbands from level 1 and 2 as shown in fig. 1 (a) and (b).

Step 2: The subbands that obtain AUMSV and recognition accuracy that is lower and higher than threshold values respectively will be selected for further decomposition to level 3. The threshold values are chosen according to the AUMSV and recognition accuracy obtained at level 1 and 2 decompositions. The threshold values determine the computational complexity of the system. This step reduces the computational complexity by avoiding decomposition of all subbands from level 2 to level 3.

Step 3: Further decompose subbands that fulfil the subband selection criteria to level 3 decomposition.

Step 4: Two best performing subbands in terms of AUMSV in level-3 decomposition will be concatenated and the optimal feature set is named as Optimal Multiband Feature (OMF).

We then further improve the robustness of multiband feature fusion approach to not only invariant to face illumination but also invariant to facial expression variation. The PCA is proposed to work in conjunction with multiband feature fusion approach. In [2], the authors showed that when the first three principle components have been removed (PCA w/o 3), the recognition performance of the eigenface (PCA) method was improved in the presence of lighting variation and facial expression. However, there are two problems in this method: 1) it does not achieve recognition rate as high as other methods tested 2) high computational is required if the dimensionality of the original image is large. Hence, we propose to make use of the advantage of PCA without the first three components and also the OMF to improve extract facial features that are robust to face illumination variation, facial expression variation. The PCA is applied to the low-frequency DWT

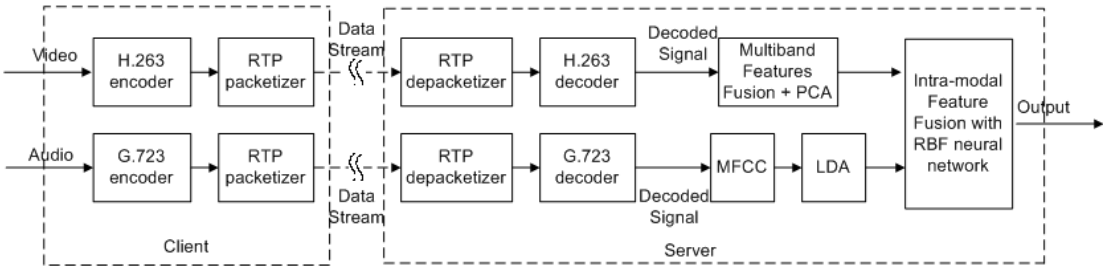

Figure 3 Block diagram of the proposed AV authentication system over IP

subband instead of to the original image to solve the high computational problem of PCA and then the principle component coefficients will be feature fused with OMF. As a consequence, OMF+PCA is expected to achieve high recognition rate in the presence of illumination and other facial variations.

Fig. 2 shows the block diagram of the proposed AV authentication system. The facial features are extracted by the multiband feature fusion with PCA approach whereas the audio features are extracted by mel-frequency cepstrum coefficient (MFCC) [12] and linear discriminant analysis (LDA) [13]. The intra-modal feature fusion proposed in [18] is employed to combine both the facial and audio features in the proposed system.

III. Proposed AV Authentication System Over Internet

Fig. 3 depicts architecture of video and audio streaming over network for proposed AV authentication system. In this figure, the client streams the captured video over the network to the server for recognition. The video and audio are first compressed by video and audio encoding. Upon the client's request, a streaming server retrieves compressed video and audio data from the client. The transport protocols packetize the compressed bit-streams and send the video and audio packets to the network. Packets may be dropped or experience delay inside the network depending on the network congestion. For packets that are delivered to the server successfully, they are passed through the transport protocols and being depacketized to bit-streams before being decoded at the video and audio decoder. The received video and audio data will be used for AV authentication purpose. There are three areas that are important to the video and audio streaming architecture. The three areas will be briefly described as follows.

1) *Video and audio encoder/decoder*: Raw video and audio must be compressed using video and audio encoding schemes before transmission to achieve efficiency. The ITU-T H.323 standard for audio-visual communication systems that has been widely used across the internet is adopted in our application [14], [15]. For video codec, H.263 that is able to achieve lower bit-rate than H.261 is selected. The H. 263 allows five standardized picture formats. These are CIF (common intermediate format), QCIF (quarter CIF), SQCIF (sub-CIF), 4CIF and 16CIF. The H.263 standard uses the discrete cosine transform (DCT) to remove spatial redundancy and motion estimation and compensation to remove temporal redundancy. For audio codec, G.723 with bit-rate of 8kbit/s and 16kbit/s that usually used for multimedia communication is selected.

2) *Protocols*: Protocols are designed and standardized for communication between clients and servers [8]. The protocols can be categorized as network protocol and transport protocol. The network-layer protocol such as IP provides basic network service support such as network addressing. The transport protocol such as user datagram protocol (UDP), transmission control protocol (TCP) and real-time transport protocol (RTP) provide end-to-end network transport functions for streaming applications. UDP and TCP support multiplexing, error control and congestion control. However, unlike UDP, TCP uses retransmission to recover lost packets. Since TCP retransmission introduces delays that are not acceptable for streaming application with stringent delay requirement [8], UDP is used as the transport protocol for video streaming in the AV recognition system. The RTP that is designed to support real-time applications [16] is employed as the upper-layer transport protocols.

3) *Packetizer*: When transmitting H.263 video streams, the output of the encoder can be packetized directly. For every video frames, the H.263 bit-streams is carried in the RTP payload without alteration. Therefore, multiplexing audio and video signals in the same packet is not accommodated [17]. In other words, the audio and video signals must be demultiplexed and sent separately. An RTP packet can use one of the three modes for H.263 video streams depending on the desired network packet size and H.263 encoding options employed [17]. For each RTP packet, the RTP fixed header is followed by the H.263 payload header, which is followed by the standard H.263 compressed stream [17]. The shortest H.263 payload header (mode A, four bytes) supports fragmentation Group of Block (GOP) boundaries. The long H.263 payload headers (mode B, eight bytes and C, twelve bytes) support fragmentation at Macroblock (MB) boundaries. Due to the simplicity of mode A, it is used as the H.263 payload header in our applications.

Fig. 3 shows the block diagram of the proposed AV recognition system over IP network. At the client side, the raw video and audio signals will be first compressed by H.263 and G.723 encoder respectively. The bit-stream will be packetized and sent over the internet by RTP. At the server side, the received packets will be depacketized and passed to the audio and video decoder. The decoded image frames and audio signal will be passed to the proposed multiband feature fusion method with PCA for facial feature extraction and the MFCC and LDA for the audio feature extraction method respectively. Both the features will be

fused by the intra-modal feature fusion method proposed in [18].

IV. Experimental Results

There are five parts of experiment in this section. The first two parts present the recognition accuracy of the OMF and OMF+PCA. The nearest neighbour classifier was used for data classification in these two parts of the experiment. The following experiment consists of three parts: audio over IP, video over IP and AV over IP.

A. Multiband Feature Fusion Approach

The first part of this section shows the experimental results for the proposed multiband feature fusion method. The YaleB database [23], total 152 cropped faces were first used in finding the OMF. The YaleB database contains illumination variations in the images that occur due to intensity and direction of the light. All images are scaled to 128x128 pixels resolution. For each individual in the set, two of their images that contain frontal illumination with different amounts of light are used for training, and the remaining two images that contain illumination from sides are used for testing. Sample face images are shown in fig. 4. To test AUMSV, N=38.

Experiments on the DWT level 1 and 2 were first carried out. As refer to fig 1a and b. The threshold value for AUMSV was 0.5 and the threshold value for recognition

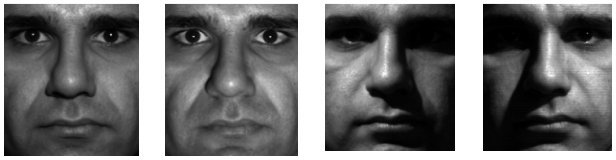

Figure 4 Sample face images containing variations in illumination from YaleB

rate was 50% of the recognition rate of the original image (44.74%) which was 22.37%. Subbands that were further decomposed to level 3 must obtain AUMSV that was lower than 0.5 and recognition rate that was higher than 22.37%. The selected subbands were ALL, HLL, VLL, DLL, ALH, HLH, AHL and HHL. It is interesting to note that high-frequency subband ALH achieved the lowest AUMSV of 0.248 and the highest recognition rate of 73.68% which yields a significant improvement of 30% as compared to the low-frequency subband LL.

The five level-3 decomposition subbands that obtained the lowest AUMSV were shown in Table I. It is shown that high-frequency subband AALH achieved the lowest AUMSV of 0.241 and the highest recognition rate of 76.32%. It shows that there were some redundant information in the ALH (32x32) subband because in AALH (16x16), the AUMSV was reduced and the recognition rate increased by 3% even with a much smaller subband. From this result, it is important to take note that high-frequency subband AALH achieved a much higher recognition rate as compared to the HALL which was found to attain the highest recognition rate in Ekenel and Sankur [4] paper. This proves that the high-frequency component is more discriminative than the low and mid-frequency components in face illumination.

Next, we carried out subband feature fusion. The subbands involved in the fusion were the components shown in Table I. Then, we compared the recognition accuracy of the best performing feature fusion subbands (named as OMF) which concatenated AALH and HALL with LL and HALL (Ekenel and Sankur) and the results were shown in Table II. We found that the highest recognition rate of 81.58% was achieved by the OMF. The improvement was around 37% as compared to HALL. Fig. 5 shows the location of HALL and AALH in frequency subband.

The OMF was then tested on two more face databases, which were ORL [19] and CUAVE [20]. The ORL database consists of 40 classes of total 400 images. For some subjects, the images were taken at different times, which contain quite a high degree of variability in facial expression, pose and facial details. The 400 images were divided by 2 to make 200 for training data set and 200 for testing data set in this experiment. The CUAVE face database was obtained from the CUAVE AV database. The CUAVE database includes subjects that involve movements such as nodding the head in different directions, moving back-and-forth and side-to-side in the field of view, and in some cases rotation of the head. The database consists of 36 classes. Three samples were chosen from each class as training samples and on the other hand, three samples were chosen from each class as testing samples. Table III shows that the OMF achieved high recognition rate on YaleB database that contained illumination variation factor. However, OMF achieved a considerable low recognition rate on databases that did not contain illumination variation factor (ORL and CUAVE).

B. Multiband Feature Fusion Approach with PCA

In this section, OMF+PCA is proposed to increase the face recognition robustness to not only face illumination variation, but also to facial expression variation and minor head pose variation. Table IV shows the recognition rate of the proposed OMF+PCA with other facial feature extraction methods tested. Feature extraction methods used were OMF, PCA, PCA w/o 3 [2], and independent component analysis (ICA) [21]. In ORL database which contains no face illumination variation, OMF+PCA achieved comparable recognition accuracy as in PCA. In YaleB database which contains severe illumination variation, OMF achieved the highest recognition rate followed by the OMF+PCA. In CUAVE database which contains no face illumination variation but with moving subject, OMF+PCA achieved the highest recognition rates. From the experiment, we can see that OMF+PCA achieved good recognition performance consistently in all three databases.

Table I AUMSV and correct recognition rate of decomposition sample subbands

Best performing Subband	AUMSV	Recognition rate (%)
AALH	0.241	76.32
HALL	0.271	44.74
HHLL	0.382	39.47
VALL	0.389	31.58
VHLH	0.399	23.68

Table II Correct recognition rate of methods

Method	Recognition rate (%)
OMF	81.58
LL	43.42
HALL (Ekenel and Sankur)	44.74

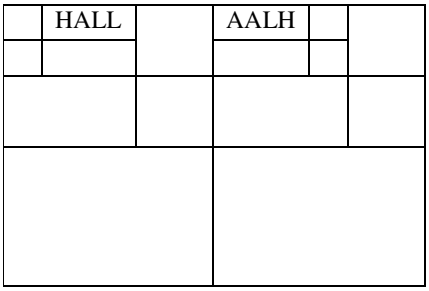

Figure 5 Location of the HALL and AALH in frequency subband

Table III Correct recognition rate of the OMF in three different databases

Database	Recognition rate (%)
ORL	64
YaleB	82
CUAVE	74

Table IV Correct recognition rate (%) of methods in the three databases

Database	OMF+ PCA	OMF	PCA	PCA w/o 3	ICA
ORL	89	64	90	85	73
YaleB	71	82	30	47	39
CUAVE	82	74	80	77	80

C. Audio over IP

For all parts of the experiment, CUAVE AV database was used. The database consists of 36 speakers. It was recorded in an isolated sound booth at a resolution of 720x480 with NTSC standard of 29.97fps. The data was then compressed into individual MPEG2 files for each speaker. The MPEG2 files were encoded at a data-rate of 5000kbps with multiplexed 16-bit, stereo audio at 44 kHz sampling rate. The network configuration that we used was the peer-to-peer network configuration with link speed of 100Mbps. JMstudio was used to transmit and receive the data [22].

In this part of the experiment, we evaluated the influence of speech compression and speech quality over IP on speaker recognition performance. At the client side, the "wav" format audio files were compressed by the audio codec G.723 to bit rate of 8kbit/s and 16kbit/s. These data were then streamed to the server side for recognition performance evaluations. MFCC and LDA were used as the feature extraction method for the audio data and RBF neural network was used as the classifier in this experiment. Table V shows that the speaker recognition performance was affected when the speech material was encoded at low bit rate. It also shows that the speaker recognition performance was less affected when the speech with 16kbit/s was streamed over IP. This is because peer-to-peer network configuration offers enough transfer speed for the data.

D. Video over IP

In this part of the experiment, we evaluated the influence of the video dimension and image quality over IP on face recognition performance. At the client side, the video files were first being encoded to three different video dimensions (Mode A): SQCIF (128x96), QCIF (176x144) and CIF (352x288) and then the files were streamed to the server side for recognition performance evaluations. Fig. 6 shows some sample images for the transcoded data at different sizes at the server side. The proposed OMF+PCA was used as the feature extraction method and the RBF neural network was used as the classifier. Table VI shows that CIF over IP degraded the recognition accuracy by 4% as compared to the recognition accuracy of the standalone system.

E. Audio-Visual over IP

In this part of the experiment, we evaluated the effect of audio and visual data over IP on AV recognition performance. G.723 with 16kbit/s was selected as the codec in this part of the experiment due to its high recognition accuracy as shown in Table V. Table VII shows that the recognition accuracy increased when video and audio data were combined as compared to individual experts for standalone system and AV authentication system over IP. Besides, lower dimensional video over IP achieved same recognition accuracy as compared to the recognition accuracy of the high dimensional original video in standalone system. While streaming the audio and visual data, the sent bytes per second were observed. For CIF, QCIF and SQCIF, the sent bytes per second were 80 kbps, 50 kbps and 40 kbps respectively. It is important to note that different video dimensions can be used according to different bandwidth requirement to avoid network congestion. For example, QCIF is recommended to be used for AV streaming because it achieves the same recognition accuracy as in CIF over IP and standalone system but with much lower bit rate.

Table V Speaker recognition results for standalone system without going through IP and transcoded data over IP

Audio bit-rate	Recognition rate (%)
Standalone (16kbit/s)	86
Over IP G.723 (16kbit/s)	85
Over IP G.723 (8kbit/s)	64

(a) (b) (c) (d)

Figure 6 Sample images for (a) standalone system with original size (b) transcoded image over IP for CIF (c) transcoded image over IP for QCIF (d) transcoded image over IP for SQCIF

Table VI Face recognition results for standalone system without going through IP and transcoded data over IP

Video Dimension	Recognition rate (%)
Standalone (720x480)	89
Over IP CIF(352x288)	86
Over IP QCIF(176x144)	61
Over IP SQCIF(128x96)	72

Table VII AV recognition results for standalone system without going through IP and transcoded data over IP

Video Dimension	Audio bit-rate	Recognition rate (%)
Standalone	Standalone(16kbit/s)	94
Over IP CIF	Over IP (16kbit/s)	94
Over IP QCIF	Over IP (16kbit/s)	94
Over IP SQCIF	Over IP (16kbit/s)	89

V. Conclusion

In this paper, the multiband feature fusion method was first proposed to search for the mid- and high-frequency subbands that were insensitive to variation in illumination based on WPT. The selective subbands were feature fused and this resulting best feature set was named to be OMF. Experiment results showed that OMF achieved higher recognition rate as compared to the recognition rate of a previous study. OMF+PCA was then proposed to increase the robustness of the face recognition system to large facial variations. The OMF+PCA achieved good recognition performance consistently in all three databases tested. The OMF+PCA was employed in the proposed AV authentication system and the proposed system was implemented over IP. The effects of speech and face compression on recognition performance on speaker recognition system over IP, face recognition system over IP and AV authentication system over IP were investigated. The experiment result showed that low bit-rate speech compression and lower dimensional video degraded the recognition performance in speaker and face recognition system. However, after combining both the audio and visual data, the AV authentication system achieved higher recognition rate as compared to the recognition rate of face and speaker recognition system. Simulation results also showed that with the combination of audio signal and lower dimensional video (QCIF), the AV authentication system achieved the same recognition rate as in high dimensional video over IP and high dimensional video in standalone system. Hence, QCIF that accounted for lower bandwidth requirement and achieved high recognition rate was recommended to be used in the proposed AV authentication system over IP to reduce network congestion.

References

[1] P.S. Aleksic and A.K. Katsaggelos, “Audio-visual Biometrics”, in Proc. Of the IEEE, vol. 94, no. 11, pp. 2025-2044, Nov. 2006.

[2] P.N. Belhumeur, J.P. Hespanha, D.J. Kriegman, “ Eigenfaces vs. fisherfaces: recognition using class specific linear projection”, IEEE Trans. Patt. Anal. Mach. Intell., Vol.19, no.7, pp. 711–720, 1997.

[3] M. J. Er, W. Chen, S.Wu, J. L, “High-speed face recognition based on discrete cosine transform and RBF neural network,” IEEE Trans. Neural Netw., vol. 16, no. 3, pp. 679–691, May 2005.

[4] H. K. Ekenel and B. Sankur, “Multiresolution face recognition”, Image and Vision Computing, vol. 23, pp. 469-477, 2005.

[5] C.Naster, B. Moghaddam, A. Pentland, Flexible images: matching and recognition using learned deformations, Comput. Vision Image Understanding 65(2) (1997) pp. 179-191.

[6] C. Naster, N. Ayache, Frequency-based non-rigid motion analysis, IEEE Trans. Pattern Anal. Mach. Intell. 18 (11) (1996) pp. 1067-1079.

[7] T. Acharya and P.S Tsai, JPEG2000 Standard for Image Compression, Wiley-interscience. pp.79-91, 2005.

[8] D. Wu, Y.T. Hou, W. Zhu, Y.-Q. Zhang and J. M. Peha, “Streaming video over internet: Approaches and Directions,” IEEE Trans. Circuits Syst. Video Technol, vol. 11, no. 3, pp. 282-300, March. 2001.

[9] L. Besacier, P. Mayorga, J.-F. Bonastre, C. Fredouille and S. Meignier, “Overview of compression and packet loss effect in speech biometric”, IEE Proc.-Vis. Image signal process, vol. 150, Dec. 2003, pp. 371-376.

[10] H. Song, S. J. Chung, and Y.-H. Park, “An online face recognition system using multiple compressed images over the internet”, LNCS, 2005, pp. 569-576.

[11] Feng, G. C., Yuen, P.C. & Dai, D. Q., “Human face recognition using PCA on wavelet subband”, Journal of Electronic Imaging, Vol. 9, No. 2, April 2000, pp. 226-233.

[12] J.P. Campbell, JR, “Speaker recognition: a tutorial”, Proc. Of IEEE, vol. 85, no. 9, no.9, pp. 1437-1462, Sept. 1997.

[13] X. Lu, Image Analysis for Face Recognition, personal notes, May 2003, 36 pages.

[14] W. Ding and B. Liu, “Rate Control of MPEG video coding and recording by rate-quantization modeling,” IEEE Trans. Circuits Syst. Video Technol., vol. 6, pp. 12-20, Fec. 1996.

[15] T. Weigand, M. Lightstone, D. Mukherjee, T. G. Campbell. and S. K. Mitra, “Rate-distortion optimized mode selection for very low bit-rate video coding and the emerging H.263 standard,” IEEE Trans. Circuits Syst. Video Technol., vol. 6, pp. 182-190, 1996.

[16] H. Schulzrinne, S. Casner, R. Frederick, and V. Jacobson, “RTP:A transport protocol for real-time applications,” Internet Engineering Task Force, RFC 1889, Jan. 1996.

[17] C. Zhu, “RTP Payload Format for H.263 Video Streams,” Intel Corp. Sept 1997.

[18] Y. W. Wong, K. P. Seng, L.-M. Ang, W. Y. Khor, H. F. Liau, “Audio-Visual Recognition System with Intra-Modal Fusion”, *2007 International Conference on Computational Intelligence and Security*, pp. 609-613, Dec 2007.

[19] http://www.cam-orl.co.uk/face-database.html.

[20] E.K. Patterson, S. Gurbuz, Z. Tufekci, and J.N. Gowdy, “CUAVE: A NEW AUDIO-VISUAL DATABASE FOR MULTIMODAL HUMAN-COMPUTER INTERFACE.

[21] Bartlett, M.S. Movellan, J.R. Sejnowski, T.J. Face recognition by independent component analysis,” IEEE Transactions on Neural Networks, ,vol. 13, pp. 1450- 1464, Nov 2002.

[22] JMstudio: http://java.sun.com/javase/technologies/desktop/media/jmf/2.1.1/samples/samplecode.html

[23] Georghiades, A. S., Belhumeur, P. N., and Jacobs, D. W. (2001). From few to many: illumination cone models for face recognition under variable lighting and pose. *IEEE Trans. Pattern Anal. Mach. Intel.*, 23(6), 630-660.