Audio-Visual Recognition System Insusceptible to Illumination Variation over Internet Protocol

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

Abstract— **In this paper, we present an audio-visual recognition system which is insusceptible to illumination variation over internet protocol. First, the multiband feature fusion method is proposed for face recognition under varying illumination. The wavelet packet transform decomposes an image into various frequency subbands. We show how to select a set of subbands that are invariant to illumination variations by using statistical method and modified Euclidean based method. More specifically, there exist a set of wavelet subbands that carry facial features which provide an effective representation for face recognition under wide range of lighting conditions. Histogram equalization is then applied on these subbands to enhance the contrast of the features. The recognition performance of the proposed method is validated on some standard data sets and high recognition accuracy is achieved. Then the audio-visual recognition system over internet protocol is developed. The compression and packet loss effects of sending the audio and video data over internet protocol on recognition performance are investigated.**

Index Terms— **Audio-Visual recognition system, wavelet packet transform, illumination invariant subband, internet protocol.**

I. Introduction

Audio-Visual (AV) recognition system is an automatic system that recognizes a person's identity using audio and visual data. However, these data are easily influenced by acoustic noise. With the combination of visual and audio data, the recognition performance of the AV recognition system is improved even in acoustic noisy environment [1], [2]. Nevertheless, the recognition performance of the AV recognition system is degraded by face illumination variation.

Research efforts to solve the illumination problem can be grouped into three streams: subspace methods [3], [4], Lambertian reflectance model [5-10], and normalization [12-14]. The first approach includes the popular face recognition methods which are the principle component analysis (PCA) [3] and the linear discriminant analysis (LDA) [4]. These approaches are highly sensitive to illumination variation. To improve the recognition performance of PCA under illumination variation conditions, [15] proposed to remove the first three eigenvectors to reduce the illumination factor in the subspace. The LDA has also been modified to handle illumination variations [16]. The second approach is based on the Lambertian reflectance model with varying albedo field. Under the Lambertian assumption, the illumination cone [5], harmonic images [6], and quotient image based methods [8-10] are developed to solve the illumination problem. The third approach preprocesses images to appear stable under illumination variations. This approach removes illumination variations while keeping the main facial features unimpaired. For example, histogram equalization based methods [12], Gamma correction, and logarithm based methods [14].

Wavelet transform has been used in face recognition system [17], [18], [26]. Some methods based on wavelet transform [19-21] are developed to solve illumination problem. In Ekenel and Sankur paper [19], they search for the subbands that are insensitive to the variation in illumination by using wavelet transform. They find that the mid-range frequency subband (HALL) is successful against variations in illumination. However, the high-frequency subbands that are less affected by illumination [22], [23] are abandoned. Du and Ward [20] performs illumination normalization in wavelet domain. Histogram equalization is applied to the approximation subband and simple amplification is applied to the detail subbands. Image reconstruction is then performed on these modified subbands. More recently, Zhang et al. [21] proposed a wavelet-based method to estimate the illuminance in logarithm domain and then extract the invariant facial features. The parameter selection making this method difficult to be applied because it depends on the illumination conditions of the training and probe images.

In this paper, we propose the audio-visual recognition system which is insusceptible to illumination variation over internet protocol. First, the multiband feature fusion method is proposed to extract the illumination invariant facial features directly from the image. The wavelet packet transform (WPT) [25] is used to decompose image into various frequency subbands. Unlike the method proposed in [19], our proposed method searches for the invariant features from not only low-frequency subband, but also high-frequency subbands. We use a statistical method [26] and a modified Euclidean distance based method to choose the frequency subbands. The selected subbands form the Optimal Multiband Feature (OMF) and the histogram equalizer (HEQ) is applied on to the selected subbands to enhance the contrast of the features. Then the audio-visual (AV) recognition system is developed. Figure 1 depicts the block diagram of the AV recognition system. The face image is first being decomposed by WPT and the features in the optimal multiband (OMF) are extracted and then the contrast of OMF is enhanced by the HEQ. Mel-frequency cepstrum coefficient (MFCC) [27] is used to extract the audio features and the LDA is applied as a feature selection technique. Intra-modal feature fusion [28] combines both the audio and facial features and radial basis function (RBF) neural network [29] performs classification. The AV recognition system is then implemented over internet protocol (IP). The recognition performance of the OMF is tested on

[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).

(Advance online publication: 22 May 2009)

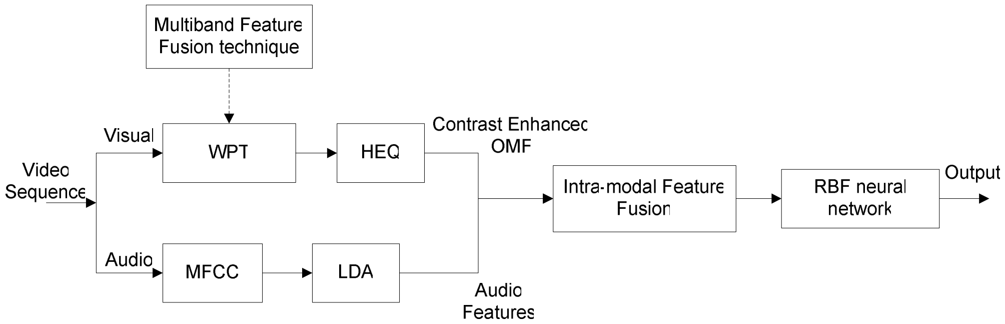

Fig. 1. Block diagram of the proposed AV recognition system

Extended YaleB database [7], YaleB database [30], and AR database [31]. Experimental results show that the proposed method attains high recognition rate. The effects of compression and streaming of audio and video files over different link capacities on recognition performance of AV recognition system are also investigated.

The paper is organized as follows. In section II, the multiband feature fusion method is explained. In section III, the important components for the implementation of the developed AV recognition system over IP are discussed. In section IV, experimental results for multiband feature fusion technique and AV recognition system over IP are presented. Finally, conclusion of the paper is given in section V.

II. Multiband Feature Fusion Method

Findings in a few published works [13], [22], [23], [32] show the evident of the existence of wavelet subbands that are invariant to illumination variations. For example, Naster et al. [22], [23] found that changes in illumination affect the low-frequency spectrum. This statement indicates that high-frequency components are invariant to illumination. We also know that smaller intrinsic structure that mostly comprised of the detail of the object such as the lines, edges, and small-scale objects is invariant to illumination variation [13]. Since the detail subbands (mid- and high-frequency subbands) carry the small intrinsic structure and high-frequency component, this led us to believe that the detail subband or mid and high-frequency subband is invariant to illumination variation.

The main goal of the proposed multiband feature fusion method is to select the frequency subbands that carry the illumination invariant features. Although there are many methods designed to extract the illumination invariant features [13], [21], their methods either fail to extract detail information from multiscale space or involve complicated parameter selection. In order to extract the illumination invariant features from multiscale space, WPT is used to decompose the image into more compact frequency subbands. Since the selected subbands are expected to be invariant to illumination, we require these subbands to satisfy the following conditions:

Condition 1. The similarity between classes should be at minimum.

Since we know that if the class separation is large, it is easier to discriminate the classes. To test the class separation, one image per class (subject) is used. The face images are chosen randomly from the face database. The similarity matrix $\rho(i, j)$ with the data size m謝 where m refers to number of classes in the database. It records the similarity between image *i* and image *j*. We propose to use Euclidean distance to form the similarity matrix $\rho(i, j)$. For $i = 1,2,...m$

$$Ed(i, j) = \left\| x_i - x_j \right\| \quad j = 1,2,...m \tag{1}$$

After that the $Ed(i, j)$ is normalized to 0 to 1, the similarity matrix is

$$\rho(i, j) = 1 - Ed(i, j) \tag{2}$$

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) [26] is defined as below,

$$AUMSV = \frac{1}{(N^2 - N)} \sum_{i=1}^{N} \sum_{j=1}^{N} \rho(i, j) \tag{3}$$

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.

Condition 2. The ratio for the between-class distance and within-class distance should be maximized.

Fisher's Linear Discriminant (FLD) [16] is a class specific method where it shapes the data scatter in such a way that the ratio of the between-class scatter and the within-class scatter is maximized. By maximizing this ratio, FLD is more reliable for classification. We use the same concept; a method based on Euclidean distance is proposed to show the subband representation in term of data scatter. For each class $k, k = 1,2,...m$, the center is generated as the mean value of the sample patterns belong to the class,

$$M^k = \frac{1}{n^k} \sum_{i=1}^{n^k} S_i^k \qquad k = 1,2,...m \tag{4}$$

where n^k is the total number of samples in class k and S_i^k is the i th sample belonging to class k . The average within-class Euclidean distance d_k from the mean M^k to the furthest point S^k of class k ,

$$d_k = \frac{1}{m} \sum_{k=1}^{m} \left\| S^k - M^k \right\| \tag{5}$$

The distance $d_c(k, j)$ between the mean of class k and the mean of other class j ,

$$d_c(k, j) = \left\| M^k - M^j \right\| \qquad j = 1,2,...,m \tag{6}$$

The average between-class distance d_c is

$$d_c = \frac{1}{m^2} \sum_{k=1}^{m} \sum_{j=1}^{m} d_c(k, j) \tag{7}$$

The ratio for the between-class distance and within-class distance is described as

$$BWR = \frac{d_c}{d_k} \tag{8}$$

The BWR term gives an idea of the scatter of the subband representation, which means the higher theBWR , the better the classification performance.

Below are the steps proposed to select the optimal subbands:

Step 1: Compute the AUMSV and BWR in each subbands from level 1 and 2.

Step 2: Three subbands that obtain the lowest AUMSV and the highest BWR will be selected for further decomposition to level 3. This step reduces the computational complexity by avoiding decomposition of all subbands from level 2 to level 3.

Step 3: Further decompose the selected three subbands to level 3.

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

After the OMF is found the histogram equalizer (HEQ) is applied on to the subband in OMF. The HEQ is one of the most useful contrast enhancement schemes which it maps the image pixel to uniformly distributed pixel values. The HEQ is normally used to enhance contrast of the global images because it does not consider the detail image [20]. Since the OMF is expected to contain the detail images or high-frequency components, we apply HEQ on the subbands to enhance the contrast of the detail images.

III. AV Recognition System over IP

After selecting the OMF, the multiband feature fusion method is implemented in the face recognition system of the AV recognition system and this system is implemented over internet protocol. Figure 2 depicts architecture of video and audio streaming over network for the AV recognition system. 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 [33], [34]. 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 [35]. 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. Unlike UDP, TCP uses retransmission to recover lost packets and it introduces delays that are not acceptable for streaming application with stringent delay requirement [35], and therefore UDP is used as the transport protocol. The RTP is employed as the upper-layer transport protocols.

3) Packetizer: 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 [36]. 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 [36]. 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.

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. 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. At the server side, the received packets will be depacketized and passed to the audio and video decoder. The decoded image frames will be decomposed by the WPT and the OMF is extracted. HEQ is then applied onto the subbands in OMF to enhance the image contrast. The audio signal will be the MFCC for audio feature extraction and LDA for feature selection. Both the audio and facial features will be fused by the intra-modal feature fusion method [28] and The RBF neural network performs classification. The RBF neural network intra-modal fusion algorithm is shown as below. Let V be the feature set from visual signal and A be the feature set from audio signal,

$$V = [v^1 \,|\, v^2 \,|\, ... \,|\, v^N] \tag{9}$$

$$A = [a^1 \,|\, a^2 \,|\, ... \,|\, a^N] \tag{10}$$

The resulting fused matrix will be x which combining V with the data matrix of size $P \times N$ and A with the data matrix size of $Q \times N$, where P and Q is the feature dimension of visual and audio respectively, putting V and A side-by-side we get

$$x = [V \,|\, A]^T \tag{11}$$

where x is a input data with the size $S \times N$ where $S = P + Q$. x will be fed into RBF neural network for training and testing.

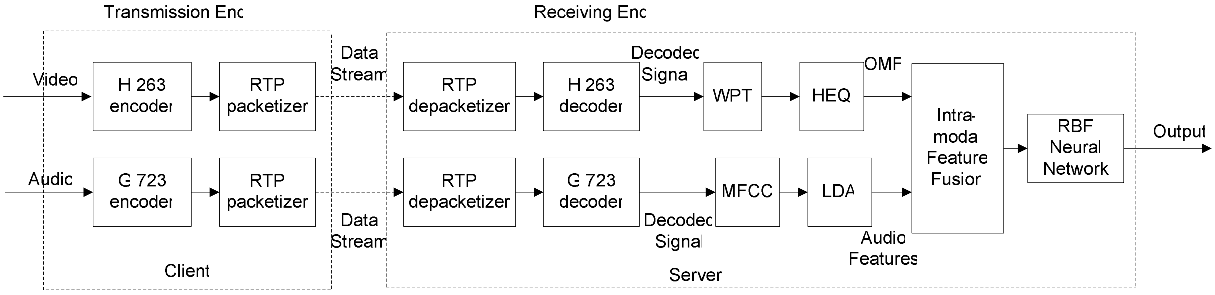

Fig. 2. Block diagram of the developed AV recognition system over IP

IV. Experiments and Results

There are two experiments in this section. The first experiment shows the recognition performance of the proposed method and the second experiment shows the recognition performance of the AV recognition system over IP.

A. *Multiband Feature Fusion*

This experiment first shows how the proposed multiband feature fusion method selects the frequency subbands that are invariant to illumination variation. The Extended Yale face database B (EYaleB) that contains 38 subjects and AR database that contains 100 subjects are used in the selection. In EYaleB and AR database, there are total 152 and 400 cropped faces respectively. Both the databases contain illumination variations in the images that occur due to intensity and direction of the light. All images are scaled to 128 × 128 pixels resolution. For each subject in the set, two of their images that contain frontal illumination with normal light are used as the gallery set, and the remaining two images that contain illumination from sides are used for testing. Sample face images are shown in Figure 3. In this experiment, the nearest neighbor classifier based on Euclidean distance is employed for classification.

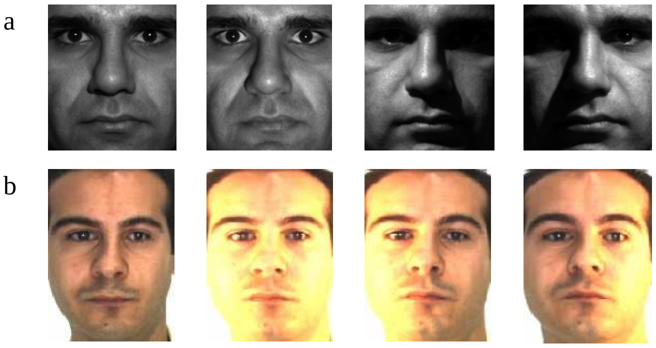

Fig. 3. Some cropped faces used in this experiment: (a) EYaleB database; (b)AR database

Experiments on the WPT level-1 and -2 are first carried out in both databases. The Haar wavelet is used in the proposed method. The AUMSV, the between-class distance and within-class distance ratio BWR and recognition error rate of all the subbands decomposed from level-1 and level-2 are generated. For notation, the A, H, V and D indicate the approximation, horizontal, vertical and diagonal subband in each of the LL, LH, HL and HH subbands. For example, ALL refers to the level-2 approximation subband decomposed from level-1 LL subband. Table I shows that the ALL, HLL and ALH achieve the lowest AUMSV and the highest BWR in both databases. These subbands potentially perform well in face classification under illumination variation conditions. Hence, these three subbands are further decomposed to level-3. Table II shows that the AALH achieves the lowest AUMSV, the highest BWR and the lowest error rate among the others in both the databases. AALH outperforms HALL which was found to be invariant to illumination variation in [19] in term of recognition accuracy. Next, we concatenate the HALL and AALH to form OMF. The OMF achieves recognition error rates of 18.4% and 14% in EYaleB and AR database respectively. The OMF achieves improvements of 5.3% and 1% as compared to AALH in term of recognition error rate in EYaleB and AR database respectively. Figure 4 shows the location of AALH and HALL in frequency subband.

After obtaining the OMF, the recognition performance of OMF with histogram equalizer (HEQ) is evaluated. Since the OMF contains the detail image, we apply HEQ on the AALH and HALL individually to enhance the contrast of the detail images and then we concatenate these AALH and HALL to form the OMF again. Figure 5 displays faces of one subject from the EYaleB database illuminated by a light source and faces after HEQ is applied on the faces. Figure 4a shows the effect of HEQ on the approximation subband AALL at level-3, the HEQ enhances the global image but not the detail image. Figure 5b and c show that the contrast of the detail images in AALH and HALL are enhanced and we can see definite detail in both the subbands.

The recognition performance of OMF+HEQ is evaluated on the Yale database B (YaleB). This database contains face images with large illumination variation. There are ten subjects under 64 different lighting conditions. The database is divided into four subsets according to the angle between the lighting source direction and the camera axis. Table III shows the subset with the corresponding angles and number of images. Comparison results with other methods dealing with illumination variations on YaleB database are shown in Table IV. Since our proposed method does not involve any training process, training images are not needed. We used subset 1 as the gallery set. Some listed results of other methods are directly taken from other papers since they are based on the same database. We can see from Table IV that the OMF+HEQ outperforms most of the existing methods in term of recognition accuracy except cone-cast method and multiscale representation + PCA. However, it should be pointed out that the cone-cast method needs much more complicated modeling steps. There is no complicated parameter selection in the proposed method as compared to the multiscale representation + PCA method.

Table I AUMSV, ratio R and recognition error rate values in EYaleB and AR databases for level-1 and -2 decomposition

	EYaleB			AR		
Subband	AUMSV	BWR	Rate (%)	AUMSV	BWR	Rate (%)
LL	0.360	0.740	56.6	0.479	1.197	52.5
ALL	0.358	1.000	67.1	0.338	1.201	57
HLL	0.295	1.097	31.6	0.451	1.299	32.5
VLL	0.438	0.866	57.9	0.490	0.707	61
DLL	0.442	0.887	69.7	0.468	0.796	69.5
LH	0.454	0.841	77.6	0.469	1.075	54
ALH	0.248	1.259	26.3	0.469	1.333	24
HLH	0.486	0.806	73.7	0.505	0.839	66.5
VLH	0.414	0.738	84.2	0.476	0.836	70
DLH	0.503	0.664	85.5	0.556	0.677	91.5
HL	0.42	0.797	65.8	0.543	0.694	73.5
AHL	0.432	0.848	47.4	0.476	0.803	24
HHL	0.359	0.890	59.2	0.479	0.869	66.5
VHL	0.499	0.868	80.3	0.479	0.776	70
DHL	0.522	0.555	84.2	0.524	0.730	91.5
HH	0.554	0.768	80.3	0.518	0.652	91

Table II AUMSV, ratio R and recognition error rate values in EYaleB and AR databases for level-3 decomposition

	EYaleB			AR		
Subband	AUMSV	BWR	Rate (%)	AUMSV	BWR	Rate (%)
AALL	0.348	0.646	39.5	0.397	1.183	64.5
HALL	0.271	1.282	55.3	0.396	1.493	15.5
VALL	0.389	0.981	68.4	0.451	0.743	64
DALL	0.363	0.933	75.0	0.398	1.200	31
AHLL	0.281	1.219	61.8	0.414	1.399	21
HHLL	0.382	1.11	60.5	0.487	1.357	43.5
VHLL	0.361	0.883	69.7	0.399	0.997	51.5
DHLL	0.329	0.904	68.4	0.398	0.907	49
AALH	0.241	1.29	23.7	0.368	1.494	15
HALH	0.312	1.13	47.4	0.453	1.346	36
VALH	0.278	0.944	68.4	0.400	1.111	49.5
DALH	0.508	0.848	71.1	0.411	0.893	57

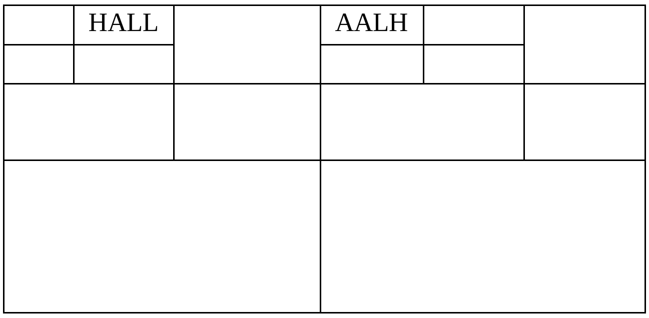

Fig. 4. Location of the HALL and AALH in frequency subband

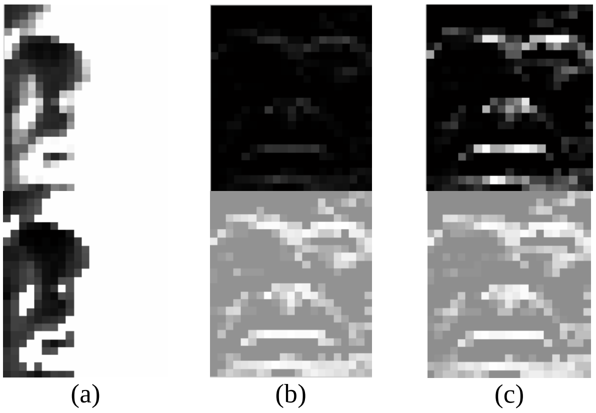

Fig.5. The top row displays the faces before the HEQ and the second row display the effect of HEQ on the faces in the top row. a) the approximation subband face image AALL, b) the AALH, and c) the HALL.

Table III Subsets divided according to light source directions

Subset	1	2	3	4
Lighting angle (°)	0~12	13~25	26~50	51~77
Number of images	70	120	120	140

Table IV Recognition error rates (%) of different methods

Method	Subset 2	Subset 3	Subset 4
PCA w/o3 [16]	4.4	27.7	-
Fisherface [16]	0	4.6	-
Linear subspace [30]	0	15	-
Cone-attached [30]	0	0	8.6
Cones-cast [30]	0	0	0
Harmonic images [6]	0	0.3	3.1
Quotient image [8]	1.7	38.1	65.9
Quotient illumination relighting [12]	0	0	9.4
Self Quotient Image [10]	2.0	1.0	3.0
Illumination ratio images [9]	0	3.3	18.6
DCT in Log domain [14]	0	0.18	1.71
Wavelet Reconstruction [20]	0	0	5.24
Multiscale representation +PCA [21]	0	0	0
OMF	0	0	10.8
OMF + HEQ	**0**	**0**	**5**

The recognition performance of the proposed OMF+HEQ is compared with the recognition performance of PCA [3], PCA w/o 3 [15] and Independent Component Analysis (ICA) [38] under larger variations. The EYaleB, AR and CUAVE databases [24] are included in this experiment. The number of distinct subjects, the number of gallery images and the number of testing images in the respective databases are tabulated in Table V. The setting of the EYaleB database is the same with the previous experiment. However, unlike the setting in the previous experiment, in the AR database, other than ligting from left and right, it also contains facial expression variations. For CUAVE AV database, it contains moving subjects with different facial expressions and constant lighting. Images with neutral facial expression and constant lighting conditions are chosen to be the gallery images from these databases. Table VI shows that the OMF+HEQ outperforms PCA, PCA w/o 3 and ICA in term of recognition accuracy in the three databases.

Table V The databases used in the experiments

	EYaleB	AR	CUAVE
Number of subjects	38	100	36
Number of gallery images	76	300	108
Number of testing images	76	400	108

Table VI Recognition error rates (%) of different methods in three databases

Database	OMF+HEQ	PCA	PCA w/o 3	ICA
EYaleB	18	70	53	61
CUAVE	14	20	23	20
AR	6	36	19	11

B. Audio-Visual Recognition System over IP

The proposed method is then implemented in the AV recognition system over IP. The recognition performance of the system is evaluated. The CUAVE AV database [24] is used in the following experiments. The database consists of 36 subjects. It is recorded in an isolated sound booth at a resolution of 720x480 with NTSC standard of 29.97fps. The data is then compressed into individual MPEG2 files for each speaker. The MPEG2 files are encoded at a data-rate of 5000kbps with multiplexed 16-bit, stereo audio at 44 kHz sampling rate. JMstudio is used to transmit and receive the data [37]. We organize our presentation as follows. Part 1 shows the recognition performance of speech compression to speaker recognition system over IP. Part 2 shows the recognition performance of video compression to face recognition system over IP. For part 1 and 2, the link capacity between the client and server sides is 10Mbits/s. At last part 3 shows the recognition performance of audio and video under varying network link capacities. Bandwidth Limiter Enterprise [11] is used to control the link capacities.

1) Audio over IP

In this experiment, we evaluate the influence of speech compression and speech quality over IP on speaker recognition performance. At the client side, the “wav” format audio files are compressed by the audio codec G.723 to bit rate of 8kbit/s and 16kbit/s. These data are then streamed to the server side for recognition performance evaluations. As shown in Figure 2, MFCC and LDA are used as the feature extraction and selection methods for the audio and RBF neural network is used as the classifier. Three training samples per subject are used for the RBF training. The width of the neuron is 10 and the number of neuron is 25. Table VII shows that the speaker recognition performance is less affected when the speech with 16kbit/s when it is streamed over IP. This is because peer-to-peer network link capacity offers enough transfer speed for the data.

Table VII Speaker recognition results for standalone system (without going through IP) and transcoded data over IP

Audio bit-rate	Error rate (%)
Standalone (16Kbits/s)	14
Over IP G.723 (16Kbits/s)	15
Over IP G.723 (8Kbits/s)	36

2) Video over IP

In this experiment, we evaluate the influence of the video dimension to image quality over IP on face recognition performance. At the client side, the video are first being encoded to three different video dimensions (Mode A): SQCIF (128x96), QCIF (176x144) and CIF (352x288) and then the files are streamed to the server side for recognition performance evaluations. All images are scaled to 128 × 128 pixels resolution for feature extraction. The OMF is extracted and the HEQ is applied, then the RBF neural network is used as the classifier. Three training samples per subject are used for the RBF training. The width of the neuron is 30 and the number of neuron is 100. The result shows that the recognition error rate increases when the video dimension decreases.

Table VIII Face recognition results for standalone system (without going through IP) and transcoded data over IP

Video Dimension	Error rate (%)
Standalone (720x480)	5.6
Over IP CIF (352x288)	19.4
Over IP QCIF (176x144)	25
Over IP SQCIF (128x96)	38.9

3) Audio-Visual over IP

In this experiment, we evaluate the recognition performance of AV recognition system under varying bandwidth. The link capacity at the client side is fixed at 10Mbits/s and the link capacity at the server sides varies from 160Kbits/s, 400Kbits/s, and 8Mbits/s. G.723 with 16Kbits/s is selected as the codec in this part of the experiment due to its low error rate as shown in Table VII. Three training samples per subject are used for the RBF training. The width of the neuron is 30 and the number of neuron is 100. Table IX shows the recognition performance of the system under low link capacity (160Kbits/s). Due to the limited link capacity, the result shows that when the high dimensional video CIF is transmitted, the packet loss ratio is the highest and causes the highest recognition error rate. Figure 6 illustrates the packet loss and delay jitter effect caused by the network congestion for CIF video. When QCIF and SQCIF are at the same limited link capacity, the packet loss ratio is 11.8% and 0% respectively and causing the same error rate at 25%. Table X shows that only CIF encounter packet loss and as the link capacity increases to 400Kbits/s, the packet loss ratio deceases and as a result, the error rate reduces. Smaller video formats QCIF and SQCIF achieve a constant error rate at 13.9% and 25% across link capacity of 400Kbits/s to 8Mbits/s as shown in Table X and XI. At larger link capacity 8Mbits/s, there is no packet loss for CIF and it achieves the lowest error rate. From all the results shown, we see that due to the large video size of CIF, the recognition performance of the system is highly influenced by the link capacity, whereas due to the small video size of SQCIF, the recognition performance of the system is not influenced by the link capacity, however, the error rate is high. We can also see that the recognition performance of QCIF is the most promising one. This is because the recognition performance of QCIF is less affected by the link capacity variations.

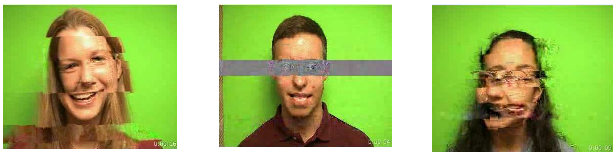

Fig. 6. Examples of CIF images contains packet loss and delay jitters taken under link capacity of 160Kbits/s and 400Kbits/s

Table IX Packet loss ratio and recognition error rate of the AV recognition system over 160Kbits/s link capacity

Video Dimension	Audio bit-rate	Packet Loss Ratio (%)	Error rate (%)
Standalone (720x480)	Standalone (16kbit/s)	0	2.8
Over IP CIF (352x288)	Over IP G.723 (16kbit/s)	65.4	47.2
Over IP QCIF (176x144)	Over IP G.723 (16kbit/s)	11.8	25
Over IP SQCIF (128x96)	Over IP G.723 (16kbit/s)	0	25

Table X Packet loss ratio and recognition error rate of the AV recognition system over 400Kbits/s link capacity

Video Dimension	Audio bit-rate	Packet Loss Ratio (%)	Error rate (%)
Standalone (720x480)	Standalone (16kbit/s)	0	2.8
Over IP CIF (352x288)	Over IPG.723 (16kbit/s)	22.6	33.3
Over IP QCIF (176x144)	Over IP G.723 (16kbit/s)	0	13.9
Over IP SQCIF (128x96)	Over IP G.723 (16kbit/s)	0	25

Table XI Packet loss ratio and recognition error rate of the AV recognition system over 8Mbits/s link capacity

Video Dimension	Audio bit-rate	Packet Loss Ratio (%)	Error rate (%)
Standalone (720x480)	Standalone (16kbit/s)	0	2.8
Over IP CIF (352x288)	Over IP G.723 (16kbit/s)	0	8.3
Over IP QCIF (176x144)	Over IP G.723 (16kbit/s)	0	13.9
Over IP SQCIF (128x96)	Over IP G.723 (16kbit/s)	0	25

V. CONCLUSION

In this paper, the audio-visual recognition system which is insusceptible to illumination variation over internet protocol is presented. The multiband feature fusion method is proposed to select the illumination invariant subbands. The selected AALH and HALL form the OMF. The OMF is proved to be invariant to illumination variation by the statistical method, the modified Euclidean based method and the recognition accuracy. Recognition performance of the proposed method achieves higher recognition accuracy as compared to most of the existing methods. Then the audio-visual recognition system over internet protocol is developed where the proposed method is implemented in the face recognition system. The compression and packet loss effects of audio and video data sent over varying link capacities on recognition performance are investigated. The result has shown that low bit-rate speech compression and lower-dimensional video degrade the recognition performance in speaker and face recognition system over IP. Besides, the results show that large data size video (CIF) encounters serious packet loss effect when the video is sent over the limited link capacity network. As a result, the recognition performance of the system is degraded. The QCIF is shown to be most suitable to be used in the AV recognition system over IP because it achieves promising recognition performance in both limited and unlimited link capacities.

REFERENCES

[1] N.A. Fox, R. Goss, P. de Chazal, J.F. Cohn, and R. B. Reilly, " Person identification using automatic integration of speech, lip, and face expert," *Proc. ACM SIGMM 2003 Multimedie Biometrics Methods and Application Workshop (WBMA'03), Berkeley, CA*, 2003. pp. 25-32.

[2] C.C. Chibelushi, F. Deravi, and J.S.D. Mason, "A review of speech-based bimodal recognition", *IEEE Trans. On Multimedie,* vol. 4, no. 1, pp. 23-37, March 2002.

[3] M. Turk, A. Pentland, "Eigenfaces for recognition", J. Cogn. Neurosci. 3 (1) (1991) 71–86.

[4] D. L. Swets and J. Weng, "Using discriminant eigenfeatures for image retrieval", *IEEE Trans. On Pattern Anal. And Mach. Intell.*, vol. 18, no. 8, pp. 831-836, August 1996.

[5] P. Belhumeur and D. Kriegman, "What is the set of images of an object under all possible lighting conditions," Int"l J. Computer Vision, vol. 28, pp. 245-260, 1998.

[6] L. Zhang and D. Samaras, "Face recognition under variable lighting using harmonic image examplers," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* vol. 1, pp. 19-25, 2003.

[7] K.-C. Lee, J. Ho, and D. J. Kriegman, "Acquiring linear subspaces for face recognition under variable lighting illuminations," *IEEE Trans. On Pattern Anal. And Mach. Intell.*, vol. 27, no. 5, pp. 684-698, May 2005.

[8] A. Shashua and T. Riklin-Raviv, "The quotient image: class-based rerendering and recognition with varying illuminations.," *IEEE Trans. On Pattern Anal. And Mach. Intell.*, 23(2):129-139, 2001.

[9] J. Zhao, Y. Su, D. Wang, and S. Luo, "Illumination ratio image: synthesizing and recognition with varying illuminations," *Pattern Recog. Lett.*, vol. 24, pp. 2703-2710, 2003.

[10] H. Wang, S. Li, and Y Wang, "Face recognition under varying lighting condition using self quotient image," *Proc. IEEE AFGR*, 2004.

[11] Bandwidth controller enterprise: wareseeker.com/Network-Tools/bandwidth-controller-enterprise-1.18.zip/280994

[12] S. Shan, W. Gao, B. Cao, and D. Zhao, "Illumination normalization for robust face recognition against varying lighting conditions," *Proc. IEEE workshop on AMFG*, 2003.

[13] T. Chen, W. Yin, X. S. Zhou, D. Camaniciu, and T. S. Huang, "Total Variation Models for Variable Lighting Face Recognition," *IEEE Trans. On Pattern Anal. And Mach. Intell.*, vol. 28, no. 9, 2006.

[14] W. Chen, M. J. Er. And S. Wu, "Illumination compensation and normalization for robust face recognition using discrete cosine transform in logarithm domain." *IEEE Trans, on Syst., Man, and Cybernetics- Part B: Cybernetics*, vol. 36, no. 2, pp. 458-466, April 2006.

[15] M. Turk, A. Pentland, "Eigenfaces for recognition", *J. Cogn. Neurosci*. 3 (1) (1991) 71–86.

[16] P. Belhumeur, J. Hespanha, and D. Kriegman, "Eigenfaces vs fisherfaces: Recognition using class specific linear projection," *IEEE Trans. On Pattern Anal. And Mach. Intell.*, 19:711-720, 1997.

[17] C. Garcia, G. Zikos, G. Tziritas, "Wavelet packet analysis for face recognition," *Image and Vision Computing*," 18 (4) (2000) 289–297.

[18] V. Pathangay and S. Das, "Selection of wavelet subbands using genetic algorithm for face recognition," *Computer Vision, Graphics and Image Processing*, vol. 4338, pp. 585-596, 2006.

[19] H. K. Ekenel and B. Sankur, "Multiresolution face recognition", *Image and Vision Computing*, vol. 23, pp. 469-477, 2005.

[20] Shan Du and R. Ward, "Wavelet-based illumination normalization for face recognition," *Proc. IEEE ICIP*, 2005.

[21] T. Zhang, B. Fang, Y. Yuan, Y. Y. Tang, Z. Shang, D. Li, F. Lang, "Multiscale facial structure representation for face recognition under varying illumination," *Pattern Recognition*, vol. 42, pp. 251-258, 2009.

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

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

[24] E.K. Patterson, S. Gurbuz, Z. Tufekci, and J.N. Gowdy, "CUAVE: A New Audio-Visual Database for Multimodal Human-Computer Interface," in *Proc. ICASSP*, pp. 2017-2020, 2002.

[25] A. Primer, *Introduction to wavelet and wavelet transform*, Prentice Hall, 1998.

[26] 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.

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

[28] 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.

[29] S. Chen, C. C. N. Cowan, and p. M. Grant, “Orthogonal least squares for radial basis function networks,” *IEEE Trans. On Neural Network*, vol. 2, no. 2 pp. 302-309, 1991.

[30] A. S. Georghiades, P. N. Belhumeur, and D. W. Jacobs, “From few to many: illumination cone models for face recognition under variable lighting and pose,” *IEEE Trans. Pattern Anal. Mach. Intel*., vol. 23, no. 6, pp. 630-660, June. 2001.

[31] A.M. Martinez and R. Benavente, “The AR face database," CVC Tech. Report #24, 1998.

[32] Y. Adini, Y. Moses, and S. Ullman, “ Face recognition: The problem of compensating for changes in illumination direction,” *IEEE Trans. Pattern Anal. Mach. Intel*., 19(7), 1997.

[33] 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.

[34] 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.

[35] 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.

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

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

[38] 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.