Multiband Curvelet-Based Technique for Audio Visual Recognition over Internet Protocol

Sue Inn Ch'ng[1], KahPhooi Seng[2], Fong Tien Ong[1], and Li-Minn Ang[1]

[1] University of Nottingham Malaysia Campus
JalanBroga, 43500 Semenyih, Selangor, Malaysia
{keyx9csi,keyx1ofe,Kenneth.Ang}@nottingham.edu.my
[2] Sunway University
No. 5, JalanUniversiti, Bandar Sunway, 46150 PetalingJaya, Selangor, Malaysia
Jasmine.Seng@nottingham.edu.my

Abstract. The transmission of the entire video and audio sequences over an internal or external network during the implementation of audio-visual recognition over internet protocol is inefficient especially when only selected data out of the entire video and audio sequences are actually used for the recognition process. Hence, in this paper, we propose an efficient method of implementing audio-visual recognition over internet protocol whereby only the extracted audio-visual features are transmitted over internet protocol. To extract the robust features from the video sequence, a multiband curvelet-based technique is employed at the client whereas a late multi-modal fusion scheme using RBF neural network is employed at the server to perform the recognition across both modalities. The proposed audio-visual recognition system is implemented on several standard audio-visual databases to showcase the efficiency of the system.

Keywords: curvelet transform, multiband technique, internet protocol, windows sockets.

1 Introduction

An audio-visual (AV) recognition system is a multi-modal biometric system that recognizes a person based on the person's audio (voice) and visual (face) data. However, most of these systems [1],[2],[3] are localized and does not permit remote authentication or recognition to be done. To solve this problem, a method of implementation over internet protocol (IP) is required. The works in [4] uses Java Media Framework API (JMF) to stream the video and audio data of the subject acquired from the client to the server where the recognition process is done. Based on the study carried out in these papers, it can be noted that this method of implementation is highly ineffective as the system is susceptible to packet losses when the network conditions are poor. According to the results reported in [4], the packet loss phenomenon has an adverse effect on the performance of the audio-visual recognition system. To overcome this drawback, the later work in [5] uses multi-frame technique to compensate the loss

V.V. Das, E. Ariwa, and S.B. Rahayu (Eds.): SPIT 2011, LNICST 62, pp. 132–138, 2012.

of information by utilizing the information contained from several frames during the recognition process. Although there is an improvement in the performance compared to using one frame for testing, there is still a slight degradation in the performance of the system implemented over internet protocol compared to the performance of the same system in its standalone state. This is because the actual problem of packet loss was not addressed but a solution to mitigate the effects of packet loss was only done. Furthermore, it is noted that the entire video and audio sequences are transmitted over the network in the above-mentioned systems while only several frames of the data are used for the recognition process at the server. This process is highly inefficient as it is a waste of network resources which is limited especially in the case of mobile networking.The use of curvelet transform [6] for face recognition has been widely explored in the works of [7-9]. These works show that curvelet transform is a viable alternative to wavelets as the latter has limited directional representation and requires more wavelets for edge representation. However, the earlier works of curvelet for face recognition only deals with face recognition on a general basis and does not particularly delve into the problem of specific visual variations such as illumination variation. The more recent work in [10] uses multiband technique to extract illumination invariant features from the decomposed curveletsubbands to improve the performance of the system to illumination variation. Nonetheless, the isolated illumination invariant features which are extracted from the high frequency subbands are less effective against expression variation because the illumination variation affects the low frequency components whereas the expression variation affects the high frequency components [11, 12]. Hence, by solely selecting the high frequency subbands to solve illumination variation will cause the current system to be susceptible to expression variation and vice versa. To overcome this problem, the work in [13] proposed a fusion scheme to harvest the benefits of both high and low frequency subbands to solve expression and illumination variations. However, the work in [13] is based on wavelets.

Hence, in this paper, we extend our previous work [10] to extract expression invariant features from the decomposed curvelets using multiband curvelet-based technique. The extracted expression and illumination invariant features are then applied to two separate radial basis function neural networks to be classified. Subsequently, the score adaptive fusion used in [13] is used to combine the classification scores of the expression and illumination invariant features. In addition to the extension of the multiband curvelet-based technique to cover expression and illumination variations, a more efficient method of implementing the overall audio-visual system over internet protocol is also proposed. The proposed implementation reduces the load on the network by transmitting only the extracted audio and visual features over internet protocol. The performance of the proposed audio-visual recognition system over internet protocol is tested with CUAVE, XM2VTS and VIER database.The outline of the paper is as follows: Section 2 describes in detail the architecture of the audio-visual system over internet protocol. The simulation results showcasing the performance of the proposed multiband curvelet-based technique and the proposed system over internet protocol is provided in Section 3. Finally, Section 4 concludes the paper.

2 Proposed AV Recognition System over Internet Protocol

The proposed audio-visual system consists of a client and server. At the client, the video of the test subject to be recognized is first captured using a webcam. The data is then demultiplexed into video and audio streams. Subsequently, the proposed multiband curvelet-based technique and MFCC[14]+LDA[15] are applied to the streams to extract the visual and audio features respectively. These extracted features are then sent to the server using the application program created using C++ and Windows Socket (WinSocks) [16]. Since the data size of the extracted features are small in comparison with the original video and audio stream, Transmission Control Protocol (TCP) is opted as the transport protocol between server and client compared to User Datagram Protocol (UDP). The use of TCP is advantageous in this case as it employs acknowledgement and retransmission to ensure that the received data stream is uncorrupted. Hence, the issue of packet loss is eliminated. At the server, the received data is passed to the RBF neural networks to be classified. To combine the audio-visual scores and the scores of the different feature sets, the three-level fusion scheme proposed in [5] is used. Figure 1 shows the block diagram of the overall proposed AV recognition system over internet protocol. The process of the proposed multiband curvelet-based technique is as follows: First, the video frames are decomposed to their subband representations using Fast Discrete Curvelet Transform (FDCT) via Wrapping [17]. The multiband selector is then applied on the decomposed curvelet to select the expression and illumination invariant features. The multiband selector uses the Average Unmatched Similarity Measure (AUMSV) and between-within class ratio to determine the level of representation contained by each subband. To determine which subbands contain the expression invariant features, we use the findings in [11, 12] to minimize our search to the approximate curvelets at different scales. Based on our simulation results on the Yale and AR database, the expression invariant features were located at the approximate subband for scale 2 and scale 3 whereas the illumination invariant feature were located at (2,1) and (2,2). Hence, in the following evaluations, the approximate subbands for scale 2 and scale 3 and the subbands at (2,1) and (2,2) are concatenated together to form the optimal curveletsubbands for expression (Curve_E) and illumination (Curve_I) respectively.

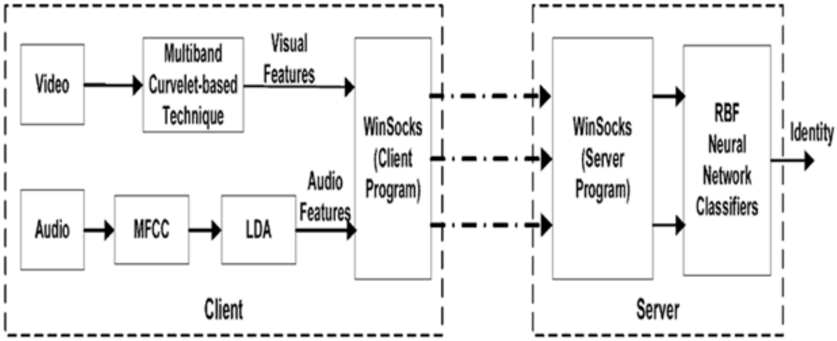
Video
Multiband Curvelet-based Technique
Visual Features
WinSocks (Client Program)
Audio
MFCC
LDA
Audio Features
Client
WinSocks (Server Program)
RBF Neural Network Classifiers
Identity
Server

Fig. 1. Block diagram of the proposed AV recognition system over internet protocol

During the implementation over IP, the visual features (both Curve_E and Curve_I) and audio features are combined to form a singular matrix, *F*, prior to transmission. At the server, these features are demultiplexed back to their original form by using the received sizes of the features to reshape the received data before it is applied to the corresponding neural network for classification.

3 Results and Discussion

There are two experiments carried out in this section. The aim of the first experiment is to showcase the effectiveness of the proposed multiband curvelet-based method to other methods, particularly gradientfaces[18], dual optimal multiband features[13] and eigenfaces[19]. This is done by first evaluating the isolated optimal curveletsubbands for expression (curve_E) and illumination (curve_I) and subsequently its fusion state using the method proposed in [13]. The experiments were carried out on three standard databases namely AR database[20], Yale database [21] and ORL database[22]. Two training samples were used for each of the databases. Only images with even lighting and neutral expressions were chosen to be used as the training samples. For the case of Yale and ORL databases, the rest of the images were used as the testing samples. These images contain a mixture of facial expressions and illumination variation. On the other hand, two images containing illumination variation and two images containing expression variations were chosen as the testing samples for the AR database. The classifier used throughout the performance evaluation including for the other methods is RBF neural network classifier. The weight value for the adaptive decision fusion used to combine both feature sets in the proposed curvelet-based multiband feature technique and dual optimal multiband features is set to 0.6 as described by the original paper. Table 1 shows the recognition rate obtained for the individual feature sets and the performance of the proposed multiband curvelet-based method while Table 3 shows the performance comparison results of the proposed method to other techniques.

From Table 1, it can be seen that by using the method of fusion proposed in [13], a higher recognition rate can be achieved compared to the individual feature sets. This shows that the fusion method proposed by Wong et. al. is effective in solving the problem of adverse effect of compensating one variation for the other. However, based on the simulation results in Table 2, it can be seen that the performance of the original DOMF method can be further improved with the use of curvelets instead of wavelets.

Table 1. Recognition performance (%) for individual feature sets and the fusion set

	Curve_E	Curve_I	Fusion Set
AR	69.50	68.50	80.25
Yale	75.00	65.00	88.15
ORL	86.56	62.19	83.44

Table 2. Recognition performance (%) of the proposed method compared to other techniques

	Gradientfaces	DOMF	PCA	Multiband Curvelet-based Method
AR	76	78.5	30.8	**80.3**
Yale	86	83.7	76.3	**88.2**
ORL	71.9	83.4	58.1	**83.4**

The aim of the second experiment is to evaluate the performance of the proposed audio-visual recognition system over IP. This experiment was carried out on the CUAVE [23], XM2VTS [24] and UNMC-VIER [25] audio-visual database. For the visual parts of the experiment on the CUAVE and XM2VTS database, the first five frames were taken for training and the sixth frame was taken for testing. On the other hand, for VIER database, the five frames of the video from the controlled environment with neutral expression were used for training. One test image with facial expression and one with illumination variation were then used as the testing samples. For the audio part of the experiment, five training samples and one testing sample of each subject speaking the utterance “zero” was used for all the databases. The fusion weights between the two modalities are set to 0.5 whereas the weight value for the adaptive fusion between the features set is set to 0.6. To evaluate the effectiveness of the proposed AV system over internet protocol, the link capacity at the client and server was set according to 10Mbit/s and 380Kbit/s respectively. This is the same settings used in [5]. For reference purposes, the standalone system was also evaluated. The performance of both systems is reported in Table 3. Based on the results obtained, it can be seen that the implemented system over internet protocol is unaffected by the network conditions. This is due to the type of transport protocol used which removes packet loss completely. Hence, the data received at the server is unaffected by the network conditions as compared to the works in [4, 5].

Table 3. Recognition performance (%) of the proposed system and its standalone counterpart

	Standalone			Proposed System		
	Audio	Visual	Combined AV	Audio	Visual	Combined AV
CUAVE	83.33	100.00	100.00	83.33	100.00	**100.00**
UNMC-VIER	71.54	95.12	97.56	71.54	95.12	**97.56**
XM2VTS	75.00	100.00	100.00	75.00	100.00	**100.00**

4 Conclusion

In this paper, an audio-visual recognition system over IP using multiband curvelet-based technique was proposed. The proposed system uses WinSocks programming to implement a reliable transmission process between the client and server. To improve the efficiency of the transmission, only the extracted audio and visual features are used. A multiband curvelet-based technique is used in the proposed system to extract robust features that is invariant to illumination and expression variation. The simulation results of the proposed system shows that the proposed system is unaffected by the network conditions and has similar results to that of its standalone counterpart.

References

1. Chibelushi, C., Deravi, F., Mason, J.S.: Voice and facial image integration for speaker recognition. In: Proc. IEEE Int. Symp. Multimedia Technologies Future Appl. IEEE, Southampton (1993)
2. Brunelli, R., Falavigna, D.: Person identification using multiple cues. IEEE Trans. Pattern Anal. Machine Intell. 10, 955–965 (1995)
3. Jourlin, P., et al.: Integrating acoustic and labial information for speaker identification and verification. In: Proc. 5th Eur. Conf. Speech Communication Technology, Rhodes, Greece (1997)
4. Yee Wan Wong, K.P.S., Ang, L.M.: Audio-visual authentication or recognition System Insusceptible to Illumination Variation over Internet Protocol. IAENG International Journal of Computer Science Mag. 36(2), IJCS_36_2_0
5. Ch'ng, S.I., et al.: Robust Video Authentication System over Internet Protocol. International Journal of Biometrics 3(4), 322–336 (2011)
6. Candes, E.J., Donoho, D.L.: Curvelets, multiresolution representation, and scaling laws. In: Proc. SPIE 2000, vol. 4119(1) (2000)
7. Aroussi, M.E., et al.: Local appearance based face recognition method using block based steerable pyramid transform. Signal Processing (91), 38–50 (2010)
8. Rziza, M., et al.: Local curvelet based classification using linear discriminant analysis for face recognition. International Journal of Computer Science 4(1), 72–77 (2009)
9. Mandal, T., Wu, Q.M.J., Yuan, Y.: Curvelet based face recognition via dimension reduction. Signal Process. 89(12), 2345–2353 (2009)
10. Ch'ng, S.I., Seng, K.P., Ang, L.-M.: Curvelet-based illumination invariant feature extraction for face recognition. In: 2010 International Conference on Computer Applications and Industrial Electronics, ICCAIE (2010)
11. Ekenel, H.K., et al.: Multiresolution face recognition. Image Vision Comput. 23(5), 469–477 (2005)
12. Nastar, C., Moghaddam, B., Pentland, A.: Flexible Images: Matching and Recognition Using Learned Deformations. Computer Vision and Image Understanding 65(2), 179–191 (1997)
13. Wong, Y.W., Seng, K.P., Ang, L.-M.: Dual optimal multiband features for face recognition. Expert Systems with Applications 37(4), 2957–2962 (2010)
14. Campbell Jr., J.P.: Speaker recognition: a tutorial. Proceedings of the IEEE 85(9), 1437–1462 (1997)
15. Lu, X.: Image Analysis for Face Recognition (2003)
16. Comer, D.E.: Internetworking with TCP/IP vol.1: Principles, Protocols, and Architecture, 4th edn. Prentice Hall Internation Inc., Upper Saddle River (2000)
17. Candes, E.J., et al.: Fast discrete curvelet transform. SIAM Multiscale Model Simul. (2007)
18. Taiping, Z., et al.: Face Recognition Under Varying Illumination Using Gradientfaces. IEEE Transactions on Image Processing 18(11), 2599–2606 (2009)
19. Turk, M., Pentland, A.: Eigenfaces for face recognition. Journal of Cognitive Neuroscience 3(1), 71–86 (1991)

20. Martinez, A.M., Benavente, R.: The AR face database (1998)
21. University, Y.: http://www.cvc.yale.edu/projects/yalefaces/yalefaces.html
22. The ORL in Cambridge, U.: http://www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.html
23. Patterson, E.K., et al.: CUAVE: A new audio-visual database for multimodal human-computer interface research. In: Proc. ICASSP (2002)
24. Messer, K., et al.: XM2VTSDB: The Extended M2VTS Database. In: Second International Conference on Audio and Video-based Biometric Person Authentication (1999)
25. Wong, Y.W., et al.: The Audio-Visual UNMC-VIER Database. In: Proceedings of the International Conference on Embedded Systems and Intelligent Technology (ICESIT 2010) (2010)