2010 Second International Conference on Computer Engineering and Applications

Feature Level Fusion of Speech and Face Image based Person Identification System

Gunawan Sugiarta YB. [1], Riyanto Bambang [2], Hendrawan [2], and Suhardi [2]
[1]Electro Technique, State Polytechnic Bandung
[2]School of Electrical Engineering and Informatics, Institute of Technology Bandung
Bandung, Indonesia
ybgsarta@yahoo.com briyanto@lskk.ee.itb.ac.id hend@stei.itb.ac.id mr.suhardi@gmail.com

***Abstract*— Person identification system that used multiple biometric can improve performance than using a single biometric. This paper presents feature level fusion of dual tree complex wavelet transform speech and face image features for person identification system. Results of experiments using a VidTIMIT database presented to show the identification rate of multibiometric system more improved compared with single biometric system.**

Keywords— speech; face image; feature; identification; multibiometric; fusion;

I. INTRODUCTION

The spread of the existing biometric systems tend to be unimodal. Acoustic based methods are susceptible to low acoustic signal-to-noise ratios or channel distortion. Similarly, face based recognition performs poorly when confronted with pose/illumination/expression variation and occlusion. Limitations imposed by unimodal biometric systems (i.e., biometric systems that rely on the evidence of a single biometric trait) can be overcome by using multiple biometric modalities [1, 3].

Multiple biometric system address the problems of non-universality, because some features can ensure adequate population coverage. Furthermore, multiple biometric systems provide anti-spoofing measures by making it difficult for an intruder to simultaneously spoof the multiple biometric traits of a legitimate user [3]. By asking the user to present a random subset of biometric traits, the system ensures that a "live" user is indeed present at the point of data acquisition. Thus, a challenge-response type of authentication can be facilitated using multiple biometric systems.

The information presented by multiple traits may be consolidated at various levels. At the feature extraction level, the feature sets of multiple modalities are integrated and a new feature set is generated; the new feature set is then used in the matching and decision-making modules of the biometric system. At the matching score level, the matching scores output by multiple matchers are integrated. At the decision level, the final decisions made by the individual systems are consolidated by employing techniques such as majority voting. Fusion at the feature extraction level is expected to perform better than fusion at the other two levels, because there are more information about person identity. However, it is not always feasible for a number of reasons [3]. First, most commercial systems do not provide access to information at this level. Second, the feature spaces of different biometric traits may not be compatible. For example, it is difficult to combine the minutiae feature set of a fingerprint image with the eigen-coefficients of a face image. Third, even if the feature sets were to be compatible, concatenation might result in a feature vector with a very large dimensionality leading to the "curse of dimensionality" problem. Fusion at the decision level is considered to be rigid due to the availability of limited information. In fact, the only type of information available at this level is an "Accept" or a "Reject" label in the verification mode, or the identity of the user in the identification mode.

The dual-tree complex wavelet transform (DT-CWT) is an enhancement of the conventional discrete wavelet transform (DWT) that has gained increasing popularity as a signal processing tool. The transform, originally proposed by Kingsbury [4] to circumvent the shift-variance problem of the decimated DWT, involves two parallel DWT channels with the corresponding wavelets forming approximate Hilbert transform pairs [5, 6]. We refer the reader to the excellent tutorial [6] on the design and application of the DT-CWT.

In this paper we deal with the problem of features extraction fusion level for building a multimodal biometric identification system, using speech and face image modalities. Features of speech and face image are extracted using DT-CWT.

II. DUAL TREE COMPLEX WAVELET TRANSFORMS

DT-CWT is one most promising decomposition that removes drawbacks from others transform like discrete wavelet transform [7], discrete cosine transform [8], and as well as Gabor wavelet transform [9]. Two classical wavelet trees (with real filters) are developed in parallel, with the wavelets forming (approximate) Hilbert pairs. One can then interpret the wavelets in the two trees of the DT-CWT as the real and imaginary parts of some complex wavelet $\psi_c(t)$. The requirement for the dual-tree setting for forming Hilbert transform pairs is the well-known half-sample delay condition. The resulting complex wavelet is then approximately analytic (i.e., approximately one sided in the frequency domain). The design of filter banks satisfying the half sample delay condition can be found in [4-6].

The properties of the DT-CWT can be summarized as:

- Approximate shift invariance;
- Good directional selectivity in 2-dimensions;
- Phase information;
- Perfect reconstruction using short linear-phase filters;

978-0-7695-3982-9/10 $26.00 © 2010 IEEE
DOI 10.1109/ICCEA.2010.196

- Limited redundancy, independent of the number of scales, 2 : 1 for 1-D (2*m* : 1 for *m*D);
- Efficient order-N computation – only twice the simple DWT for 1-D (2*m* times for *m*D).

The transform has the ability to differentiate positive and negative frequencies and produces six sub-bands oriented in $\pm 15^0$, $\pm 45^0$, $\pm 75^0$. However these directions are fixed unlike the Gabor case where the wavelets can be oriented in any desired direction. Complex wavelet transform are inspired by Fourier representation, but with a complex-valued scaling function and complex-valued wavelet [4].

$$\psi_c(t) = \psi_r(t) + j\psi_i(t) \tag{1}$$

Projecting the signal onto $2^{j/2}\psi_c\left(2^j t - n\right)$, we obtain the complex wavelet coefficient:

$$d_c(j,n) = d_r(j,n) + jd_i(j,n) \tag{2}$$

with magnitude:

$$|d_c(j,n)| = \sqrt{[d_r(j,n)]^2 + [d_i(j,n)]^2} \tag{3}$$

and phase:

$$\angle d_c(j,n) = \arctan\left(\frac{d_i(j,n)}{d_r(j,n)}\right) \tag{4}$$

when $|d_c(j, n)| > 0$.

The dual-tree CWT is easy to implement that only employs two real DWTs in parallel, for real part and imaginary part trees as illustrated in Fig. 1 [4].

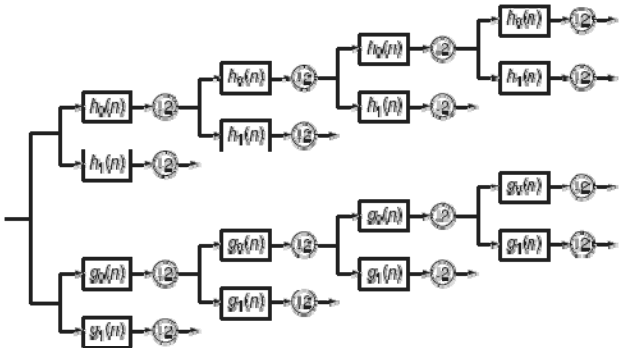

Figure 1. Analysis FB for the dual-tree discrete CWT.

Dual Tree Complex Wavelet Packet Transform (DT-CWPT) is a packet form of the DT-CWT with each of the sub bands should be repeatedly decomposed using low-pass/high-pass Filter Banks (FBs) [18]. The FBs should be chosen so that the response of each branch of the second wavelet packet FB is the discrete Hilbert transform of the corresponding branch of the first wavelet packet FB. Then each sub band of the DT-CWPT will be analytic.

Fig. 2 illustrated of DT-CWPT. H_1, F_1 implements of high-pass filters at every branch, and also H_0, F_0 implements for low-pass filter.

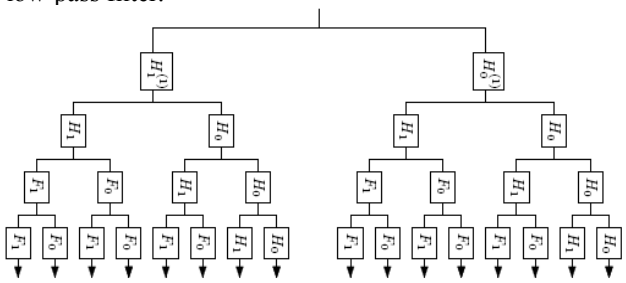

Figure 2. Analysis FB for the DT-CWPT.

We used DT-CWT and DT-CWPT for extraction of feature from face image and speech signals respectively. This is because we must consolidated two coefficient of feature extraction from two modalities in order to get compatibility new feature.

III. PROPOSED METHOD

Feature extraction method and fusion for two features from speech and face image are described below.

A. Face Image feature Extraction

The directional multiscales decomposition of the gray level face image is performed up to level 5. The DT-CWT feature vector X is formed by concatenating the results of the multiscale representation (magnitude of wavelet coefficients). Given an image $I(x,y)$ and a wavelet $\psi_{\mu,v}(x,y)$, of level μ and direction , vector X can be form by:

$$X = \left(O_{0,0} \quad O_{0,1} \quad .... \quad O_{3,5}\right)^t \tag{5}$$

where $O_{\mu,v}(x,y) = I(x,y) * \psi_{\mu,v}(x,y)$ and $O_{\mu,v}$, μ=0,..,4, v=0,1,..,5 is formed by concatenating the rows or columns of $O_{\mu,v}(x,y)$. Here * and t denote, the convolution and transpose operators respectively. This representation encompasses different scales, spatial location and 6 fixed orientations similar to Gabor [20, 21] representation.

The size of such a feature vector for a 112x92 face image is 20748 pixels which is much smaller than the corresponding 2D-DWT feature vector where the size is 59294. In order to reduce the dimensionality of the feature vector to manageable sizes. For the complex wavelets due to the intrinsic downsampling of the multiscale transform, we employed an extra dyadic downsampling strategy to further reduce the size of the feature vector. The feature vectors even after downsampling are of very high dimension and therefore not very convenient to be used directly for recognition. To reduce the dimensionality of the feature

vector space we employed PCA on DT-CWT, 2D-DWT and Eigenface.

Coefficients on different sub bands representing different frequency scales and orientations have different affects on feature extraction, and thus on face recognition. First, coefficients on each sub band are normalized to take magnitude of complex coefficients to balance the effects of different sub bands, especially, the low and high frequencies. Second, those sub bands providing more distinguished information for recognition should be emphasized, here we choose only 24 vector coefficients from decomposition in 6 direction and 5 level. For all of images, the DT-CWT features are aligned to be a vector, and taking principal component for classification.

B. Speech Feature Extraction

According to Siafarikas, et al. [10], the value of the Equal Error Rate (EER) achieved smaller in the two lowest frequency (0 ÷ 500 ÷ 500 Hz and 1000 Hz) which means that most speakers of certain information contained in two frequency bands, as expected.

Frequency range (1500 ÷ 1000 Hz, 2500 Hz ÷ 2000, and 3500 ÷ 3000 Hz) has a lower EER and thus seems to contain more specific information than the frequency speakers (1500 ÷ 2000 Hz, 2500 ÷3000 Hz, and 3500 ÷ 4000 Hz), which shows the worst results. In addition to disclosing specific speakers with different sub-band, our approach to exploit these results in a constructive way to improve the wavelet packet analysis. Having known that wavelet packet level at a depth of 7, 6, and 5, the result is 64, 32, and 16 intervals with a frequency resolution of 31.25, 62.5 Hz, and 125 Hz, respectively, and consider the speaker verification performance analysis each sub-band frequency, we construct a wavelet packet tree coefficient as follows: 32 sub-band resolution of 31.25 Hz (level 7 starting node 3), 16 sub-band resolution of 62.5 Hz (level 6) and 16 sub-bands 125 Hz resolution (level 5), so the whole 64 sub-bands which represent the feature coefficients of each speakers.

This features coefficient was then used to identify the identity of a speaker. Identification system based on unimodal speech sound signal or face image are shown in Fig. 3.

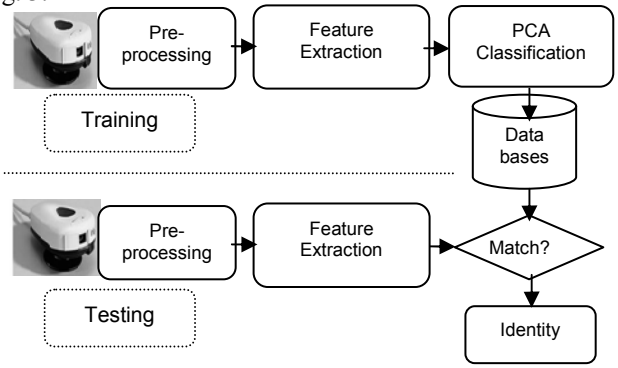

Figure 3. General Person Identification System Block Diagram.

C. Fusion of Face Image and Speech Feature

Vectors coefficient from speech and face image features are consolidated, so the new feature vector is obtained. This new feature vector has dimension 31493 x 1 aligned vector. If we have 20 face image and 8 speech utterance for training, then we have 31493 x 160 feature vector for one person. VidTIMIT database [12] is comprised of audio and video data of 43 people, so we have more dimension to process that database. To reduction this dimension, first we cluster the feature vector of one person to be more visible in calculating (31493 x 32, and 31493 x 64) using Vector Quantization: K-Means Algorithm [22]. Fig. 5 explain of multimodal biometric identification system with feature extraction fusion.

IV. SIMULATION RESULTS AND DISCUSSION

We conduct an experiment using VidTIMIT database [12] and for feature extraction decomposition of two biometric we used DT-CWT Matlab toolbox 4.3 from Kingsbury [15].

A. VidTIMIT Audio-Visual Database

The VidTIMIT database [12], created by Sanderson, is comprised of video and corresponding audio recordings of 43 people (19 female and 24 male), reciting short sentences selected from the NTIMIT corpus. It was recorded in3 sessions, with a mean delay of 7 days between Session 1 and 2, and 6 days between Session 2 and 3.

There are 10 sentences per person. The first six sentences are assigned to Session 1. The next two sentences are assigned to Session 2 with the remaining two to Session 3. The first two sentences for all persons are the same, with the remaining eight generally different for each person. The mean duration of each sentence is 4.25 seconds, or approximately 106 video frames.

The recording was done in a noisy office environment using a broadcast quality digital video camera. The video of each person is stored as a sequence of JPEG images with a resolution of 512 ←384 pixels (columns ←rows); the corresponding audio is stored as a mono, 16 bit, 32 kHz WAV file.

We used 20 face images of session one which represent of most poses of each person for training and 10 face images poses of others session for testing, as an example in Fig. 4. The face image are manually cropped and resize to obtained 125 x 100 pixels (columns ←rows) of face region. For speech signal we used 8 speech utterance from session 1 and 2 for training and 2 utterance from session 2 for testing (text dependent scenario) and also 2 utterance from session 3 for testing (text independent scenario). Speech signal was resample by 8 KHz, and removes silence speech using parametric Voice Activity Detector (VAD) [14].

B. Experiment

After the fusion of two modality features that yields in (31493 x 160 double vector) for each person, we conduct

clustering to reducing the dimension of feature using Vector Quantization: K-Means Algorithm [22] as follow:

1. Find the Mean X
2. Split each centroid to two
3. Assign Each Data to a centroid
4. Find the Centroids
5. Calculate The Total Distance
6. If the Distance has not changed much
   if the number of Centroids is smaller than L2
   Goto Step 2
   else Goto 7
   Else (the Distance has changed substantialy)
   Goto Step 3
7. If the number of Centroids is larger than L
   Discard the Centroid with (highest distortion OR lowest population)
   Goto 3
8. Calculate the Variances and Cluster Weights if required
9. End

Where:

X: a matrix each column of which is a data vector

L: codebook size (preferably a power of 2)

Figure 4. Example face image from VidTIMIT databases.

Second reduction of dimension, PCA based is performed as follows. A given size feature vector (31493 x 64 double vector) is represented by a matrix containing feature coefficients values; the matrix is then converted to a vector, $\vec{v}$, by concatenating all the columns; a D-dimensional feature vector $\vec{x}$, is then obtained by:

$$\vec{x} = U^T\left(\vec{v} - \vec{v}_\mu\right) \quad (6)$$

where U contains D eigenvectors (corresponding to the D largest eigenvalues) of the training data covariance matrix, and $\vec{v}_\mu$ is the mean of training vectors. In our experiments we use training features from all clients, find U and $\vec{v}_\mu$; moreover, D = 860. Preliminary experiments showed that while D = 860 obtained optimal identification rate.

We also conduct the experiment using a unimodal [16, 17, 19] (speech or face image) based person identification systems as depict in Fig. 3. The results are showed in Table I for speaker identification systems using YOHO [23] and VidTIMIT database and in Table II for face image identification systems.

Identification rate (IR) are calculated using Euclidean distance:

$$\%\,IR = \frac{true\ identify}{total\ number\ of\ testing\ data} \times 100\,\% \quad (7)$$

TABLE I. SPEAKER IDENTIFICATION SYSTEMS

Method	YOHO (Text Independent)	VidTIMIT (Text Independent)	VidTIMIT (Text Dependent)
MFCC	65%	60%	90%
DT-CWPT	70%	65%	93%

TABLE II. FACE IMAGE IDENTIFICATION SYSTEMS.

Method	Identification rate
Eigenface PCA	70%
2D-DWT PCA	75%
DT-CWT PCA	91%

Table III show the experiment result from multimodal identification system using VidTIMIT database, with 32 clusters and 64 clusters (an optimal cluster) that could be arranged for experiment using Matlab.

TABLE III. SPEAKER IDENTIFICATION SYSTEMS

Method	VidTIMIT	
	Face Image Feature + Speech feature (Text Independent)	Face Image Feature + Speech feature (Text Dependent)
32 cluster	87%	90%
64 cluster	90%	93,7%

C. Discussion

The experiment described above showing that using unimodal biometric with DT-CWT and DT-CWPT feature extraction, identification rate are increased ± 15 % compared with others. Identification rate using multimodal are also increased average ± 4 % compared with unimodal biometric identification systems.

V. CONCLUSION AND FUTURE WORK

This paper provides initial results obtained on a multimodal biometric identification system that uses DT-

CWT face, and DT-CWPT speech features for biometric identification purposes. Our experiments indicate that clustering algorithm and PCA performs better to reduce dimension of multimodal feature for fusion at feature extraction level.

We will conduct an experiment using others multimodal database, that could show more comprehensive performance of identification system in fusion at feature extraction level.

ACKNOWLEDGMENT

The author thanks to Direktorat Jenderal Pendidikan Tinggi (Directorate-General of Higher Education), Ministry of National Education for supporting of this research.

REFERENCES

[1] A. Ross and A. K. Jain, "Information Fusion in Biometrics," Pattern Recognition Letters, Vol. 24, No. 13, pp. 2115-2125, September 2003.

[2] Robert Snelick, Mike Indovina, James Yen, and Alan Mink, "Multimodal Biometrics: Issues in Design and Testing," to appear in Proc. of International Conference on Multimodal Interfaces, Vancouver, B.C., November 5-7, 2003.

[3] A. K. Jain and Arun Ross, "Multibiometric Systems," to appear in Communications of the ACM, Special Issue on Multimodal Interfaces, January 2004.

[4] N. G. Kingsbury, "Complex wavelets for shift invariant analysis and filtering of signals," Journal of Applied and Computational Harmonic Analysis, vol. 10, no. 3, pp. 234–253, May 2001.

[5] I. W. Selesnick, "Hilbert transform pairs of wavelet bases," IEEE Signal Process. Lett., vol. 8, no. 6, pp. 170–173,2001.

[6] I.W. Selesnick, R.G. Baraniuk, and N.G. Kingsbury, "The dual-tree complex wavelet transform A coherent framework for multiscale signal and image processing," IEEE Signal Processing Magazine, 22(6):123–151, November 2005.

[7] L. Shen Z. Ji, L. Bai, and C. Xu, "DWT based HMM for face recognition," *Journal of Electronics*, vol. 24, no. 6, pp. 835-837, November 2007.

[8] A. Manjunath, R. Chellappa, and C. von der Malsburg, "A feature based approach to face recognition," in Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR '92), pp. 373-378, Champaign, IL, USA, June 1992.

[9] L. Shen, L. Bai, and M. Fairhurst, "Gabor wavelets and general discriminant analysis for face identification and verification," *Pattern recognition Letters,* vol. 25, no. 5, pp. 553-563, May 200.

[10] M. Siafarikas, T. Ganchev, N. Fakotakis, "Objective Wavelet Packet Features for Speaker Verification," 2004.

[11] J. Campbell, and A. Higgins, "*YOHO* Speaker Verification," Linguistic Data Consortium, Philadelphia, 1994.

[12] C. Sanderson, "Biometric Person Recognition: Face, Speech and Fusion," VDM-Verlag, 2008. ISBN 978-3-639-02769-3.

[13] I.W. Selesnick, R.G. Baraniuk, and N.G. Kingsbury, "The dual-tree complex wavelet transform A coherent framework for multiscale signal and image processing," IEEE Signal Processing Magazine, 22(6):123–151, November 2005.

[14] F.O. Olutope, "SpeechCore Matlab program," London Metropolitan University, 2007.
http://www.mathworks.com/matlabcentral/fileexchange/19298-speechcore

[15] N. G. Kingsbury, "Dual-Tree Complex Wavelet Transform Matlab Toolbox, Pack - version 4.3," Cambridge University, June 2003.

[16] Gunawan Sugiarta,YB., Riyanto, B., Hendrawan, Suhardi, "Dual Tree Complex Wavelet Transform for Face Image Identification," International Conference on Rural Information and Communication Technology, 2009, 413-417.

[17] Gunawan Sugiarta,YB., Riyanto, B., Hendrawan, Suhardi, "Metoda Ekstraksi Ciri untuk Identifikasi Penutur menggunakan Dual Tree Complex Wavelet Packet Transform," The 7th National Conference on Design and Application of Technology 2008, 119-124.

[18] Gunawan Sugiarta,YB., Riyanto, B., Hendrawan, Suhardi, "KOEFISIEN DUAL TREE COMPLEX WAVELET PACKET TRANSFORM SEBAGAI CIRI UNTUK SISTEM IDENTIFIKASI PENUTUR," Seminar Nasional Ilmu Komputer dan Aplikasinya – SNIKA 200 (26/11/2009) ISSN 1907-882X, in press.

[19] İ. Bayram, and I. W. Selesnick, "On the Dual-Tree Complex Wavelet Packet and *M*-Band Transforms," IEEE Trans. Signal Processing, 56(6) : 2298-2310, June 2008.

[20] X. Xie, and K. Lam, "Gabor-based kernel PCA with doubly nonlinear mapping for face recognition with a single face image", *IEEE Trans. Image Process.*, 06(9):2481–2492, September 2006.

[21] C. Liu., "Gabor-based kernel PCA with fractional power polynomial models for face recognition", *IEEE Trans. Pattern Anal. Mach. Intell.*, 04(5):572–581, May 2004.

[22] E. Zavarehei, "K-Means Algorithm with Spliting Method for Training," Brunel University, May-2006.
http://www.mathworks.ch/matlabcentral/fileexchange/10943

[23] J. Campbell, and A. Higgins, "*YOHO* Speaker Verification," Linguistic Data Consortium, Philadelphia, 1994.

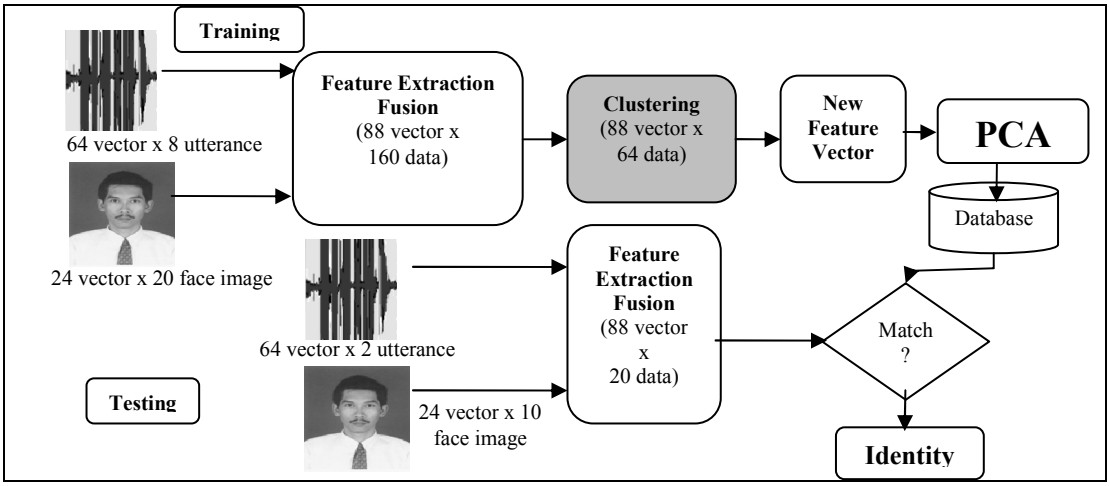

Figure 5. Multimodal Biometric Identification System with Feature Extraction Fusion.