MULTI-LEVEL LIVENESS VERIFICATION FOR FACE-VOICE BIOMETRIC AUTHENTICATION

Girija Chetty and Michael Wagner

School of Information Sciences and Engineering, University of Canberra, Australia
girija.chetty@canberra.edu.au

ABSTRACT

In this paper we present the details of the multilevel liveness verification (MLLV) framework proposed for realizing a secure face-voice biometric authentication system that can thwart different types of audio and video replay attacks. The proposed MLLV framework based on novel feature extraction and multimodal fusion approaches, uncovers the static and dynamic relationship between voice and face information from speaking faces, and allows multiple levels of security. Experiments with three different speaking corpora VidTIMIT, UCBN and AVOZES shows a significant improvement in system performance in terms of DET curves and equal error rates(EER) for different types of replay and synthesis attacks.

1. INTRODUCTION

Despite increased use of computers and the Internet in most of the day-to-day activities, traditional person authentication and identity verification methods based on PINs, passwords and tokens have met with limited success. This could be due to several reasons, such as difficulty in remembering several PINs and passwords, and vulnerability of token-based methods of identity verification like passports and driver's licenses to forgery, theft, or loss. By using biometric traits on the other hand, which involves confirming or denying a person's claimed identity based on his/her physiological or behavioral characteristics, such as person's fingerprint or voice, some of the problems with PIN and password based systems can be addressed [1]. However, all biometric traits do not enjoy the same level of user-acceptance, because of traditional usage of biometrics for criminal identification and anti-terrorism measures.

According to an EU study, by the year 2015, there will be an enormous diffusion of biometric based identity verification in day-to-day civilian and e-commerce applications [2]. Hence there is a need to investigate the usage of biometric traits that have better user-acceptance and does not have a social stigma associated with it. Face and/or voice based biometric systems rate high in terms of better user acceptance due to less intrusiveness, and lower deployment costs due to ease of availability of low-cost off-the shelf system components. However, the very use of user-friendly face and voice biometric traits exposes the system to different types of fraudulent attacks, involving audio and/or video playback, or surreptitious replay of client-specific biometric information by an impostor (pre-recorded audio or still-photo).

Various studies have indicated no single modality can provide an adequate solution against impostor attacks [1,3]. The use of multiple modalities such as acoustic and visual speech can overcome the limitations of techniques based on a single modality. The visual manifestation of speech in speaking faces for example, in terms of synchronous speech signal and the facial movements associated with an utterance provide powerful cues for robust authentication approaches. The correlation between the visual and acoustic speech components is exhibited by the human ability to "make-up" for the loss in the acoustic component using visual information, and can be used for both purposes, that is verifications of person's identity as well as "liveness". Liveness verification in a biometric system means the capability to detect and verify, whether or not the biometric sample presented is alive, and from the right person, during training/enrolment and testing phases. Until now, although there has been some published research on the liveness, for example, of fingerprints [4], research on liveness verification in face-voice bases person authentication systems has been very limited. We propose that for biometric systems based on face-voice biometric traits, novel multimodal feature extraction and fusion techniques that uncover the static and dynamic relationship between face-and voice information from a speaking face, allow verification of liveness and hence equip the system with multiple levels of security to thwart different types of audio/video replay attacks.

The next section describes the proposed multi-level liveness verification framework (MLLV), followed by speaking face data used for different experiments in section 3. The description of different approaches used for modeling speaking faces is given in section 4, followed by some experimental results and conclusions in section 5 and 6.

1-4244-0487-8/06/$20.00 ©2006 IEEE

2. THE MLLV FRAMEWORK

The MLLV framework comprises several approaches for verifying the liveness in speaking faces, with novel feature extraction and fusion approaches namely - bimodal feature fusion (BMF), cross-modal fusion (CMF), and 3D multimodal fusion (3MF). The BMF approach offers Level-1 security and verifies liveness based on modeling speakers with features in a subspace that preserves face-voice synchrony. At this level system detects replay attacks involving still photo and/or pre-recorded audio. The system can however be cheated with video-replay and synthesis attacks. The CMF approach offers Level-2 security and performs liveness detection based on modeling the speaker with features in cross-modal space, that extract hidden face-voice synchrony information during speech production. The system can hence detect video replay attacks, involving pre-recorded video playback in addition to still-photo replay attacks. The system can however be cheated with synthesis attacks with 3D synthetic talking heads. Level-3 security is achieved with the 3MF approach which performs liveness checks based on modeling the speaker in 3D space with 3D shape and texture features. The speaking faces modeled with the three - BMF, CMF and E3MF approach allow detection and verification of liveness at an increasing detail, and thus offer enhancement in security-in multiple levels to combat different types of fraudulent attacks. Moreover, it allows system implementation as a complete software-based solution based on novel feature extraction and fusion techniques without any special hardware requirements such as the use of high-cost thermal or infra-red cameras [5], or demanding user requirements in terms of strong user co-operation required in challenge-response systems [6].

3. SPEAKING FACE CORPORA

The speaking face data from three different corpora, VidTIMIT, UCBN and AVOZES was used for conducting replay and synthesis attack experiments. The VidTIMIT multimodal person authentication database [7] consists of video and corresponding audio recordings of 43 people (19 female and 24 male). The mean duration of each sentence is around 4 seconds, or approximately 100 video frames. A broadcast quality digital video camera in a noisy office environment was used to record the data. The video of each person is stored as a sequence of JPEG images with a resolution of 512×384 pixels with corresponding audio provided as a 16-bit 32-kHz mono PCM file.

The second type of data used is the UCBN database, a free to air broadcast news database. The broadcast news is a continuous source of video sequences that can be easily

Figure 2: Faces from (a) VidTIMIT, (b) UCBN, (c) AVOZES

obtained or recorded, and have optimal illumination, colour, and sound recording conditions. However, some of the attributes of broadcast news database such as near-frontal images, smaller facial regions, multiple faces and complex backgrounds require an efficient face detection and tracking scheme to be used. The database consists of 20-40 second video clips for anchor persons and newsreaders with frontal/near-frontal shots of 10 different faces (5 female and 5 male). Each video sequence is 25 frames per second MPEG2 encoded stream with a resolution of 720 × 576 pixels, with corresponding 16 bit, 48 kHz PCM audio.

The third database used is the AVOZES database, an audiovisual corpus developed for automatic speech recognition research [8]. The corpus consists of 20 native speakers of Australian English (10 female and 10 male speakers), and the audiovisual data was recorded with a stereo camera system to achieve more accurate 3D measurements on the face. The recordings were made at 30 Hz video frame rate and 16bit 48 kHz mono audio rate in a controlled acoustic environment with no external noise, and some background computer and air-conditioning noise. For each speaker there were 3 spoken utterances, 10 digit sequences, 18 phoneme sequences (CVC words in a carrier phrase), and 22 VCV phoneme sequences(VCV words in a carrier phrase.

Figure 1a, 1b and 1c, show sample faces from VidTIMIT, UCBN and AVOZES corpus. The three types of corpora represent very different types of speaking face data, VidTIMIT with original audio recorded in a noisy environment and clean visual environment, UCBN with clean audio and visual environments, but complex visual backgrounds, and AVOZES with stereo face data for better 3D face modeling. The VidTIMIT corpus has profile face view (side-view) of each person, in addition to frontal face video sequences.

4. MULTIMODAL FUSION APPROACHES FOR MLLV FRAMEWORK

A speaking face is a kinematic-acoustic system in motion, and the shape, texture and acoustic features during speech production are correlated in several ways as reported in a

1-4244-0487-8/06/$20.00 ©2006 IEEE

number of studies carried out by Yehia et.al. [9,10]. This correlation is based on anatomical facts, with single neuromotor source controlling the vocal tract behavior, which in turn is responsible for both the acoustic and the visible attributes of speech production. Hence, for a speaking face the facial motion and speech acoustics are correlated in many ways, and this person-specific anatomical visual speech relationship cannot be imitated or faked easily.

The proposed MLLV framework is based on extraction this correlation in several representative subspaces, and involves different types of feature extraction and fusion techniques that uncover the static and dynamic relationship during speech production A brief description of each type of approach used is given here.

4.1. Bimodal Feature Fusion (BMF)

The classical approaches to multimodal fusion are based on late fusion and its variants, and has been investigated in great depth [11,12]. Late fusion approach involves combining the scores of different classifiers, each of which has made an independent decision. For extracting information from speaking faces however, this means, that many of correlation properties of the joint audio-video data are lost. Fusion at feature-level on the other hand, can substantially improve the modeling of liveness in the speaker as the joint feature sets provide a richer source of information than the matching scores, and because in this mode, features are extracted from the raw data and subsequently combined. In addition, feature-level fusion allows synchronisation between closely coupled modalities for a speaking face, such as voice and lip movements to be preserved throughout various stages of authentication. The BMF approach is based on feature fusion of voice and lip-region features from speaking face video sequences.

The audio and visual features corresponding to lip-region were extracted from each frame of the video clip, and the joint audio-visual feature vector was formed by direct concatenation of acoustic and visual features from the lip-region. The acoustic features used were Mel Frequency Cepstral Coefficients (MFCC) derived from cepstrum information. The pre-emphasized audio signal was processed using a 30ms Hamming window with one-third overlap, yielding a frame rate of 50 Hz, to obtain the MFCC acoustic vectors. An acoustic feature vector was determined for each frame by warping 512 spectral bands into 30 mel-spaced bands, and computing the 8 MFCCs. Cepstral mean normalization was performed on all MFCCs before they were used for training, testing and evaluation.

The visual features used were geometric and eigen-lip features from lip-region of faces in the video sequence. Before the lip-region features were extracted, faces were recognised based on well known eigen-faces technique [13]. Before recognition faces were detected based on an approach involving skin colour analysis in red-blue chrominance colour space, followed by deformable template matching with an average face, and finally verification with rules derived from the spatial/geometrical relationships of facial components. The lip region was determined using derivatives of hue and saturation functions, combined with geometric constraints.

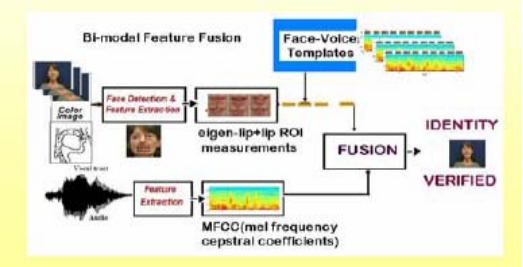

Figure 3: Bi-modal Feature Fusion (BMF) approach

Figure 3 shows various processing stages of BMF approach. The audio-visual fusion vector was obtained by direct concatenation of the audio features (MFFCs-8) and visual features (eigen-lip projections (10) + lip dimensions (6)), and the combined feature vector was used for building the speaker models. The audio features acquired at 50 Hz, and the visual features acquired at 25Hz were appropriately rate interpolated to obtain synchronized joint audiovisual feature vectors. A Gaussian mixture model (GMM) is then trained for each client using the synchronized joint audio-visual fusion vectors.

4.2 Cross-modal Fusion (CMF)

For modeling speaking faces using cross-modal fusion (CMF) approach, face-voice features obtained from BMF approach are transformed in a new cross-modal space. The cross-modal features were based on latent semantic analysis (LSA) involving singular value decomposition of joint face-voice feature space, and canonical correlation analysis (CCA), based on optimising cross-correlations in a rotated audio-visual subspace.

LSA is a powerful tool used in text information retrieval to discover underlying semantic relationships between different textual units [14]. The LSA technique achieves three goals: dimension reduction, noise removal and the uncovering of the semantic and hidden relation between different objects such as keywords and documents. In our current context, we used LSA to uncover the synchronism between image and audio features in a video sequence.

CCA, an equally powerful multivariate statistical technique, attempts to find a linear mapping that maximizes the cross-correlation between two features sets [15]. It finds the transformation that can best represent (or identify) the coupled patterns between features of two different subsets. A set of linear basis functions, having a direct relation to maximum mutual information, is obtained in each signal space, such that the correlation matrix between the signals

described in the new basis is diagonal [15]. A subset of vectors containing the first k pairs defines a linear *rank-k* relation between the sets that is optimal in a correlation sense. t has been shown that finding the canonical correlations is equivalent to maximizing the mutual information between the sets if the underlying distributions are elliptically symmetric [15]. Figure 4 shows the processing stages for cross-modal feature extraction.

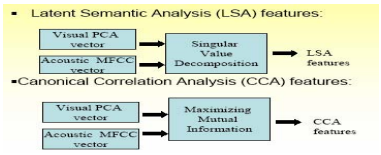

Figure 4: Cross Modal Fusion (BMF) approach

The LSA and CCA features are computed from low-level visual and audio features. The visual features are 20 PCA (Eigen lip) coefficients, and the audio features are 12 MFCC coefficients. Based on preliminary experiments, fewer than 8-10 LSA/CCA features were found to be sufficient to model more than 95% variations in speaking faces. A Gaussian mixture model (GMM) is then trained for each speaker using the cross-modal fusion vector obtained from direct concatenation of LSA and CCA features. The optimal LSA/CCA features with their reduced dimensions compared with the 32-dimensional bimodal feature fusion vector with 20 PCA and 12 MFCC vectors, allows replay attacks of higher complexity to be detected.

4.2 3D multimodal fusion (3MF)

In this approach, a speaking face is modeled with features comprising 3D shape and texture features directly concatenated with MFCC acoustic features. The 3D shape and texture features were obtained by pre-processing of facial images from multiple views, involving the stages of face detection and facial feature extraction, followed by creation of 3D face models, and subsequent extraction of shape and texture features corresponding to lip region. Since each speaking corpus consists of different types of face data, the 3D face modeling technique used was different based on methods reported in [16,17]. The VidTIMIT data base for example, consists of frontal and profile view images of the faces, AVOZES data comprises left and right stereo faces, and UCBN corpus comprises near frontal face images. For VidTIMIT and AVOZES faces, the 3D face modeling algorithm starts by computing 3D coordinates of automatically extracted facial feature points. Correspondence between feature points in both images was established using epipolar constraints, and then depth information from front and profile views for VidTIMIT faces, and, left and right views for AVOZES faces, was computed using perspective projection. The 3D coordinates of the selected feature points are then used to deform a 3D generic face model to obtain a person specific 3D face model. Figure 5 shows the sample 3D face model developed from a face in VidTIMIT corpus.

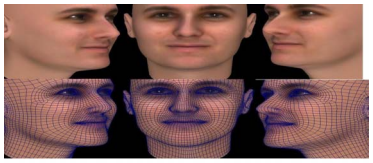

Figure 5: 3D face model for a VidTIMIT face image

The major deformations for the speaking face are in the lip-region of the face. The lip-region of the face was modeled using 36 vertices and 20 surfaces. From initial experiments, with principal component analysis (PCA) of the 3D shape vector and the texture vector for lip-region , we learnt that about 6-8 principal components of shape vector and 3-4 components of texture vector explains more than 95% of variations in lip shapes and appearances in visemes corresponding to English language. The 8 eigen-values for shape vector correspond to jaw opening/closing, lip protrusion/retraction, lip opening/closing, and jaw protrusion/retraction as shown in Figure 6.

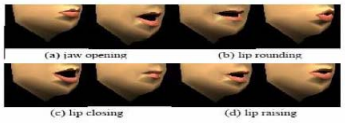

Figure 6: Principal Visemes in English language

Similarly, the 3-4 Eigen values of texture vector describe most of the appearance variations mainly those corresponding to one rounded viseme with closed lips, (e.g. [‘u’]), one rounded viseme with open lips, and one spread viseme with spread lips, (e.g. [‘i’]). The 18-dimenaional audio-visual feature vector for 3D multimodal fusion module was constructed by concatenating 8 MFCCs + 1 F0 feature, 6 eigen-shape and 3 eigen-texture features.

5. LIVENESS VERIFICATION EXPERIMENTS

To investigate the performance of face-voice fusion approaches in MLLV framework, different sets of liveness verification experiments were conducted consisting of training and test phases. In the training phase, a 10-mixture Gaussian mixture model of each client’s face voice fusion vector for each approach was built by constructing a gender-specific universal background model (UBM) and then adapting each UBM by MAP adaptation [19]. In the test phase, clients’ live test recordings were evaluated against a client’s model λ by determining the log-likelihoods *(log* $p(X|\lambda)$) of the time sequences X of audiovisual feature vectors. For testing replay attacks, three types of replay/synthetic attack experiments were conducted. For

still-photo replay attacks, a number of "fake" recordings were constructed by combining the sequence of audio feature vectors from each test utterance with ONE visual feature vector chosen from the sequence of visual feature vectors. Such a fake sequence represents an attack on the authentication system, which is carried out by replaying an audio recording of the client's utterance while presenting a still photograph to the camera. Four such fake audiovisual sequences were constructed from different still frames of each client test recording. Log-likelihoods *(log $p(X'|\lambda)$)* were computed for the fake sequences X' of audiovisual feature vectors against the client model λ.

For *video-replay* attacks, "fake" video clip was constructed by reducing the resolution of face images in the video sequence and introducing a random shift (delay) in the face frames (8-10 frames) relative to acoustic frames. This scenario simulates *video-replay* attacks where an impostor has surreptitiously acquired client's video recording, and synchrony between face and voice in the fake sequence is not the same as original speaking face video sequence of the client. The *video-replay* attacks also included in-synch scenarios, where the original synchrony information between face and voice frames were kept intact, no artificial shift or delay was introduced. For third type of attacks, *synthetic-replay* attacks, synthetic video clip was constructed from the still photo of each speaker. This represents a scenario of a replay attack, with an impostor presenting a fake video clip constructed from pre-recorded audio and a still photo of the client animated with facial movements and voice-synchronous lip movements. We constructed several fake video clips by extracting faces corresponding to few key-frames from the original video sequence, animated the corresponding lip-regions by phoneme-to-viseme mapping, and then added random deformations and movements in the face and finally rendered lip and face movements with speech, all together as a new video clip. The synthesized fake clip visually emulates a normal speaking face with certain facial and head movements in synchronism with spoken utterance. Performance in terms of DET curves and EER rates was examined with different subsets of data corpus such as text-dependent/text-independent and male/ female only cohorts. The results of liveness verification experiments for each fusion approach of MLLV framework is described next.

6. RESULTS AND DISCUSSION

The performance of MLLV framework was examined for different replay-attack scenarios with three scenarios testing each fusion approach with different corpus. As described in section 5, the three scenarios were *still-photo* replay attacks, *video-replay* attacks, and *synthetic-replay* attacks For *still-photo* and *video-replay* attack scenarios, the speaking face data was from VidTIMIT and UCBN corpus with pre-normalization of pose and illumination. The first two utterances for all speakers in the VidTIMIT corpus being common were used for text dependent experiments and 6 different utterances for text independent verification experiments. For UCBN, the training data for both text dependent and text independent experiments contained 15 utterances from 5 male and 5 female speakers, and 5 utterances for testing, each recorded in a different session.

For *synthetic-replay* attack scenario, AVOZES corpus was also used in addition to VidTIMIT and UCBN. The average performance in terms of EERs and DET curves for the three scenarios are shown in Table 2 and Figure 7-9.

Scenarios vs. Approaches	Still-photo Attacks	Video replay attacks	Synthetic attacks
BMF	2.4 %	6.54 %	9.23 %
CMF	0.29%	2.25%	3.96%
3MF	0.0155 %	.611%	1.18%

Table 1: EERs for different scenarios of MLLV framework

As can be seen from Table 1 and Figure 7, with BMF approach, system can address *still-photo* attacks reasonably well (2.4% EER), however performs poorly for video replay attacks (6.54 %) and synthesis attacks (9.23 %). This is level-1 security offered by the MLLV framework. With CMF approach, the system performs well for *still-photo* attacks with EER of 0.229%, and *video-replay* attacks (2.25% EER), but fails to perform for *synthesis* attacks... With 3MF approach, system can detect almost all still-photo and video replay attacks with EERs of 0.0155% and 0.611 % respectively, and can also address *synthetic* attacks very well (1.18%). Thus three approaches in the MLLV framework provide increasing levels of security and can address increasingly complex replay attacks. It should be noted that the EER figures reported here are average figures for all subsets of data in three corpora, and the performance of each subset in each corpus has a variation of around 2-3 standard deviations around the reported EERs.

The performance of *in-synch* video replay attacks drops by about 27-42 %, a deterioration of 42% for BMF approach, followed by 36% for CMF approach, and around 27% for 3MF approach. The performance for text-dependent experiments was better for UCBN corpus and AVOZES corpus as compared to text-independent experiments. For VidTIMIT corpus however, the text-independent male-only cohort had best performance compared to text-dependent male and female only cohorts. This is due to availability of large speaking face data in this subset. The computational cost of each approach given by $O(mn)$, where m is the size of speaker model in terms of audio-visual fusion vector size, and n is the original size of lip-region image and voice samples is different for each approach. The computation cost is least for CMF approach, increases by around 10% for BMF approach, and around 35% for 3MF approach. Also, there was a significant enhancement in performance for detecting impostors – *the primary function* of the system, in

1-4244-0487-8/06/$20.00 ©2006 IEEE

addition to detecting replay attacks. An improvement of the order of 18-37% was achieved as compared to corresponding baseline late-fusion and single-modal features. The EER values of less than 2% was achieved for 3MF and CMF approach, and less than 3.5% for BMF approach for impostor attack experiments. Further experiments are in progress to compare the improvements in terms of new performance measures such as genuine accept rates vs. false accept rates in addition to conventional EERs and DET curves.

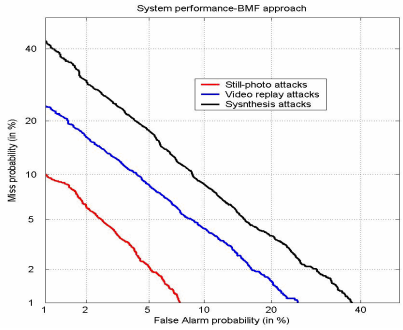

Figure 7: DET curves for BMF approach

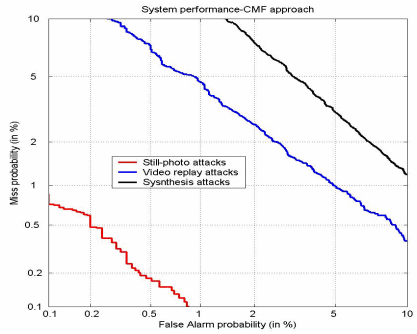

Figure 8: DET curves for CMF approach

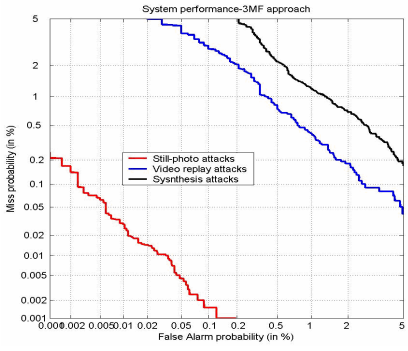

Figure 9: DET curves for 3MF approach

7. CONCLUSION

In this paper we show the potential of proposed MLLV framework for thwarting different types of replay attacks on face-voice biometric authentication, the frame-work based on several new features and fusion techniques verifies the liveness in the speaking face data at multiple levels by extracting static and dynamic relationship in face-voice information in novel subspaces.

8. REFERENCES

[1] A. Ross and A. K. Jain, "Multimodal Biometrics: An Overview", Proc. of 12th European Signal Processing Conference (EUSIPCO), (Vienna, Austria), pp. 1221-1224, September 2004.

[2] Ioannis Maghiros et. al, "Biometrics at the Frontiers: Assessing the impact on Society", Technical Report Series, http://europa.eu.int/comm/justice_home/doc_centre/freetravel/doc/biometrics_eur21585_en.pdf.

[3] Poh, N., and J. Korczak, "Hybrid biometric person authentication using face and voice features," Proc. of Int. Conf. on Audio and Video-Based Biometric Person Authentication, Halmstad, Sweden, June 2001,pp.348-353.

[4] Sujan T. V. Parthasaradhi, Reza Derakhshani, Larry A. Hornak, and Stephanie A. C. Schuckers: "Time-series detection of perspiration as a liveness test in fingerprint devices". IEEE Transactions on Systems, Man, and Cybernetics, Part C 35(3): 335-343 (2005).

[5] Friedrich Gil et. al.; "Seeing People in the Dark: Face Recognition in Infrared Images"; http://www.math.tau.ac.il/~hezy/papers/c27.PDF

[6] Frischholz Robert et.al ; "Avoiding Replay-Attacks in a Face Recognition System using Head-Pose Estimation" in AMFG 2003: IEEE International Workshop on Analysis and Modeling of Faces and Gestures, October 17, 2003, Nice, France.

[7] Sanderson, C. and K.K. Paliwal, "Fast features for face authentication under illumination direction changes", Pattern Recognition Letters 24, 2409-2419, 2003.

[8] Goecke, R., and J.B. Millar, "The Audio-Video Australian English Speech Data Corpus AVOZES", Proceedings of the 8th International Conference on Spoken Language Processing INTERSPEECH 2004 - ICSLP, Volume III, pages 2525-2528, 4-8 October 2004.

[9] Yehia, H., Rubin, P. and Vatikiotic-Bateson E. (1998), "Quantitative association of vocal tract and facial behavior", Journal of Speech Communication 26(1-2), 23-43.

[10] Yehia Hani, Takaaki Kuratate, Eric Vatikiotic-Bateson, "Linking Facial Animation, Head Motion and Speech Acoustics", Journal of Phonetics, Vol.30, Issue 3, 2002.

[11] Kittler, J., G. Matas, K. Jonsson, and M. Sanchez,"Combining evidence in personal identity verification systems," Pattern Recognition Letters, vol.18, no.9,pp.845–852,Sept. 1997.

[12] Poh, N., and J. Korczak, "Hybrid biometric person authentication using face and voice features," Proc. of Int. Conf. on Audio and Video-Based Biometric Person Authentication, Halmstad, Sweden, June 2001, pp. 348--353.

[13] M. Turk and A. Pentland, "Eigenfaces for recognition," Journal of Cognitive Neuroscience, Vol. 3, No. 1, pp. 71-86, Winter 1991.

[14] Deerwester, S., Dumais, S.T., Frunas, G.W., Landauer, T.K., and Harshman, R. Indexing by Latent Semantic Analysis, Journal American Society for Information Sci.,2001, 41(6),391-407.

[15] M. Borga, H. Knutsson, "An Adaptive Stereo Algorithm Based on Canonical Correlation Analysis" ICIPS 1998, Gold Coast, Austalia August 1998

[16] Blanz, V. and Vetter, T.; "A Morphable Model for the Synthesis of 3D Faces", SIGGRAPH'99 Conference Proceedings Component-Based Face recognition with 3D Morphable Models,

[17] Hsu, R.L. and A.K.Jain, "Face Modeling for Recognition," Proceedings Int'l Conf. Image Processing, ICIP, Greece, Oct. 7- 10, 2001.