ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/4219276

Audio-Video Person Authentication Based on 3D Facial Feature Warping

Conference Paper · January 2006

DOI: 10.1109/DICTA.2005.13 · Source: IEEE Xplore

CITATIONS
0

READS
49

2 authors, including:

Girija Chetty
University of Canberra
141 PUBLICATIONS **766** CITATIONS

SEE PROFILE

Some of the authors of this publication are also working on these related projects:

Surgical Simulation and Assistance Technologies View project

Multimodal Systems View project

All content following this page was uploaded by Girija Chetty on 16 May 2014.

The user has requested enhancement of the downloaded file.

Audio-Video Person Authentication based on 3D Facial Feature Warping

Girija Chetty and Michael Wagner
HCC Laboratory, School of ISE, University of Canberra, Australia
girija.chetty@canberra.edu.au

Abstract

In this paper we propose a novel feature warping technique based on thin-plate-spline (TPS) analysis for 3D audio-video person authentication systems. The TPS warp features model information related to non-rigid variations on speaking faces, such as expression lines, gestures, and wrinkles, enhancing the performance of the system against impostor and spoof attacks. Experiments with multimodal fusion of acoustic and TPS shape features for two different speaking face data corpus, VidTIMIT and AVOZES, allowed equal error rates (EERs) of less than 0.5 % for imposter attacks, less than 1 % for type-1 replay attacks (still photo and pre-recorded audio) and less than 2% for more complex type-2 replay attacks (pre-recorded video or fake CG animated video).

1. Introduction

Current audio-video person authentication systems based on 2D face models can achieve satisfactory performance in highly constrained environments, and encounter difficulties in handling large amounts of facial variations due to head pose, lighting conditions and facial expressions [1]. Because the human face is a three-dimensional (3D) surface, and is a detailed internal anatomical structure, instead of just the external appearance, utilizing 3D face information should improve the performance of the system against pose, illumination and expression variations [2]. By including voice information in addition to 3D face models, audio-video biometric systems can be made less vulnerable to different types of imposter and replay attacks. This is because of differential difficulty in spoofing a person's voice, in synchronism with 3D shape and texture of a person's face [3, 4]. However, certain subtle and non-rigid variations on speaking faces due to variations in expression lines, gesture, and wrinkles while talking, cannot be modeled by methods that simply extract the rigid 3D shape and texture information. Modeling of subtle and non-rigid facial variations allows liveness verification and can lead to substantial reduction in impostor and spoof attacks, as it is almost impossible to imitate such fine details of a human face, and it can be ensured that the person trying to access a facility is an authorized "live" person and not an impostor or a fake client.

In this paper, novel feature warping of 3D facial shape features based on thin plate spline(TPS) analysis is proposed, allowing robustness to pose, illumination and expression variations, leading to significant enhancement in performance of audio-video biometric systems against imposter and replay attacks. An equal error rate (EER) of less than 0.5 % was achieved for imposter attacks, less than 1% for *type-1* replay attacks (pre-recorded audio and still photo) and less than 2% for *type-2* replay attacks(pre-recorded video or fake speaking faces created with CG animation and other similar techniques). The TPS warp features extract non-rigid deformations such as wrinkles and expression lines, and other subtle facial gestures on a 3D speaking face.

The speaking face corpus used for examining the performance of the proposed technique is described in next section. The detail of 3D face modeling technique used is given in section 3, followed by description of TPS warp features in section 4. The details of impostor and replay attack experiments is described in section 5, with conclusions in section 6.

2. Speaking Face Data Corpus

The speaking face data from two different data corpus, VidTIMIT and AVOZES was used for conducting imposter and spoof attack experiments. The VidTIMIT multimodal person authentication database [4] 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

Proceedings of the Digital Imaging Computing: Techniques and Applications (DICTA 2005)
0-7695-2467-2/05 $20.00 © 2005 IEEE

IEEE COMPUTER SOCIETY

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 database used is the AVOZES database, an audiovisual corpus developed for automatic speech recognition research [5]. 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.

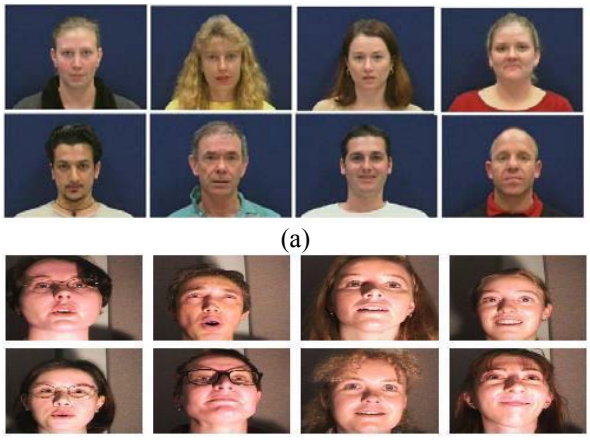


Figure 1: Faces from (a) VidTIMIT, (b) AVOZES

Figure 1a and 1b show sample data from VidTIMIT and AVOZES corpus. The two types of databases represent very different types of speaking face data, VidTIMIT with original audio recorded in a noisy environment and clean visual environment, and AVOZES with stereo face data for better 3D face modeling. The technique for 3D face modeling for three data bases is described in next section, before the description of details of proposed TPS warp features.

3. 3D Face modeling

The VidTIMIT data base consists of frontal and profile view images of the faces, and AVOZES data comprises left(top) and right(bottom) images of the faces, as shown in Figure 1(a) and (b). We used a unified approach for 3D face modeling of faces from the databases, based on [1, 6, 7, and 8]. The algorithm starts by computing 3D coordinates of automatically extracted facial feature points. Correspondence between feature points in both images is established using epipolar constraints, and then depth information from front and profile views for VidTIMIT faces, and, left and right views for AVOZES faces, is 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.

The wire-mesh for face modeling can be created either by finite element modeling of vertices and surfaces, or based on graphics software created by an artist. We used second option, and used discreet 3DSMAX™, a commercial 3D graphics software for creating the generic head model based on polygons. Figure 2 shows the generic head model created by the software.

The generic model then undergoes global alignment and local refinement. The global alignment stage brings the generic model and facial measurements into same coordinate system. Then, local refinement is performed by generating 3D spline curves for each facial component and adjusting corresponding vertices of the 3D model accordingly. Further details of the face modeling and automatic facial feature extraction are given in [9].

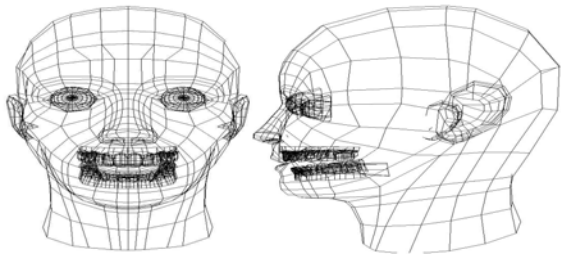


Figure 2: Front and profile view of a generic head model

The global alignment of generic head model shown in Figure 2 for each person's head shape involves deformation of 15 vertices. The entire 3D generic model is brought as close as possible to the corresponding 3D coordinates of anchor points calculated from the images of the person's face. This is done by rotating, translating and scaling to match the calculated 3D points by minimizing the sum squared error criteria shown in equation 1.

IEEE COMPUTER SOCIETY

$$\min Error(S,R,T) = \sum_{po\,\mathrm{int}\,sS} (P_I - P_M)^2 \qquad (1)$$

Where

$$P_M = \begin{bmatrix} x_M \\ y_M \\ z_M \end{bmatrix} = S.R.\begin{bmatrix} x_{M0} \\ y_{M0} \\ z_{M0} \end{bmatrix} + T$$

Subscripts M, I and M0 correspond to model points, calculated image points, and initial model points. S is scaling factor, R is the Rotation matrix and T is the translation vector. Equation 1 can be solved by adjusting the parameters of S, R and T matrices.

Given two sets of 3D points, namely 15 calculated (x_I, y_I, z_I) anchor points, and 15 corresponding (x_M, y_M, z_M) model points, global alignment algorithm finds the translation and rotation matrices that best match the corresponding data points. That is, it calculates the best fit of two similar sets of 3D data points, the best least squares translation is computed from the centers of mass of the two sets of data points. The rotation matrix is computed using a linearized iterative least-squares algorithm. The global rigid alignment deforms successfully, scales and aligns the generic model to the 3D feature points calculated from the face images.

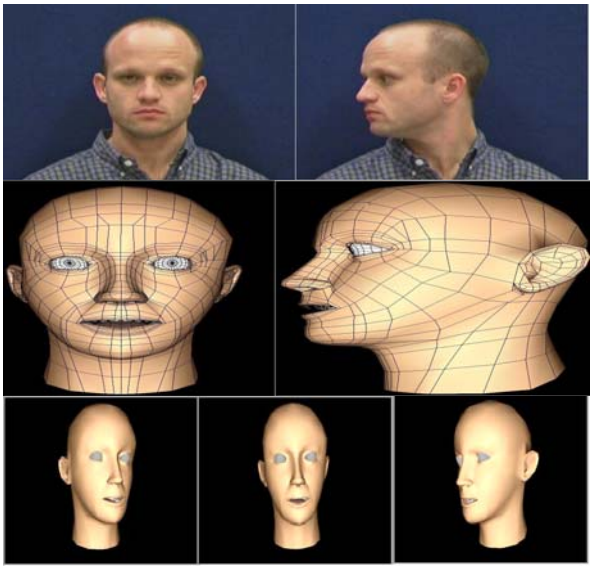


Figure 3: 3D face model of VidTimit face by global and local alignment of generic face model shown in Figure 2

Local refinement is then implemented by treating each of the facial features as separate non-rigid components, and the vertices of the generic model are brought closer to the calculated 3D anchor points of the person's face. Facial texture for all the vertices is computed by blending the R, G, B color components of two views of the face. Figure 3 shows the textured 3D face model for a male subject in VidTIMIT database by global and local alignment of generic face model.

4. Thin plate spline warp

The use of thin plate spline warping features was first proposed by Bookstein [10]. A speaking face is characterized by certain person-specific subtle non-rigid facial deformations, such as nasolabial folds, horizontal wrinkles between eyes and forehead, furrows on the fore head and the brows, as shown in Table 1 [11,12]. Such subtle variations on the face cannot be captured with current methods, though there is a rich literature on methods for modeling non-rigid deformations, ranging from contour, shape, appearance to optical flow [13-19]. Thin plate spline(TPS) formulation, a powerful surface interpolation approach finds a "*minimally bended*" smooth surface that passes through all control points, and in the process maps the facial regions with high curvatures, such as nasofabial folds and furrows, to a subspace that are more discriminatory. Due to better modeling of non-rigid facial deformations, it is easier to distinguish genuine clients, from imposters and fake clients (replay attacks). The name "Thin Plate" comes from the fact that a TPS more or less simulates how a thin metal plate would behave if it was forced through the same control points.

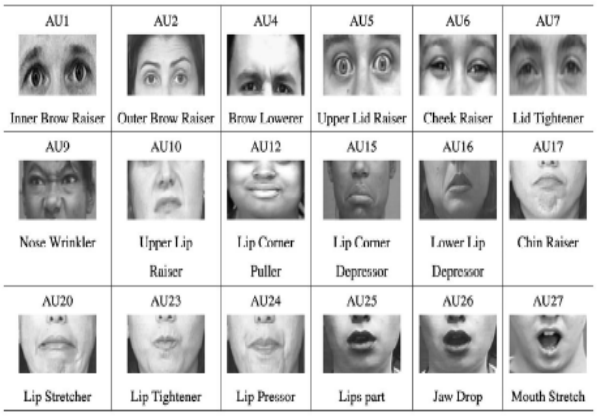


Table 1: Facial expression changes for a speaking face [adapted from [11]

IEEE COMPUTER SOCIETY

Thin plate spline warp allows parameterization of a warping transformation based on a set of fiducial or anchor points. It effectively generates a minimum bending energy (approximate curvature) solution to a point constrained warping. The transformed warp features allow extraction of non-rigid facial deformations such as expression lines and wrinkles, as facial expressions can be thought of as the deviation of facial action from neutral zero energy position.

We show that by selecting the anchor points on the face that correspond to separate facial action units, it is possible to extract person-specific expressions, wrinkles and gestures. In addition, differential TPS warp field discriminates clearly the intra-subject variations due to affine pose and illumination variations, and non affine expression variations, from the warp field corresponding to imposters and spoof attack scenarios. This is due to the ability of TPS warp formulation to modeling rigid(affine) head movements as well, and the solution to TPS model equations comprise an affine part for rigid deformations, and a non-linear part for non-rigid deformations. For sake of simplicity, 2D TPS model is described here, though we included the depth information for TPS warping.

The TPS model is initialized from the neutral face region with '*n*' control point window surrounding the facial feature anchor points as shown in Figure 2.

Given *n* control points $(\hat{x}_i, \hat{y}_i) \in IR^2$, in a plane, and their corresponding function values $\hat{v}_i \in IR, i = 1, \cdots, n$, thin plate spline warp transformation $f(x, y)$, specifies a mapping $f : IR^2 \rightarrow IR$, whose bending energy E_f is minimal,

$$I_f = \arg\min_f \iint_{R^2} \left(\left(\frac{\partial^2 f}{\partial x^2} \right)^2 + 2\left(\frac{\partial^2 f}{\partial x \partial y} \right)^2 + \left(\frac{\partial^2 f}{\partial y^2} \right)^2 \right) \quad (1)$$

And the warped or interpolated value at a point *(x, y)* is given by

$$f(x,y) = a_1 + a_2 x + a_3 y + \sum_{i=1}^{n} w_i U(\|(\hat{x}_i, \hat{y}_i) - (x,y)\|) \quad (2)$$

Where '*U*' is a cost function defined as:

$$Z = -U = -r^2 . \log(r^2)$$
$$and,\ r^2 = x^2 + y^2 \quad (3)$$

The interpolated spline function consists of two parts: affine transformation parameterized by *'a'*, and non-affine warping specified by '*w*'. Since the warp transform here is a spline function *f(x, y)*, it needs to have square-integrable second derivatives with following constraints:

$$\sum_{i=1}^{n} w_i = 0,\ and\ \sum_{i=1}^{n} w_i \hat{x}_i = \sum_{i=1}^{n} w_i \hat{y}_i = 0 \quad (4)$$

The TPS parameters **a** and **w** can be computed by solving the following system of equations:

$$\begin{bmatrix} K & P \\ P^T & \mathrm{O} \end{bmatrix} \begin{bmatrix} w \\ a \end{bmatrix} = \begin{bmatrix} v \\ 0 \end{bmatrix} \quad (5)$$

Where $K = U(\|(\hat{x}_i, \hat{y}_i) - (\hat{x}_j, \hat{y}_j)\|)$, the *i-th* row of *P* is $(1, \hat{x}_i, \hat{y}_i)$, **O** and **0** are 3×3 and 3×1 zero padding matrices, and **ω**, **a**, and **v**, are vectors formed from $w_i, a_1, a_2, a_3\ and\ \hat{v}_i$, respectively.

Here, we show warping of 2D points using TPS defined by '*pairs*' of control points, i.e.., we want to map points (x, y) to (x', y') given *n* control point correspondences $(\hat{x}_i, \hat{y}_i) : (\hat{x}'_i, \hat{y}'_i)$. For that, we need two TPS functions for *x* and *y* coordinates separately. From equation (5), the TPS warp which maps $(\hat{x}_i, \hat{y}_i)$ onto $(\hat{x}'_i, \hat{y}'_i)$ can be recovered by

$$\begin{bmatrix} w_x & w_y \\ a_x & a_y \end{bmatrix} = L^{-1} \begin{bmatrix} \hat{x}' & \hat{y}' \\ 0 & 0 \end{bmatrix} \quad (6)$$

Where 'L' is defined as:

$$L = \begin{bmatrix} K & P \\ P^T & 0 \end{bmatrix} \quad (7)$$

And $\hat{x}'$ and $\hat{y}'$ are vectors formed with $\hat{x}'_i$ and $\hat{y}'_i$ respectively. The transformed coordinates (x'_j, y'_j) of points (x_j, y_j) are given by

$$[x' \quad y'] = [B \quad Q] \begin{bmatrix} w_x & w_y \\ a_x & a_y \end{bmatrix} \quad (8)$$

Where $B_{ji} = U(\|(x_j, y_j) - (\hat{x}_i, \hat{y}_i\|)$, the *j-th* row of *Q* is $(1, x_j, y_j)$, and the *j-th* row of the resulting vectors x' and y' are the interpolated *x* and *y* coordinates x'_j and y'_j, respectively. We denote the matrix $[B \quad Q]$ as '*M*'. So, the TPS warp procedure can be described in two steps:

IEEE COMPUTER SOCIETY

1. Given the correspondence of control points in the two images, the TPS coefficients are first estimated and then,
2. Points of interest are transformed to new locations using interpolated spline from TPS coefficients estimated above.

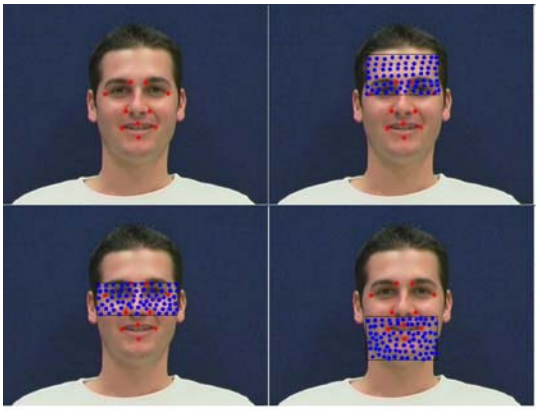


Figure 4: VidTIMIT face with anchor points And regions with control points for TPS warp.

The initial 10 facial anchor points shown in Figure 4 corresponding to the neutral face for VidTimit and Avozes, automatically extracted similar to method used for 3D face modeling [9]. The algorithm extracts 10 anchor points corresponding to the left and right corner of the lips, the top and bottom of the lips, the left and right side of nostrils, the outer corner of the eyes, and inner corner of the eyebrows. For TPS warping a square window with *10*n* (around 40) control points around anchor points corresponding to different facial feature regions are chosen as shown in Figure 4. The control points in three windows undergo TPS warp as in equations(1) to (8) above. The differential warp field (DWF) magnitude is computed as the difference between the magnitude of *x, y, z* (3D) of original control points and warped control points. The DWF computations are repeated for all frames of entire speaking face video, by projecting the markers (anchor points) from neutral face by deformable template matching technique described in [9]. The DWF magnitude for same person with gesture and expression variations is different from DWF magnitudes corresponding to an imposter and fake client trying to spoof the system. For genuine clients with gesture and expression variations, the DWF magnitude is strong around cheeks and chin, between two eyebrows and nasolabial regions, while it is low around nose and eye areas. For imposters, the DWF magnitude is noisy and strong deformation is obtained near nose, eyes and mouth, due to facial features of an imposter being different from a client, and DWF around these key facial features dominates more than the regions corresponding to expression and gesture variations. For fake clients trying to spoof the system (simulated here with photo-realistic CG animated talking head), the DWF magnitude is random and very noisy around the facial regions corresponding to expression and gesture variations such as nasogenian wrinkles, horizontal wrinkles between eyes and forehead, furrows on the fore head and the brows. Figure 5 shows simulated facial deformations for genuine client, fake client (spoof attack) and impostor for an utterance "aah" with a neutral, smiling and angry face. The differential warp field (DWF) for fake client and impostor is shown in Figure 6.

Figure 5: Facial deformations for genuine client, fake client and impostor for an utterance "aah"

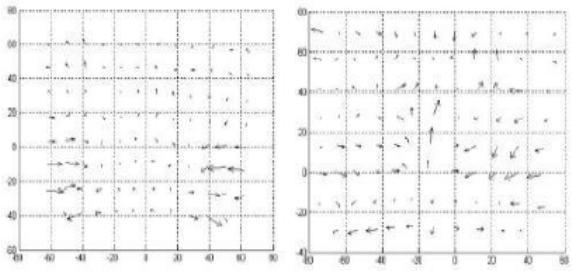


Figure 6: DWF field for fake client and impostor

The 120 dimensional DWF vector (40 control points*3 windows) being high can make the DWF classifier very weak. However, from initial experiments, we learnt that more than 97% non-rigid deformations for a speaking face can be modeled by

IEEE COMPUTER SOCIETY

Eight-ten principal differential warps or Eigen values of differential warp field (DWF magnitudes), two-four each for the three windows shown in figure 4. These exigent values correspond to each of the regions from AU1-AU7, AU9-AU17, and AU20-AU27, shown in Table 1.

For acoustic features, the Mel frequency cepstral coefficients (MFCC) as derived from the cepstrum information were used. The MFCC features were obtained by pre-emphasizing the audio signal first, and then processed with a 30ms Hamming window with one third overlap, yielding a frame rate of 50 Hz. 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. In addition, log energy and pitch information computed by autocorrelation method was used.

5. Authentication Experiments

To investigate the potential of DWF warp features for thwarting impostor attacks and spoof attacks, different sets of experiments were conducted using 16 dimensional multimodal audio-visual feature vector(8 MFCCs +1 log-Energy + 1 pitch+ 6 TPS warp features).

In the training phase, a 10-Gaussian mixture model of each client's feature vectors in the three dimensional space was built by constructing a gender-specific universal background model (UBM) and then adapting each UBM by MAP adaptation [11]. Both text-dependent and text-independent experiments were conducted with VidTIMIT corpus and text-dependent experiments with AVOZES data. For all experiments, the threshold was set using data from the test data set. Table 2 shows the notation used for different experimental modes.

Notation	True Description
EER	Equal Error Rate
DB1	Vitamin corpus
DB2	AVOZES corpus
TDMO	Text dependent male only cohort
TDFO	Text dependent female only cohort
TIMO	Text independent male only cohort
TIFO	Text independent female only cohort

Table 2: Notation for different experimental modes

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 in cross-modal space. A Z-norm based approach [12] was used for score normalization.

For testing replay/spoof attacks, two types of replay-attack experiments were conducted. For Type-1 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 Type-2 replay attacks, a video clip was constructed from a still photo of each speaker. This represents a scenario of a spoof 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. The still photo of each client was voice-synched with the speech signal of each speaker, using a set of commercial software tools (Adobe Photoshop Elements, Discreet 3DSMax, and Adobe After Effects). We constructed several fake video clips by extracting ONE face (the first face) from the video sequence, which acts as a key frame, animated the lip region of the key frame 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. Such a fake clip emulates a normal talking head with certain facial and head movements in three dimensional spaces in synchronism with spoken utterance.

Different sets of experiments were conducted to evaluate the performance of the system in terms of DET curves and equal error rates. The results for only two types of data, that is DB1TIMO (VidTIMIT database text-independent male-only cohort) and DB2TDFO (AVOZES database text dependent female-only cohort) are reported here. For both types of data, both late-fusion and feature-level fusion of shape and texture features were examined. For late-fusion equal weights for shape and feature fusion was used.

For VidTimit corpus in text-independent mode there were 144 client trials (24×6) and 3312 impostor

IEEE COMPUTER SOCIETY

trials (24×23×6) for male subjects. For AVOZES there were 53 client trials and 4770(10×9×53) imposter trials. Next set of experiments were for testing the Type-1 replay attacks. For the VidTimit database in text-independent mode, there were 144 client trials (24×6) and 576(24×6×4) replay attacks for male subjects. For AVOZES data, there were 53 client and 2120 (10×53×4) replay attack trials for both female subjects in text dependent mode. The third set of experiments is to test Type-2 replay attacks, where the numbers of client and spoof attack trials were same as client trials. Table 3 shows the number of client, imposter and replay attack trials for each set.

The DET curve and EER results in Table 4, and Figures 7 and 8, show the potential of the proposed fusion of principal TPS warp features with acoustic features (MFCC+f0) to thwart imposter and replay attacks for VidTIMIT data and AVOZES data.

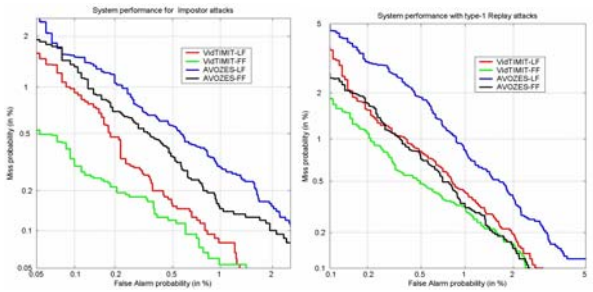
Corpus	DB1TIMO	DB2TDFO
Client Trials	144 (24 clients × 6 Utterances per client)	530 trials (10×53)
Imposter Trials	3312 trials (24×23 ×6)	4770 trials (10×9×53)
type-1 Replay-attack Trials	576 trials (24×6×4)	2120 trials (10×53×4)
type-2 Replay attack Trials	144 trials	530 trials

Table 3: Number of Client, Imposter and Replay attack trials

For VidTIMIT corpus, less than 0.5% EER achieved for impostor attacks, with 0.28% for late fusion and 0.19% for feature fusion. Feature fusion performs better, around 32% improvement as compared to late fusion, due to synchronous processing of principal warp and acoustic features. For AVOZES corpus, EER achieved is 0.42% with feature fusion as compared to 0.50% for late fusion, about 16% EER improvement. For type-1 replay attacks, less than 1 % EER is achieved for VidTIMIT and AVOZES, with feature-fusion performing better than late fusion (27% improvement for VidTIMIT data vs. 33% for AVOZES data). Less than 2% EER is achieved for type-2 replay attacks for both VidTIMIT and AVOZES data, with best EER equal to 0.77% for VidTIMIT TIMO data and worst EER of 1.57% for AVOZES TDFO data.

% EER achieved	ViDTIMIT TIMO	ViDTIMIT TIMO	AVOZES TDFO	AVOZES TDFO
Fusion Type	*Late Fusion*	*Feature Fusion*	*Late Fusion*	*Feature Fusion*
Imposter Attacks	0.28	0.19	0.50	0.42
Type1 RA attacks	0.65	0.47	0.90	0.60
Type-2 RA attacks	1.1	0.77	1.57	1.25

Table 4: EERs for impostor and replay attacks



(a) (b)

Figure 7: DET curves for testing (a) imposter, and (b) type 1 replay attacks

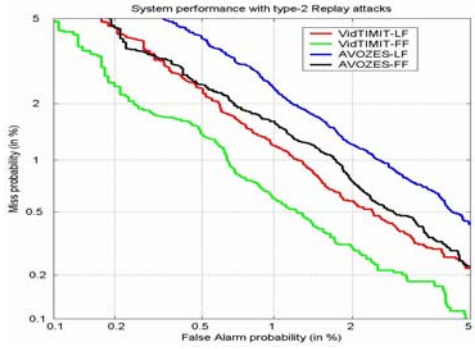


Figure 8: DET curves for testing type-2 replay attacks

The fusion of acoustic features with three dimensional TPS warp features allowed a significantly enhanced performance for both impostor and spoof attacks, including type-2 replay attacks, which are more complex to detect. VidTIMIT data in general performs better than AVOZES data for all experiments. This can be due to several reasons, better quality of images in VidTIMIT, and difference in accuracy of depth computations with availability of frontal and side faces

IEEE COMPUTER SOCIETY

in VidTIMIT, and left and right images in AVOZES. Figures 7 and 8 shows the DET curves corresponding to the EERS in table 4.

6. CONCLUSIONS

The potential of TPS warp features to thwart imposter and still-photo/video-replay spoof attacks for audio-video biometric system has been shown in this study. The multimodal feature fusion of acoustic and TPS warp features allowed less than 0.5 % EERs to be achieved for imposter attacks, and less than 1% for *type-1* replay attacks. With less than 2 % EER for *type-2* replay attacks, a significant enhancement in performance was achieved for more difficult type-2 replay attacks.

10. References

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

[2] Iguana Lu and A.K.Jain, "Deformation analysis for 3D face matching", Proc. WACV (Workshop on Applications of Computer Vision), pp. 99-104, Breckenridge, Colorado, January 2005

[3] Chetty, G. and Wagner, M., "Liveness detection using cross-modal correlations in face-voice person authentication, Inter Speech 2005, Lisbon, Portugal, Sept 4- 7 2005.

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

[5] R. Goecke 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*, Juju, Korea, 4 - 8 October 2004.

[6]Bland, V. and Vetter, T., "Face recognition based on fitting a 3D morphable model", *IEEE trans. Pattern Analysis and Machine Intelligence*, 25(9):1063-1074, 2003.

[7] Beumier C. and McKay N., "Automatic 3D face authentication", *Image and Vision Computing*, 18(4):315-321, 2000.

[8] G.Gordon, "Face Recognition from Frontal and Profile Views," *Proceedings Int'l Workshop on Face and Gesture Gesture Recognition*, Zurich, 1995, pp.47-52.

[9] Chetty, G. and Wagner, M., "Automated lip feature extraction for liveness verification in audio-video authentication", *Proc. Image and Vision Computing 2004*, New Zealand, pp 17-22.

[10] F.L.Bookstein, "Principle Warps: Thin-Plate Splines and the decomposition of deformations", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 11, No.6, June 1989.

[11]Ying-li Tian, Takeo Kanade, Jeffrey Kohn, "Recognizing lower face action units for facial expression analysis", *Proceedings of the 4th International Conference on Automatic Face and Gesture Recognition (FG00)*, May 2000, pp. 489-490.

[12]Ekman,P,"Facial Expressions", *In T. Dalgleish and T. Power (Eds.) The Handbook of Cognition and Emotion* Sussex, U.K.: John Wiley & Sons, Ltd., 1999, . Pp. 301-320.

[13] M. Black and Y. Yacoob. Tracking and recognizing rigid and non-rigid facial motions using local parametric models of image motion. In *Proceedings of the Fifth IEEE International Conference on Computer Vision*, pages 374–381,1995.

[14] H. Chui and A. Rangarajan. A new algorithm for non-rigid point matching. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, pages 44–51, 2000.

[15] T. Cootes, C. J. Taylor, D. Cooper, and J. Graham. Active shape models - Their training and application. *Computer Vision and Image Understanding*, 61:38–59, 1995.

[16] T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(6):681–685, 2001.

[17] L. Wiskott, J. Fellous, N. Kruger, and C. von der Malsburg, Face recognition by elastic bunch graph matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):775–779, 1997.

[18] M. Isard and A. Blake. Contour tracking by stochastic propagation of conditional density. In B. Buxton and R. Cipolla, editors, *Proceedings of the Fourth European Conference on Computer Vision*, LNCS 1064, pages 343–356. Springer Verlag, 1996.

[19] N. Paragios and R. Deriche. Geodesic active contours and level sets for the detection and tracking of moving objects. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(3):266–280, 2000.

IEEE COMPUTER SOCIETY

View publication stats