

Audio-Video Biometric System with Liveness Checks

Girija Chetty and Michael Wagner

HCC Laboratory, School of ISE, University of Canberra, Australia.

Email: girija.chetty@canberra.edu.au

Abstract

In this paper we propose multimodal fusion of super resolved texture (SRT) features and 3D shape features with acoustic features for 3D audio-video person authentication systems with liveness checks. The proposed SRT features allow 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 super-resolved texture and 3D shape features for two different speaking face data corpus, VidTIMIT, and AVOZES, allowed equal error rates (EERs) of less than 0.5 % for impostor and type-1 replay attacks (still photo and pre-recorded audio) and less than 3% for more complex type-2 replay attacks (pre-recorded video or fake CG animated video).

Keywords: audio-video, biometrics, liveness, 3D fusion, super-resolution

1 Introduction

Current audio-video biometric 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 based on the 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 impostor 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 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 features based on super-resolving texture (SRT) information of faces is proposed,

allowing a significant enhancement in performance of audio-video biometric systems against impostor and replay attacks. An equal error rate (EER) of less than 0.5% was achieved for impostor and type-1 replay attacks (pre-recorded audio and still photo) and less than 3% for type-2 replay attacks (pre-recorded video or fake speaking faces created with CG animation and other similar techniques). The SRT 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 details of 3D face modeling technique used are given in section 3, followed by description of SRT features in section 4. The details of impostor and replay attack experiments are described in section 5, with conclusions in section 6.

2 Speaking Face Data Corpus

The speaking face data from three different data corpus, VidTIMIT and AVOZES was used for conducting impostor 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 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 SRT features.

3 3D Face modeling

The VidTIMIT data base consists of frontal and profile view images of the faces, and AVOZES data comprises left and right view images of the faces, as shown in Figure 1(a) and (b). We used a common approach for 3D face modeling of faces from the databases, based on [1, 6, 7, and 8]. The algorithm used output of SuperRes algorithm described in next section and 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.

For this, the generic model 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 shows the computational steps involved in constructing 3D head model for a speaker from VidTIMIT database.

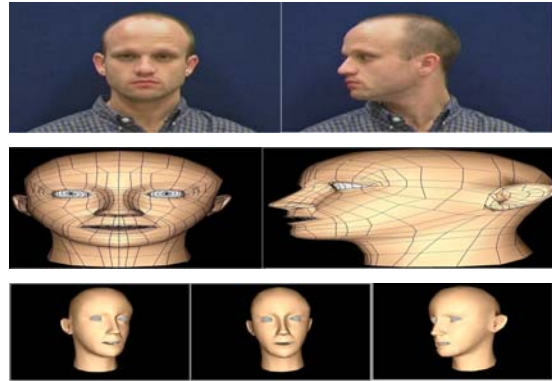


Figure 2: 3D face model of VidTIMIT face by global and local alignment of generic face model

4 Super-Resolved Texture (SRT) features

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 forehead and the brows, as shown in Table 1 [10,11]. 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.

AU1 Inner Brow Raiser	AU2 Outer Brow Raiser	AU4 Brow Lowerer	AU5 Upper Lid Raiser	AU6 Cheek Raiser	AU7 Lid Tightener
AU9 Nose Wrinkler	AU10 Upper Lip Raiser	AU12 Lip Corner Puller	AU15 Lip Corner Depressor	AU16 Lower Lip Depressor	AU17 Chin Raiser
AU20 Lip Stretcher	AU23 Lip Tightener	AU24 Lip Pressor	AU25 Lips part	AU26 Jaw Drop	AU27 Mouth Stretch

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

We propose a super-resolution method to recover subtle details in the facial region. We first transform each image into a new parametric vector space characterised by the image edges, instead of matching image patches in the spatial domain. Then, we create a database of source edges, and replace low resolution data around each target edge with high resolution detail from the database.

The image content from orthogonal pixel space is transformed to a parametric space structured around edges, based on a super-resolution transform. The super-resolution (SuperRes) consists of a number of computational stages as illustrated in Figure 4. These computational stages can be logically divided into two parts: the Absorption phase, through which both source and target image frames are transformed and added to edges database, and the Synthesis phase, in which a single target image frame is reconstructed at a much higher resolution than its original pixels provided.

Absorption: Before commencing with the true SuperRes Algorithm, the target image frame is obtained by averaging several consecutive image frames, and interpolating the average frame up to the desired output size (e.g., 400%) using the simple bicubic interpolation. Because we are only interested in analysing and matching regions of high spatial detail for extracting non-rigid facial deformations such as wrinkles, the high-frequency images for both source image and target image frames are obtained by using a high-pass filtering with first order Gaussian derivative filter. The high pass filtered versions of the source and target image frames are further processed with SuperRes algorithm. To construct an image frame around its edges, we first perform edge detection by using Sobel edge detector, and then apply the cubic splines to fit to the edges. This parametric curve representation allows us to establish a bidirectional mapping between two coordinate spaces: the orthogonal space in which the image pixels reside along x , and y coordinates, and the parametric space, a nonlinear transformed space relative to the curvature of each edge. Each edge has its own local parametric space in the form of a warped rectangle, with the u coordinate running parallel to the edge and the v coordinate running out along the edge's normal as evaluated at u . We restrict the u coordinate to the uniform range $[0; 1]$ such that it represents the normalized distance along the edge length we are examining.

The coordinate of a point $\vec{A}(u)$ at a parametric coordinate u in the span between any two known edge points \vec{B}_n and \vec{B}_{n+1} is defined as

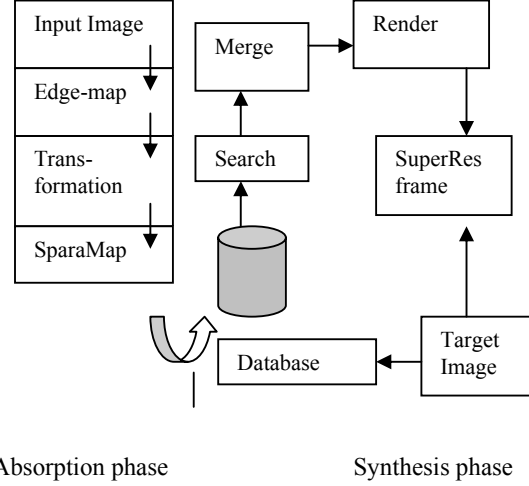


Figure 4: Stages of SuperRes algorithm

$$\vec{A}(u) = \frac{1}{2} \begin{pmatrix} 1 \\ \left(\frac{u-U_n}{S_n}\right) \\ \left(\frac{u-U_n}{S_n}\right)^2 \\ \left(\frac{u-U_n}{S_n}\right)^3 \end{pmatrix} \begin{bmatrix} 1/6 & 2/3 & 1/6 & 0 \\ -1/2 & 0 & 1/2 & 0 \\ 1/2 & -1 & 1/2 & 0 \\ -1/6 & 1/2 & -1/2 & 1/6 \end{bmatrix} \begin{bmatrix} \vec{B}_{n-1} \\ \vec{B}_n \\ \vec{B}_{n+1} \\ \vec{B}_{n+2} \end{bmatrix}$$

This span is defined as starting at $u = U_n$, with a span length of $S_n = \vec{B}_{n+1} - \vec{B}_n$. Through the use of this cubic spline approach, a smooth transfer function, $\vec{A}(u)$ from the parametric coordinate u to x ; y orthogonal coordinates is realized. To completely define the parametric space, we must also find the v coordinate, which runs parallel to the curve's normal at a given value of u . The complete transform from a parametric coordinate u ; v to an orthogonal coordinate $\vec{T}(u, v) \rightarrow x, y$ is thus:

$$\vec{T}(u, v) = \vec{A}(u) + v \frac{\partial \vec{P}(u)}{\partial u} \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}$$

where

$-v_\sigma \leq v \leq v_\Sigma$, $\vec{T}(u, v) = [T_x(u, v), T_y(u, v)]$ is the final coordinate in the orthogonal space.

$$\vec{T}(u, v) = [A_x(u), A_y(u)]$$

$$\text{and } \frac{\partial \vec{A}(u)}{\partial u} = \left[\frac{\partial \vec{A}_x(u)}{\partial u}, \frac{\partial \vec{A}_y(u)}{\partial u} \right]$$

After fitting each edge to a parametric curve representation, the curve area around each edge is

transformed to a standard rectangle (u, v) area, thus forming a ‘‘Super parametric map’’(called SparaMap).

Synthesis: In this stage we use two techniques: key generation and hierarchical decomposition.

Key generation refers to the construction of fixed size keys for each edge based on its Sparamap; these keys are then entered into the database and searched.

Then, a hierarchical decomposition model is used for mapping edges to keys with the start by creating a single key for an entire edge, then recursively splitting the edge into two segments. We place the split point at the value of u where the following is maximized:

$$\max \left(\left| \frac{\partial}{\partial u} \int_{+v_\sigma}^{-v_\sigma} e^{-\frac{4}{v_\sigma^2} \sigma^2} |A(u, v)| dv \right| \right)$$

It divides the SparaMap at the u value where sharpest change of the average Sparamap values along v occurs; this naturally subdivides edges at inflection points or where other edges intersect. As a result, we can always attempt to match the longest possible edge first, then subdivide as needed to optimize for accuracy.

The results of the search stage are in the form of key pairs: a target key and its matching source key. In the Merge stage, we map the source key back to the corresponding subsection of the Sparamap owning it, and merge that parametric pixel data back into the correct place in the target edge’s Sparamap. The reconstructed Sparamap is further rendered onto the curvature of the appropriate target edge, which is implemented by projecting the u, v coordinates back into orthogonal x, y space. As a result, the detail-reconstructed high-frequency image is created. Finally, we add this high- frequency image back into the original low resolution image, just as if we were performing an unsharp masking operation, to generate the new resolution-increased image at the completion of the SuperRes algorithm. Figure 5 show one example of the SuperRes results.



Figure 5: left: original face region 64*64 pixels; middle: enlarged by cubic interpolation (256*256); right: enlarged by SuperRes (256*256).

The fusion of high dimensional shape and SRT features (texture extracted from SuperRes image frames) for all vertices face can make the speaker models very weak. However, from initial experiments, we learnt that more than 97% non-rigid deformations for a speaking face can be modeled by

eight to ten principal shape and SRT features. These eigenvalues 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 3D shape and SRT features against impostor and spoof attacks, different sets of experiments were conducted using 20 dimensional multimodal audio-visual feature vector(8 MFCCs + log-Energy + pitch+ 5 SRT + 5 three dimensional shape 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. Both text-dependent and text-independent experiments were conducted with VidTIMIT corpus and text-dependent experiments with AVOZES data.

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 3D shape-texture space. A Z-norm based approach 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 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. 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, experiments without SuperRes enhancement of texture features were first conducted for baseline comparison, and improvement achieved with fusion of acoustic, shape and SRT features were examined. Table 3 shows the number of client, impostor and replay attack trials for each set.

The DET curve and EER results in Table 4 and Figures 6, 7 and 8, show the potential of the proposed fusion of principal shape and SRT features with acoustic features (MFCC+f0) to thwart impostor 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)
Impostor Trials	3312 trials (24×23 ×6)	4770 trials (10×9×53)
type-1RA Trials	576 trials (24×6×4)	2120 trials (10×53×4)
type-2RA trials	144 trials	530 trials

Table 4: Number of Client, Impostor and Replay attack trials

For VidTIMIT corpus, less than 1% EER achieved, with 0.92% for plain texture features and 0.64% for

SRT feature fusion, a 30% improvement, due to synchronous processing of SRT, 3D shape and acoustic features. For AVOZES corpus, EER achieved is 1.24% with SRT features as compared to 1.53 %, about 20% EER improvement. For type-1 replay attacks, less than 1 % EER is achieved for VidTIMIT and AVOZES, with SRT feature-fusion performing better than plain texture fusion (48% improvement for VidTIMIT data vs. 38% for AVOZES data). Less than 3% EER is achieved for type-2 replay attacks for both VidTIMIT and AVOZES data, with best case EER equal to 1.9% for VidTIMIT TIMO data and worst case EER of 2.45% for AVOZES TDFO data. The fusion of acoustic features with three dimensional shape and super-resolved texture features allowed a significantly better performance, though type-2 replay attacks are more complex replay attacks to detect. VidTIMIT data in general performs better than AVOZES data for all experiments. Figures 6 and 7 shows the DET curves corresponding to the EERs of two experiments.

% EER achieved	VidTIMIT TIMO	VidTIMIT TIMO	AVOZES TDFO	AVOZES TDFO
<i>Fusion Type</i>	<i>Plain Texture</i>	<i>SRT Texture</i>	<i>Plain Texture</i>	<i>SRT Texture</i>
Impostor Attacks	0.92	0.64	1.53	1.24
Type-1attacks	0.44	0.23	0.95	0.59
Type-2attacks	1.4	1.9	2.45	2.3

Table 5: EERs for impostor and replay attacks

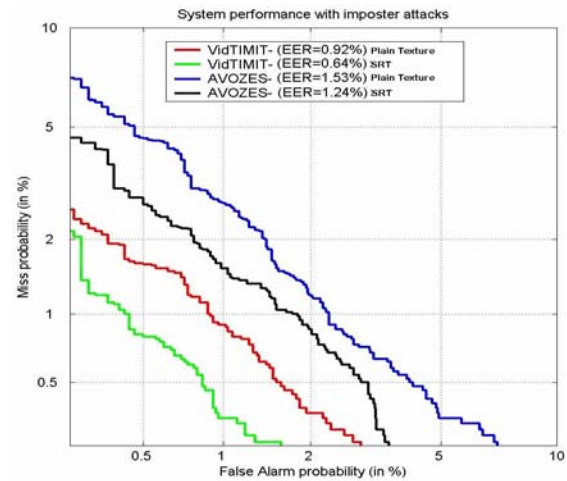


Figure 6: DET curves for impostor attacks

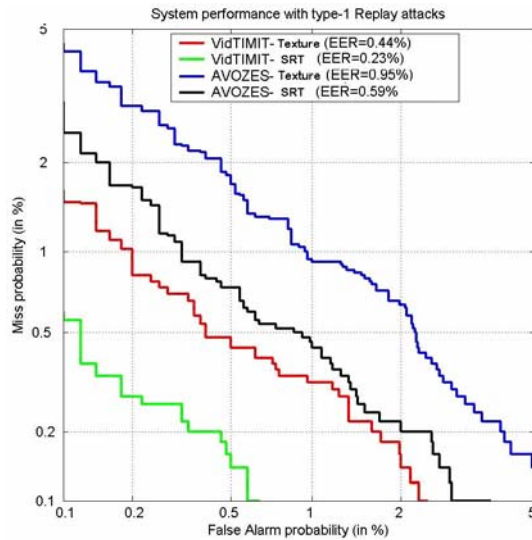


Figure 7: DET curves for type-1 replay attacks

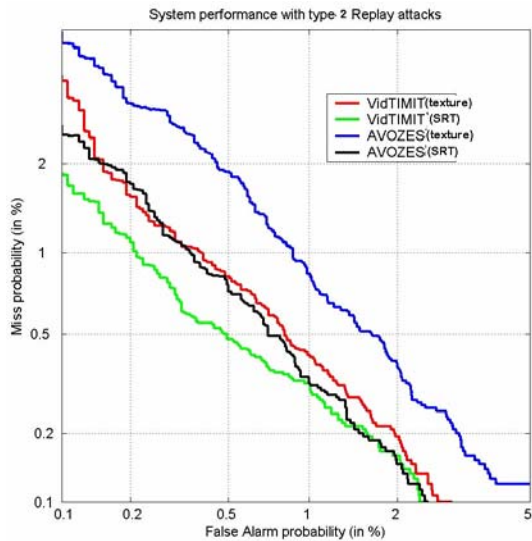


Figure 8: DET curves for type-2 replay attacks

6 Conclusions

The potential of super-resolved texture features for thwarting impostor and still-photo/video-replay spoof attacks for audio-video biometric system has been shown in this study. The multimodal feature fusion of acoustic, 3D shape and SRT features allowed less than 2 % EERs to be achieved for impostor attacks, and less than 1% for type-1 replay attacks. With less than 3% EER, significantly better performance was achieved for more difficult type-2 replay attacks compared to baseline plain texture fusion experiments.

7 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] Xianguang 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, "Fast features for face authentication under illumination direction changes", *Pattern Recognition Letters* 24, 2409-2419, 2003
- [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* pp. 2525-2528, 2004.
- [6] Blanz, 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] Ying-li Tian, Takeo Kanade, Jeffrey Kohn, "Recognizing lower face action units for facial expression analysis", *Proceedings of the Conference on Automatic Face and Gesture Recognition*, pp. 489-490, 2000.
- [11] Ekman, P, "Facial Expressions", In T. Dalgleish and T. Power (Eds.) *The Handbook of Cognition and Emotion Sussex*, U.K.: John Wiley & Sons, Ltd., Pp. 301-320, 1999.