Detecting Liveness in Multimodal Biometric Security Systems

Girija Chetty

Faculty of Information Sciences and Engineering, University of Canberra, Australia

Summary

In this paper a novel liveness checking technique is proposed for multimodal biometric authentication systems based on face and voice biometrics. Liveness detection ensures that biometric cues are acquired from a live person who is actually present at the time of capture for authenticating the identity. The proposed liveness checking technique based on corellation modelling that involves fusion of acoustic and visual speech features which measure the degree of synchrony between the lips and the voice extracted from speaking face video sequences.

Key words:

Multimodal, face-voice, liveness detection biometric security

1. Introduction

Most of the commercial biometric security systems currently deployed are based on modeling the identity of a person based on unimodal biometric information, i.e. face, voice, or fingerprint features. Also, many current interactive civilian human computer interaction applications are based on speech based voice features, which achieve significantly lower performance for operating environments with low signal-to-noise ratios (SNR). Use of both visual and audio information can lead to better robustness, as they can provide complementary secondary clues that can help in the analysis of the primary biometric signals [1]. In extreme cases, primary biometric (visual or acoustic) information can even be used on its own. For instance, it is well known that deaf people can learn how to lip read. The joint analysis of acoustic and visual speech improves the robustness of automatic speech recognition systems [2, 3].

There have been several systems proposed on use of joint face-voice information for improving the performance of identity authentication systems. However, most of these state-of-the-art approaches are based on independently processing the voice and face information and then fusing the scores – score fusion [4,5,6]. A major weakness of these systems is that they do not take into account fraudulent replay attack scenarios into consideration, leaving them vulnerable to spoofing by recording the voice of the target in advance and replaying it in front of the microphone, or simply placing a still picture of the target's face in front of the camera. This problem can be addressed with liveness checking, which ensures that biometric cues are acquired from a live person who is actually present at the time of capture for authenticating the identity. With the diffusion of Intenet based authentication systems for day-to-day civilian scenarios at a astronomical pace [7], it is high time to think about the vulnerability of traditional biometric authentication approaches and consider inclusion of liveness checks. Though there is some work in finger print based liveness detection techniques [8,9], there is hardly any work in liveness checks based on user-friendly biometric identifiers (face and voice), which enjoy more acceptability for civilian access control scenarios.

A significant progress however, has been made in independent processing of face only or voice only based authentication approaches [1,2,3,4,5,6], without taking into consideration an inherent coupling that exists between jointly occurring some primary biometric identifiers. Some preliminary approaches (such as the one described in [7, 8] address liveness checking problem by jointly modeling the acoustic and visual speech features for testing liveness. They involve the fusion of acoustic, appearance and shape based lip features for jointly modeling the co-occurring face-voice dynamics in speaking face video sequences.

In this paper we propose correlation models for joint analysis of acoustic and visual speech features for incorporating liveness information in the authentication approach. The rest of the paper is organized as follows. Section 2 describes the motivation for using correlation models, and the proposed liveness check approach is described in Section 3. Section 4 details the data corpora used and the experimental evaluation of the proposed correlation models and subsequent fusion approach, with Section 5 summarizing the conclusions drawn from this work and plans for further research

2. Correlation modeling

The motivation to use correlation models is based on the following two observations: The first observation is in relation to any video event, for example a speaking face video, where the content usually consists of the co-occurring audio and the visual elements. Both the elements

Manuscript received October 5, 2009
Manuscript revised October 20, 2009

carry their contribution to the highest level semantics, and the presence of one has usually a "priming" effect on the other: when hearing a dog barking we expect the image of a dog, seeing a talking face we expect the presence of her voice, images of a waterfall usually bring the sound of running water etc. A series of psychological experiments on the mutually dependent cross-modal influences [9, 10] have proved the importance of synergistic fusion of the multiple modalities in the human perception system. A typical example of this kind is the well-known McGurk effect [9]. Several independent studies by cognitive psychologists suggest that the type of multi-sensory interaction between acoustic and orafacial articulators occurring in the McGurk effect involves both the early and late stages of integration processing [9,10]. It is likely that a human brain uses a hybrid form of fusion that depends on the availability and quality of different sensory cues.

Yet, in audiovisual speech and speaker verification systems, the analysis is usually performed separately on different modalities, and the results are brought together using different fusion methods. However, in this process of separation of modalities, we lose valuable cross-modal information about the whole event or the object we are trying to analyze and detect. There is an inherent association between the two modalities and the analysis should take advantage of the synchronised appearance of the relationship between the audio and the visual signal. The second observation relates to different types of fusion techniques used for joint processing of audiovisual speech signals. The late-fusion strategy, which comprises decision or the score fusion, is effective especially in case the contributing modalities are uncorrelated and thus the resulting partial decisions are statistically independent. Feature level fusion techniques, on the other hand, can be favoured (only) if a couple of modalities are highly correlated.

However, jointly occurring face and voice dynamics in speaking face video sequences, is neither highly correlated (mutually dependent) nor loosely correlated nor totally independent (mutually independent). A complex and nonlinear spatiotemporal coupling consisting of highly coupled, loosely coupled and mutually independent components may exist between co-occurring acoustic and visual speech signals in speaking face video sequences [11, 12]. The compelling and extensive findings by authors in [11] validate such complex relationship between external face movements, tongue movements, and speech acoustics when tested for consonant vowel (CV) syllables and sentences spoken by male and female talkers with different visual intelligibility ratings. They proved that the there is a higher correlation between speech and lip motion for C/a/ syllables than for C/i/ and C/u/ syllables. Further, the degree of correlation differs across different places of articulation, where lingual places have higher correlation than bilabial and glottal places. Also, mutual coupling can vary from talker to talker; depending on the gender of the talker, vowel context, place of articulation, voicing, and manner of articulation and the size of the face. Their findings also suggest that male speakers show higher correlations than female speakers. Further, the authors in [12] also validate the complex, spatiotemporal and non-linear nature of the coupling between the vocal-tract and the facial articulators during speech production, governed by human physiology and language-specific phonetics. They also state that most likely connection between the tongue and the face is indirectly by way of the jaw. Other than the biomechanical coupling, another source of coupling is the control strategy between the tongue and cheeks. For example, when the vocal tract is shortened the tongue does not get retracted.

Due to such a complex nonlinear spatiotemporal coupling between speech and lip motion, this could form a good candidate for detecting liveness, and modelling the speaking faces by capturing this information can make the biometric authentication systems less vulnerable to spoof and fraudulent replay attacks, as it would be almost impossible to spoof a system which can accurately distinguish the artificially manufactured or synthesized speaking face video sequences from the live video sequences. We propose an approach based on correlation models and subsequent Bayesian fusion to address this problem. Next section briefly describes the proposed approach.

3. Correlation Models

Correlation modelling based on Canonical Correlation Analysis (CCA), as first proposed by Hotelling [13], is a method of determining a linear space where the correlations between two sets of variables are maximized. This approach has been successfully applied to sets of variables that are manifestations of a set of hidden variables, examples of this are fMRI and image retrieval[14]. There is an obviously similarity with audio-visual speaking face modelling since the motions of articulators and the speech produced are fundamentally linked. However, CCA is derived as a linear process and this limitation becomes apparent in the cases where the underlying relationship is non-linear [15], such as the complex nonlinear spatiotemporal correlations between the speech and lip-motion in speaking face video sequences. To circumvent this linearity constriction, we have used a "kernel trick", which allows replacing an inner product by a projection of the data into a higher dimensional space, and performing CCA in this realized dual representation [15].

We perform a kernel Canonical Correlation Analysis (kCCA) on Mel Frequency Cepstral Coefficients (MFCC)

voice features and the lip motion features extracted from a biological inspired optical flow algorithm called Multi Channel Gradient Model (MCGM).

The MCGM is a neurophysiological and psychophysical inspired unified motion algorithm [15]. In MGCM approach, the behaviour of V1/V2 cells is modelled by MGCM functions and the ratio of temporal and spatial gradients is computed to establish local velocity estimates. From one sequence of lip region images it is possible to derive two sets of visual information from MCGM, initially a sequential series of frames are anlaysed by MCGM algorithm, calculating the relative motions between successive frames. Additionally, a current frame of data is processed against a fixed open mouth frame, calculating the absolute motions of the mouth. MCGM processing results in a matrices of equal size to the input frames, each containing speed and angular information for a given pixel. Applying (linear) Principal Component Analysis (PCA) produces a linear space onto which the motions can be mapped, reducing the dimensionality of the visual features.

Mel-Frequency Cepstral Coefficients (MFCC) are classical acoustic speech features used in automatic speech processing [16]. They are state-of-the-art features in many applications, including automatic speech recognition and speaker verification systems. For obtaining a MFFC feature vector, the voice signal is transformed into the frequency domain via windowed Fast Fourier Transform and then mapped on to the Mel scale, a human perceptual scale of frequency [16]. A (logarithmically spaced) filter bank is constructed over this Mel frequency spectrum, and from this the logarithm of the power spectrum is determined. A discrete time cosine transform is performed over the power spectrum and the MFCCs are calculated. Most of the information about human voice from speech can be captured by retaining 10-12 most significant MFCC features, the first-order time-derivatives(delta features), the pitch and the signal energy.

To account for the lack of synchronization between speech features and lip motion features, rate interpolation can be done by up sampling the MCGM features to obtain the synchronized MCGM-MFCC features. Once the acoustic MFCC features and MCGM lip motion features are obtained, kCCA is implemented by first mapping them onto the kernel space using polynomial kernels and then performing CCA. Since, the kCCA involves, implementing CCA in a higher dimensional nonlinear space, it has the capability to capture and track the nonlinear correlations between different features. Parameter tuning for kCCA can be performed offline on an independent data set.

For extracting the mutually independent components of the audio and visual signals, another powerful statistical technique called independent component analysis (ICA) is performed, which treats the observed variables as a mixture of independent sources. Two different approaches can be used for Independent Component Analysis, ICA1 and ICA2 [17, 18]. In ICA1, the basis images are independent, whereas in ICA2 the mixing coefficients are independent. We utilize the ICA2 approach, where each pixel for lip images are considered as a mixture of independent coefficients. If X is a data matrix incorporating the measured variables, then it can be split as: X = AS where A is the mixing matrix and S contains the independent coefficients. The columns of A form a basis for the database and the columns of S provide ICA-features for the corresponding lip images residing in the columns of the data matrix X.

For each pixel, all x and y coordinates of a lip image are concatenated to a single vector. Its dimensionality is then reduced by applying PCA to the training set of x-y co-ordinate vectors. Each face is then represented by the first K PCA coefficients. The columns of the data matrix X for the ICA analysis are constituted of PCA coefficient vectors. Then, the Fast ICA algorithm described by [17, 18] is applied to obtain the basis A and the independent coefficients S. Next section describes the subsequent fusion technique used to combine various features.

4. Bayesian Audio-Visual Fusion

First, we derive the algorithm for performing the Bayesian fusion for liveness checks using multiple features described in the previous Section. Let us denoted the projection of audio and lip features in each of the closely coupled (kCCA), and mutually independent (ICA) subspaces as f_{kCCA} and f_{ICA}. We also include the projection of visual information in the PCA subspace as Eigenlip features f_{PCA} as the static spatial information in face images contains identity specific information. In Bayesian framework, the most generic way of performing the fusion is to compute the joint scores expressed as a weighted summation [19, 20]:

$$\rho(\lambda_r) = \sum_{n=1}^{N} w_n \log P(f_n \mid \lambda_r)$$
$$for\ r = 1,2,\ldots,R \quad (1)$$

where $\rho_n(\lambda_r)$ is the logarithm of the class-conditional probability, $P(f_n|\lambda_r)$, for the n^{th} modality f_n given class λ_r, and w_n denotes the weighting coefficient for modality n, such that $\Sigma_n w_n = 1$. Here f_n could be f_{kCCA}, f_{ICA} or f_{PCA} features. Then the fusion

problem reduces to a problem of finding the optimal weight coefficients for the nonlinear highly correlated components, loosely coupled f_{PCA} components and mutually independent f_{ICA} components. Though an adaptive fusion weight calculation would be ideally required, we selected the weights empirically and fused them using RWS (Reliability Weighted Summation) rule [19]. Since the statistical and the numerical range of these likelihood scores can vary from one modality to another, the likelihood scores were normalised within the (0, 1) interval before the RWS fusion process using a sigmoid and variance normalization as described in [20].

4. Experimental Results

Preliminary experimental results with an audio-visual speaking face video corpora VidTIMIT [21] and DaFEx [22,23] showed a significant improvement in liveness checking performance due to the detailed modelling of speaker liveness based on multiple correlation features. Figure 1 shows some images from the two corpora. The details of the two corpora are given in [21] , [22] and [23].

In this section, different experiments conducted to evaluate the performance of the proposed correlation features, and the Bayesian fusion of the MFCC, lip features in different subspaces PCA, kCCA and ICA for liveness checking are described. The testing stage for the liveness checking scenario is different from the tradition biometric identity verification scenarios, where the replay attack test data emulating fraudulent attacks needs to be artificially synthesised. Two different types of replay attacks were tested, one static replay attacks used in and other dynamic replay attacks, where artificial speaking face sequences are synthesised from still photo, few key frames from the video sequences, lip-synched with pre-recorded speech signals.

Liveness checking experiments involved two phases, the training phase and testing phase. In the training phase a 10-mixture Gaussian mixture model λ of a client's audiovisual feature vectors was built, reflecting the probability densities for the combined phonemes and visemes (lip shapes) in the audiovisual feature space. In the testing phase, the clients' live test recordings were first evaluated against the client's model λ by determining the log likelihoods log $p(X|\lambda)$ of the time sequences X of audiovisual feature vectors under the usual assumption of statistical independence of successive feature vectors.

(a) VidTIMIT corpus images

(b) DaFeX corpus images

Figure 1: Face Images from VidTIMIT and DaFex Corpus

For testing static replay attacks, a number of "fake" or synthetic 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 and keeping that visual feature vector constant throughout the utterance. Such a synthetic sequence represents an attack on the authentication system, carried out by replaying an audio recording of a 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. In order to obtain suitable thresholds to distinguish live recordings from fake recordings, detection error trade-off (DET) curves and equal error rates (EER) were determined. For testing dynamic replay attacks artificially synthesized speaking face video sequences were used instead of actually recorded video sequences in the data corpora.

Since the liveness checking is a two-class decision task, the system can make two types of errors. The first type of error is a False Acceptance Error (FA), where an impostor (fraudulent replay attacker) is accepted. The second error

is a False Rejection (FR), where a true claimant (genuine client) is rejected. Thus, the performance is measured in terms of False Acceptance Rate (FAR) and False Reject Rate (FRR), as defined as (Eqn. 2) :

$$FAR\ \% = \frac{I_A}{I_T} \times 100\ \%$$
$$FRR\ \% = \frac{C_R}{C_T} \times 100\ \% \qquad (2)$$

where I_A is the number of impostors classified as true claimants, I_T is the total number of impostor classification tests, C_R is the number of true claimants classified as impostors, and C_T is the total number of true claimant classification tests. The implications of this is minimizing the FAR increases the FRR and vice versa, since the errors are related. The trade-off between FAR and FRR is adjusted using the threshold θ, an experimentally determined speaker-independent global threshold from the training/enrolment data. The trade-off between FAR and FRR can be graphically represented by a Receiver Operating Characteristics (ROC) plot or a Detection Error Trade-off (DET) plot. The ROC plot is on a linear scale, while the DET plot is on a normal-deviate logarithmic scale. For DET plot, the FRR is plotted as a function of FAR. To quantify the performance into a single number, the Equal Error Rate (EER) is often used. Here the system is configured with a threshold, set to an operating point when FAR % = FRR %.

It must be noted that the threshold θ can also be adjusted to obtain a desired performance on test data (data unseen by the system up to this point). Such a threshold is known as the aposteriori threshold. However, if the threshold is fixed before finding the performance, the threshold is known as the apriori threshold. The apriori threshold can be found via experimental means using training/enrolment or evaluation data, data which has also been unseen by the system up to this point, but is separate from test data.

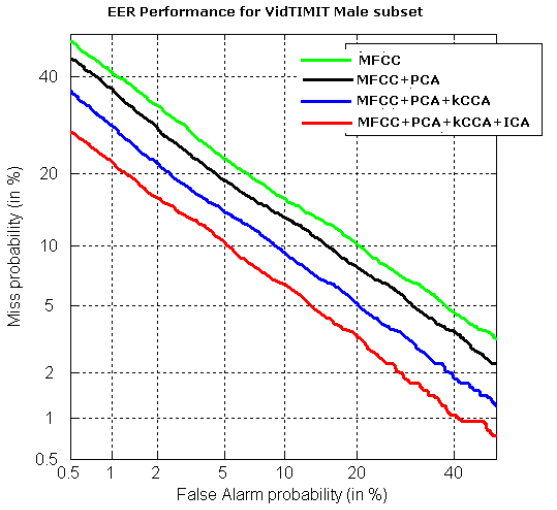
Practically, the a priori threshold is more realistic. However, it is often difficult to find a reliable apriori threshold. The test section of a database is often divided into two sets: evaluation data and test data. If the evaluation data is not representative of the test data, then the apriori threshold will achieve significantly different results on evaluation and test data. Moreover, such a database division reduces the number of verification tests, thus decreasing the statistical significance of the results. For these reasons, many researchers prefer to use the aposteriori and interpret the performance obtained as the expected performance.

Different sets of experiments were conducted to evaluate the performance of the audio-visual correlation features based on proposed mutual dependency models (kCCA, PCA and ICA), and their fusion, The performance evaluation in terms of DET curves and equal error rates (EER) for different features based on mutual dependency models in terms of DET curves and EERs is shown n Table 1 and Figure 2.

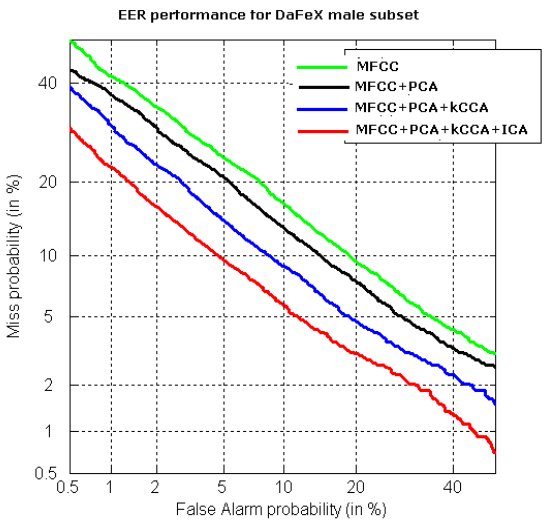
Table 1: EERs for audio visual features based on mutual dependency models

	VidTIMIT		DaFeX	
Audio/Visual Features	MALE EER (%)	FEMALE EER (%)	MALE EER (%)	FEMALE EER (%)
f_{mfcc}	16.8	16.88	15. 7	15. 7
f_{eigLip}	16.2	16.2	16.64	16.64
f_{MGCM}	17.2	17.87	15.9	15.54
f_{kCCA}	14.7	15.18	14.81	15.28
f_{ICA}	13.03	14.12	13.12	14.4
$f_{mfcc} + f_{eigLip}$	11.68	11.86	11.79	11.17
$f_{mfcc} + f_{eigLip} + f_{kCCA}$	10.26	10.26	10.46	10.46
$f_{mfcc} + f_{eigLip} + f_{kCCA} + f_{ICA}$	**8.06**	**8.85**	**9.23**	**9.31**

As can be seen from Table 1 and Figure 2 the results are quite promising for correlation features in kCCA space and their fusion with features in ICA and PCA space. The single mode MFCC features and PCA or Eigen lip features results in worse EERS. Further, the MGCM features on their own do not result in a good EER performance. However, when they are fused with the kCCA projected features, they result in improved performance. Further, use of correlation features in different subspaces, PCA, ICA and kCCA result in best EERs as complete mutual dependency components (closely coupled, loosely coupled and uncoupled components are included in the modelling). Further work involves, developing an automatic fusion computation technique based on reliability scores.



(a)



(b)

Figure 2: DET curves for audio visual features based on mutual dependency models for (a): VidTIMIT data set, (b): DaFeX dataset

5. Conclusions

In this paper we proposed a novel method of extracting audio visual features based on correlation models for checking in biometric identity authentication systems. Performance evaluation in terms of DET curves and EERs on VidTIMIT and DaFeX corpora, showed a significant improvement in performance of proposed features as compared to traditional single mode face or voice features.

6. References

[1] Gerasimos Potamianos, Chalapathy Neti, Juergen Luettin, and Iain Matthews. Audio-Visual Automatic Speech Recognition: An Overview. Issues in Visual and Audio-Visual Speech Processing, 2004.

[2] Xiaoxing Liu, Luhong Liang, Yibao Zhaa, Xiaobo Pi, and Ara V. Nefian. Audio-Visual Continuous Speech Recognition using a Coupled Hidden Markov Model. In Proc. International Conference on Spoken Language Processing, 2002.

[3] Sabri Gurbuz, Zekeriya Tufekci, Tufekci Patterson, and John N. Gowdy. Multi-Stream Product Modal Audio-Visual Integration Strategy for Robust Adaptive Speech Recognition. In Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, Orlando, 2002.

[4] Claude C. Chibelushi, Farzin Deravi, and John S.D.Mason. A Review of Speech-Based Bimodal Recognition. IEEE Transactions on Multimedia, 4(1):23–37, 2002.

[5] Hao Pan, Zhi-Pei Liang, and Thomas S. Huang. A New Approach to Integrate Audio and Visual Features of Speech. In Proc. IEEE International Conference on Multimedia and Expo., pages 1093 – 1096, 2000.

[6] U.V. Chaudhari, G.N. Ramaswamy, G. Potamianos, and C. Neti. Information Fusion and Decision Cascading for Audio-Visual Speaker Recognition Based on Time- Varying Stream Reliability Prediction. In IEEE International Conference on Multimedia Expo., volume III, pages 9 – 12, Baltimore, USA, July 2003.

[7] Chetty G., and Wagner M,, Robust face-voice based speaker identity verification using multilevel fusion, Image and Vision Computing, Volume 26, Issue 9, 1 September 2008, Pages 1249-1260.

[8] R.Goecke and J.B. Millar. Statistical Analysis of the Relationship between Audio and Video Speech Parameters for Australian English. In J.L. Schwartz, F. Berthommier, M.A. Cathiard, and D. Sodoyer (eds.), Proceedings of the ISCA Tutorial and Research Workshop on Auditory-Visual Speech Processing AVSP 2003, pages 133-138, St. Jorioz, France, 4 - 7 September 2003.

[9] S. Molholm, et al., "Multisensory Auditory-visual Interactions During Early Sensory Processing in Humans: a high-density electrical mapping study," Cognitive Brain Research, vol. 14, pp. 115-128, June 2002.

[10] J. MacDonald, & H. McGurk, “Visual influences on speech perception process”. Perception and Psychophysics, 24, 253-257, 1978.

[11] J.Jiang, A. Alwan, P.A.Keating, E.T. Auer Jr., L. E. Bernstein, “On the Relationship between Face Movements, Tongue Movements, and Speech Acoustics,” EURASIP Journal on Applied Signal Processing 2002:11, 1174–1188.

[12] H. C. Yehia, T. Kuratate, and E. Vatikiotis-Bateson, “Using speech acoustics to drive facial motion,” in Proc. the 14th International Congress of Phonetic Sciences, pp. 631–634, San Francisco, Calif, USA, 1999.

[13] H. Hotelling. “Relations between two sets of variates.” Biometrika, 28:321 377, 1936.

[14] D. R. Hardoon, S. Szedmak and J. Shawe-Taylor, “Canonical Correlation Analysis: An Overview with

Application to Learning Methods", in Neural Computation Volume 16, Number 12 2004, Pages 2639–2664.

[15] P.W. McOwan, and A. Johnston, "The algorithms of natural vision: The Multi-channel Gradient Model". Proc. IEE/IEEE Genetic Algorithms in Engineering Systems. Sept' 95.

[16] S. Chauhan and P. Wang and C.S. Lim and V. Anantharaman "A computer-aided MFCC-based HMM system for automatic auscultation" Comput. Biol. Med., Vol. 38, No. 2, 2008, Pages 221–233.

[17] K. W. Bowyer, K. Chang, and P. Flynn, "A survey of approaches and challenges in 3D and multi-modal 3D + 2D face recognition," Computer Vision and Image Understanding, vol. 101, no. 1, pp. 1–15, 2006.

[18] A. Srivastava, X. Liu, and C. Hesher, "Face recognition using optimal linear components of range images," Image and Vision Computing, vol. 24, no. 3, pp. 291–299, 2006.

[19] V. Nefian, L. H. Liang, X. Pi, X. Liu, and K. Murphy, "Dynamic Bayesian Networks for Audio-visual Speech Recognition," EURASIP Journal on Applied Signal Processing, vol. 2002, pp. 1274-1288, Nov. 2002.

[20] Movellan, J. and Mineiro, P., "Bayesian robustification for audio visual fusion". In Proceedings of the 1997 Conference on Advances in Neural information Processing Systems 10 (Denver, Colorado, United States). M. I. Jordan, M. J. Kearns, and S. A. Solla, Eds. MIT Press, Cambridge, MA, 742-748, 1998.

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

[22] Battocchi, A.; Pianesi, F.. 2004. DaFEx: Un Database di Espressioni Facciali Dinamiche. In Proceedings of the SLI-GSCP Workshop, Padova (Italy) 30 Novembre - 1 Dicembre 2004.

[23] Mana N., Cosi P., Tisato G., Cavicchio F., Magno E. and Pianesi F., An Italian Database of Emotional Speech and Facial Expressions, In Proceedings of "Workshop on Emotion: Corpora for Research on Emotion and Affect", in association with "5th International Conference on Language, Resources and Evaluation (LREC2006), Genoa, Italy, 24-25-26 May 2006.