

This is an extract from:

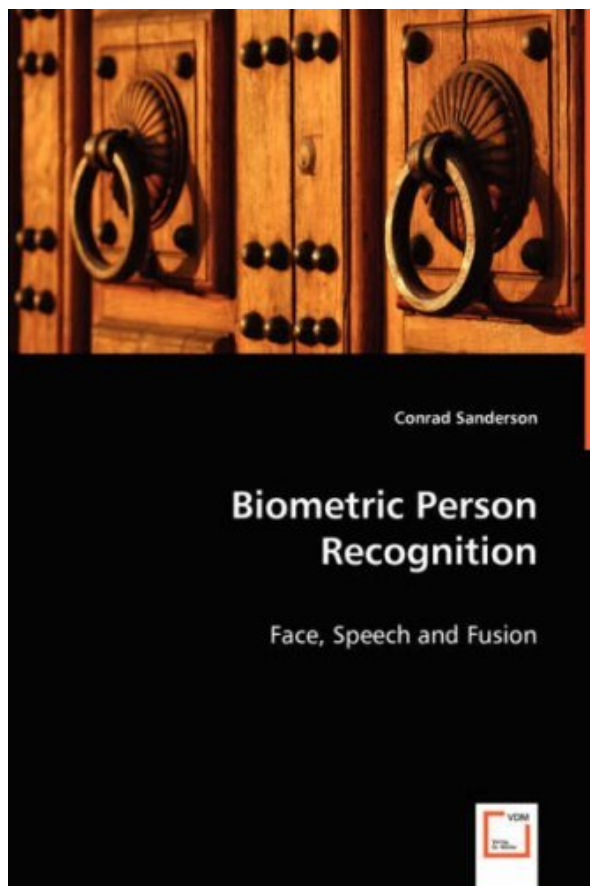
Conrad Sanderson.

**Biometric Person Recognition: Face, Speech and Fusion.**

VDM Verlag, 2008.

ISBN 978-3-639-02769-3.

<http://www.amazon.com/dp/3639027698?tag=bookref-20>



## Chapter 1

### Introduction

Identity verification systems are part of our every day life – one example is the Automatic Teller Machine (ATM) which employs a simple identity verification scheme: the user is asked to enter their<sup>1</sup> password after inserting their card. If the password matches the one prescribed to the card, the user is allowed access to their bank account. This scheme suffers from a drawback: only the validity of the combination of a certain possession (the ATM card) and certain knowledge (the password) is verified. The ATM card can be lost or stolen, and the password can be compromised. New verification methods have hence emerged, where biometrics such as the person's speech, face image or fingerprints can be used in addition to the password. Such biometric attributes cannot be lost and typically vary considerably from person to person.

Apart from the ATM example described above, biometrics can be applied to other areas, such as telephone & internet based banking, passport control (immigration checkpoints), as well as forensic work (to determine whether a biometric sample belongs to a suspect) and law enforcement applications (e.g. surveillance) [11, 35, 44, 67, 98, 105, 112, 128, 190].

While biometric systems based on face images and/or speech signals can be effective [111, 128, 149], their performance can degrade in the presence of challenging conditions. For speech based systems this is usually in the form of channel distortion and/or ambient noise. For face based systems it can be in the form of a change in the illumination direction and/or a change in the pose of the face [3, 159].

Multi-modal systems use more than one biometric at the same time. This is done for two main reasons: **(i)** to achieve better robustness (where the impact of a biometric affected by environmental conditions can be decreased) and **(ii)** to increase discrimination power (as complementary information can be used). Multi-modal systems are often comprised of several modality experts and a decision stage [22].

This work overviews relevant backgrounds and reports research aimed at increasing the robustness of single- and multi-modal biometric verification systems, in particular those based on speech and face modalities.

<sup>1</sup>The word 'their' is used as gender neutral replacement of the words 'his' and 'her' [125].

## 1.1 Overview

This work is comprised of three major parts: (i) verification using speech signals (Chapter 3), (ii) verification using face images (Chapters 4 and 5), (iii) verification using fused speech and face information (Chapter 6). It is supported by Chapter 2, which provides an overview of relevant pattern recognition theory. To ease reading, each chapter is largely self contained. The chapters are summarised as follows:

- **Chapter 2 – Statistical Pattern Recognition** – first draws distinctions between closed set identification, open set identification and verification. Relevant pattern recognition theory is then used to derive a two-class decision machine (classifier) used in the verification system. The machine is implemented using the Gaussian Mixture Model (GMM) approach. The  $k$ -means, Expectation Maximisation (EM) and maximum a-posteriori (MAP) adaptation algorithms, which are used for finding GMM parameters, are described. Two methods for finding the impostor likelihood are presented: the Background Model Set (BMS) and Universal Background Model (UBM). Next, error measures for finding the performance of a verification system, such as the Equal Error Rate (EER), are described. The chapter is concluded by a discussion on implementation issues, where practical limitations and experimental requirements are taken into account. The implementation of the decision machine is tested in the following chapter.
- **Chapter 3 – Verification using Speech Signals** – first describes the difference between text-dependent and text-independent systems. A review of the human speech production process is then given, followed by a review of feature extraction approaches used in a speaker verification system. Mel Frequency Cepstral Coefficients (MFCCs), delta features and Cepstral Mean Subtraction (CMS) are covered. An alternative feature set, termed Maximum Auto-Correlation Values (MACVs), which uses information from the source part of the speech signal, is also covered. A parametric Voice Activity Detector (VAD), used for disregarding silence and noise segments of the speech signal, is briefly described. The implementation of the Gaussian Mixture Model classifier (described in Chapter 2) is tested. The use of MACVs is evaluated for reducing the performance degradation of a verification system used in noisy conditions.
- **Chapter 4 – Verification using Frontal Face Images** – first overviews important publications in the field of frontal face recognition. Geometric features, templates, Principal Component Analysis (PCA), pseudo-2D Hidden Markov Models (HMM), Elastic Graph Matching (EGM), as well as other points are covered. Relevant issues, such as the effects of an illumination direction change and the use of different face areas, are also covered. Several new feature extraction approaches are proposed – their robustness and performance is evaluated against three popular methods (PCA, 2D DCT and 2D Gabor wavelets) for use in an identity verification system subject to illumination direction changes. It is also shown that when using the GMM classifier with local

features (such as Gabor wavelets or DCT derived features), the spatial relationships between face parts (e.g. eyes and nose) are disregarded. Such a face recognition system can surprisingly still provide good performance and provides advantages such as robustness to translations. The fragility of PCA derived features to illumination changes is addressed by introducing a pre-processing step which involves applying local feature extraction to the original face image – it is shown that the enhanced PCA technique is robust to illumination changes as well as retaining the positive aspects of traditional PCA, i.e. robustness to compression artefacts and noisy images, which might be important in forensic and law enforcement applications.

- **Chapter 5 – Verification using Faces with Pose Variations** – deals with faces subject to pose variations, in contrast to the previous chapter which dealt with frontal faces. A framework is developed for addressing the pose mismatch problem that occurs when there is only a single (frontal) face image available for training and non-frontal faces are presented during testing. In particular, the mismatch problem is tackled through building multi-angle models by extending each frontal face model with artificially synthesised models for non-frontal views. The synthesis methods are based on several implementations of Maximum Likelihood Linear Regression (MLLR), as well as standard multi-variate linear regression (LinReg). We stress that instead of synthesising images, model parameters are synthesised. The synthesis and extension approach is evaluated by applying it to two face verification systems: a holistic system (based on PCA-derived features) and a local feature system (based on DCT-derived features). It is also shown that the local feature system is less affected by view changes than the holistic system.
- **Chapter 6 – Verification Using Fused Speech and Face Information** – first provides an overview of key concepts in the information fusion area, followed by a review of important milestones in audio-visual person identification and verification. Several adaptive and non-adaptive techniques for reaching the verification decision, based on combined speech and face information, are then evaluated in clean and noisy audio conditions on a common dataset. It is shown that in clean conditions most of the non-adaptive approaches provide similar performance and in noisy conditions most exhibit a severe deterioration in performance. It is also shown that current adaptive approaches are either inadequate or use restrictive assumptions. A new category of classifiers is then introduced, where the decision boundary is fixed but constructed to take into account how the distributions of opinions are likely to change due to noisy conditions. Compared to a previously proposed adaptive approach, the proposed classifiers do not make a direct assumption about the type of noise that causes the mismatch between training and testing conditions.