Image and Vision Computing xxx (2013) xxx–xxx

Contents lists available at ScienceDirect

Image and Vision Computing

journal homepage: www.elsevier.com/locate/imavis

Bi-modal biometric authentication on mobile phones in challenging conditions[☆]

Elie Khoury [a,*], Laurent El Shafey [a], Christopher McCool [b], Manuel Günther [a], Sébastien Marcel [a]

[a] Idiap Research Institute, Martigny, Switzerland
[b] NICTA, Queensland Research Laboratory, Australia

ARTICLE INFO

Article history:
Received 30 April 2013
Accepted 3 October 2013
Available online xxxx

Keywords:
Face authentication
Speaker authentication
Bi-modal authentication
Gaussian mixture model
Session variability
Inter-session variability
Total variability
I-vector
Fusion

ABSTRACT

This paper examines the issue of face, speaker and bi-modal authentication in mobile environments when there is significant condition mismatch. We introduce this mismatch by enrolling client models on high quality biometric samples obtained on a laptop computer and authenticating them on lower quality biometric samples acquired with a mobile phone. To perform these experiments we develop three novel authentication protocols for the large publicly available MOBIO database. We evaluate state-of-the-art face, speaker and bi-modal authentication techniques and show that inter-session variability modelling using Gaussian mixture models provides a consistently robust system for face, speaker and bi-modal authentication. It is also shown that multi-algorithm fusion provides a consistent performance improvement for face, speaker and bi-modal authentication. Using this bi-modal multi-algorithm system we derive a state-of-the-art authentication system that obtains a half total error rate of 6.3% and 1.9% for Female and Male trials, respectively.

1. Introduction

Mobile phones have become an integral part of many people's daily life. They are used not just for telephonic communication, but also to send and receive emails, take photos or even have video conversations. This has led to the mobile phone being an inherently multimedia device, which often has a front-facing camera in addition to the standard microphone. Hence, it forms an exciting new device that allows researchers to explore the applicability of bi-modal (face and speaker) authentication in challenging mobile phone environments.

This exciting challenge of bi-modal authentication in the mobile phone environment has begun to receive more attention. An international competition was organised in 2010 [1], where researchers evaluated state-of-the-art algorithms for face and speaker authentication using Phase I of the MOBIO database [2]. In this evaluation, enrolment was exclusively performed with mobile phone data. It was shown that a combination of these systems produced an impressive bi-modal authentication system. Since then other researchers have examined methods to perform face [3,4], speaker [5,6] and bi-modal [7,8] authentication in the challenging mobile phone environment.

A theme common to some of the prior work on biometric authentication in a mobile environment is the idea of *session variability* modelling [5,4], which achieves state-of-the-art results for bi-modal authentication [8]. Session variability modelling aims to estimate and suppress any variability such as audio or image noise that may cause confusion between different observations of the same biometric identity. In [5] session variability modelling was used to cope with audio channel variability, while [4] introduced this concept to face authentication, where its application was supposed to reduce the impact of pose and illumination variation. Finally, in [8] state-of-the-art face and speaker authentication systems that used *inter-session variability* (ISV) modelling were combined to derive a state-of-the-art bi-modal authentication system. However, this prior work applied ISV modelling in the case of matched acquisition conditions, i. e., where biometric test samples are acquired using the same device as employed for client model enrolment. Furthermore, they did not use the most recent advances such as *total variability* (TV) modelling, which has been applied to speaker [9] and face [10] authentication.

In this paper we explore three issues of applying bi-modal authentication to the challenging mobile phone environment. First, we examine the issue of mismatched conditions between enrolment and testing. In particular, we examine the effect of enrolling users on high quality biometric samples acquired with a laptop computer and then authenticating them using lower quality biometric samples acquired with a mobile phone. As a significant contribution, we develop three new protocols for the MOBIO database [2] with respect to prior work

[☆] This paper has been recommended for acceptance by Lei Zhang and Stefanos Zafeiriou.
[*] Corresponding author.
E-mail addresses: elie.khoury@idiap.ch (E. Khoury), laurent.el-shafey@idiap.ch (L. El Shafey), chris.mccool@nicta.com.au (C. McCool), manuel.guenther@idiap.ch (M. Günther), marcel@idiap.ch (S. Marcel).

0262-8856/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.imavis.2013.10.001

[1,2,4,8] that was exclusively using mobile phone data both for enrolment and testing.

Second, we extend the work of [8] by examining the effectiveness of TV modelling for bi-modal authentication. Third, we show the effectiveness of multi-algorithm fusion to further improve the results for face, speaker and bi-modal authentication in the mobile phone environment. The final outcome of this work is the development of a state-of-the-art bi-modal (face and speaker) authentication system that improves upon the previous state-of-the-art [8] with a relative performance gain of 35% for Female and 27% for Male trials on the MOBIO database.

The remainder of this paper is structured as follows: In Section 2 we outline the employed face and speaker authentication systems, while Section 3 combines these into bi-modal and multi-algorithm authentication systems. Section 4 presents the new protocols for the MOBIO database that are used in our experiments, which we discuss in Section 5. Finally, Section 6 concludes the paper.

2. Face and speaker authentication systems

We examine the effectiveness of state-of-the-art *Gaussian mixture model* (GMM) based approaches for face, speaker and bi-modal authentication. GMMs have formed the basis of state-of-the-art speaker authentication systems for over a decade [11,9] and it was recently shown that incorporating session variability modelling into a GMM system produces state-of-the-art results for face authentication [12]. Also, the combination of GMM-based systems that use session variability modelling produces a state-of-the-art bi-modal (face and speaker) authentication system [8].

When using GMMs and session variability for speaker and face authentication, the same underlying approach is taken. The main difference is how the feature vectors are extracted from the image (face) and audio (speech) samples. Below we describe the feature extraction process for both face and speaker authentication followed by a description of the GMM and the associated session variability modelling approaches that we examine.

2.1. Feature extraction

Two separate feature extraction processes are used for image (face) and audio (speech) data. For both modalities, a biometric sample $\mathcal{O}$ (image or audio) is decomposed into a set O of K feature vectors ($O = \{o^1, o^2, \ldots, o^K\}$), where each feature vector is of dimensionality M. This decomposition is performed in the spatial domain for the image data, and in the time domain for the audio data.

2.1.1. Face-based features

For the image data, we rely on parts-based features that were proposed for the task of face authentication in [13]. These features have since been successfully employed by several researchers [14,15]. The key idea is to decompose the face image into a set of overlapping blocks before extracting a feature vector from each of them. The feature vectors extracted from these blocks are then considered as observations of the same signal (the same face), and can be modelled in a generative way.

The feature extraction process is similar to the approach described in [16]. First, each image is rotated, scaled and cropped to 64 × 80 pixels such that the eyes are 16 pixels from the top and separated by 33 pixels. Second, to reduce the impact of illumination, each cropped image is preprocessed with the multi-stage algorithm of Tan & Triggs [17], using their default parameterisation. Third, 12 × 12 blocks of pixel values are extracted from the preprocessed image using an exhaustive overlap, leading to $K = 3657$ blocks per image. Fourth, pixel values of each block are normalised to zero mean and unit variance, prior to extracting the $M + 1$ lowest frequency *2D discrete cosine transform* (2D-DCT) coefficients [13] and removing the zero frequency coefficient as it is redundant. Fifth, the resulting 2D-DCT feature vectors are normalised to zero mean and unit variance in each dimension with respect to the other feature vectors of the image. As in previous work [16,8], M was set equal to 44.

2.1.2. Speaker-based features

For the audio data, observations are extracted at equally-spaced time instants using a sliding window approach. First, audio segments are denoised using the Qualcomm-ICSI-OGI front end [18]. Second, *voice activity detection* (VAD) is performed jointly using the normalised log energy and the 4 Hz modulation energy [19]. The aim of the 4 Hz modulation energy is to discriminate speech from other audio sources such as noise and music. An adaptive threshold is applied on both the 4 Hz modulation energy and the normalised log energy. In our experiments, this approach provided a relative improvement of up to 16% compared to the common energy-based VAD. Third, 19 *mel frequency cepstral coefficient* (MFCC) and log energy features together with their first- and second-order derivatives are obtained by computing 24 filter bank coefficients over 20 ms Hamming windowed frames every 10 ms. This results in acoustic feature vectors of dimensionality $M = 60$. Finally, feature normalisation based on *cepstral mean and variance normalisation* (CMVN) is applied on the remaining speech. The number of feature vectors K extracted from each audio sample depends on the duration of the sample and the number of segments that the VAD classifies to be speech.

2.2. GMM-based modelling

We use the same generative probabilistic framework that models the observed feature vectors using *Gaussian mixture models* (GMMs) for both image (face) and audio (speech) modalities. GMMs have been successfully applied first to speaker authentication [20,11] and then to face authentication [13,21,14–16]. One of the main challenges with GMMs is to reliably estimate a client model with limited enrolment data. This enrolment process is sensitive to the conditions, in which the data was captured. To address this issue, several session variability modelling techniques built on the GMM baseline have been proposed that constrain client models to be in a restricted subspace. In this work, we consider two approaches to session variability modelling, *inter-session variability* (ISV) modelling [22] and *total variability* (TV) modelling [9]. Both methods were initially proposed for speaker authentication [9,22] before being applied to face authentication [10,12]. In the remainder of this section, we first describe the GMM baseline system, followed by the more advanced ISV and TV techniques.

2.2.1. Gaussian mixture modelling

The distribution of the observed feature vectors (face or speech) is modelled using a GMM. A GMM is the weighted sum of C multi-variate Gaussian components $\mathcal{N}$:

$$p\left(o\middle|\Theta_{gmm}\right) = \sum_{c=1}^{C} \omega_c \mathcal{N}\left(o; \mu_{c,} \sum\nolimits_c\right), \tag{1}$$

where $\Theta_{\text{gmm}} = \{\omega_c, \mu_c, \sum_c\}_{c=\{1,\ldots,C\}}$ are the parameters of this distribution: the weights, the means and the covariance matrices, respectively.

To use GMMs for authentication we need to learn a GMM $\mathcal{S}_i$ for each subject i from a set of enrolment samples. One of the main challenges is that the number of enrolment images or audio recordings per client is usually limited, possibly to a single sample. In practise, it has been shown that for both speaker [11] and face authentication [14,15] an efficient enrolment method is to use a subject-independent prior GMM $\mathcal{M}$, called the *universal background model* (UBM), and to adapt this prior to the enrolment samples of the subject i to generate the client model $\mathcal{S}_i$. The UBM $\mathcal{M}$ is learnt beforehand by maximising the likelihood of observations extracted from a large independent training set of several identities using the iterative *expectation–maximisation* (EM)

Please cite this article as: E. Khoury, et al., Bi-modal biometric authentication on mobile phones in challenging conditions, Image Vis. Comput. (2013), http://dx.doi.org/10.1016/j.imavis.2013.10.001