Facial Dynamics in Biometric Identification

L. Benedikt [1], V. Kajic [1], D. Cosker [2], P. L. Rosin [1], D. Marshall [1]
[1] School of Computer Science, Cardiff University, UK
[2] School of Computer Science, University of Bath, UK
l.truong-benedikt@cs.cardiff.ac.uk

Abstract

This paper investigates the use of facial gestures for identity recognition. This is the first time that such a quantitative evaluation is conducted, comparing the analyses of 2D versus 3D dynamic data of verbal and non-verbal facial actions. Suitable data processing and feature extraction methods are examined, then a number of pattern matching techniques including the Fréchet distance, Correlation Coefficients, Hidden-Markov Models, Dynamic Time Warping and its derived forms are compared, in light of which an improved algorithm is proposed. Finally, a face recognition prototype using facial dynamics is built, achieving an Equal Error Rate EER=1.6%.

1 Introduction

Biometric identification refers to the recognition of individuals based on their physiological characteristics (e.g. fingerprint, iris, DNA) or behavioral peculiarities (e.g. voice, gait, signature). While the robustness of behavioural biometrics has often been questioned with regard to their sensitivity to human emotional conditions, physiological biometrics also exhibit limitations, being either too expensive and intrusive (iris, DNA), or not accurate enough for high security applications (mug-shots). For this reason, there is still a great deal of interest in exploring novel biometric features [8].

The human face was originally regarded as a physiological biometric only. Early works on face recognition mainly relied on the well-established 2D methods such as Eigenfaces [16] and Active Appearance Model [4], until advances in imaging technologies made it possible to develop more accurate 3D-based algorithms such as the Morphable Model [2]. Both 2D and 3D algorithms however share one common drawback which is their sensitivity to face expressions. While Chang *et al.* [3] proposed an expression invariant recognition technique, Luettin *et al.* [11] exploited the behavioural aspect of facial dynamics and used lipreading for speaker recognition. This latter approach has inspired a number of studies, of which the most recent is the work of Faraj *et al.* [7] where the authors extract the velocity of lip motions for speaker verification.

The objective of this study is to explore further the use of facial gestures for identity recognition, and compare the accuracies of analysing 2D versus 3D dynamic data. Unlike previous works which employed long audio-video sequences, we investigate *very short* facial actions in order to increase the processing speed, which is paramount for a real-time application. Such a constraint requires the use of pattern matching techniques

which can perform accurately on short data sequences. Therefore we will evaluate a number of pattern matching methods including the Fréchet distance, Correlation Coefficients, Hidden-Markov Models, Dynamic Time Warping and its derived forms Continuous DTW and Derivative DTW, in light of which we will propose an improved DTW-based algorithm combining weighted higher-order derivatives.

2 Data acquisition

Unlike related works where long sequences of uttered digits from '0' to '9' are analysed [7, 11], we seek to identify *very short* facial actions which satisfy the fundamental biometric requirements such as repeatability and distinctiveness [8]. Two directions have been considered. First, we investigate a number of facial Action Units (AUs) as described by the Facial Action Coding System [6], for example Brow Raiser (AU1+2), Upper Lid Raiser (AU5), Nose wrinkler (AU9) and Lip corner puller (AU12). Secondly, we conduct a literature review in order to identify a speech posture involving the most discriminative and reproducible facial movements [9]. The word *'Puppy'* was chosen not only for its reproducibility, but also because its utterance affects mainly the lips, which would give a good representation of facial movements in the lower third of the face while leaving stable the upper thirds, allowing accurate alignment of sequential frames.

A 3D video camera operating at 48 frames per second is used to capture 3D dynamic data from 36 participants (20 males and 16 females, 23 are natural English speakers). All face scans have an ear-to-ear coverage and no physical markers are used. Each participant is asked to utter the word *'Puppy'* several times in a normal and relaxed way, and to perform an AU12-based standardised smile. Smaller tests are performed for other AUs. At this stage, we are still investigating suitable facial actions, so we are not able to conduct tests on a large database. However, we believe that our data is statistically significant since it is triple the size of the database used in the work of Luettin *et al.* [11].

During the recording sessions, we observe that it is very difficult for non-actors to produce accurate AUs, let alone to repeat identical performances. For example, although the participants have been asked to produce an AU12-based maximum smile, some performances also exhibit cheek raiser (AU6) and lips part (AU25). Thus, AU-based expressions fail the repeatability requirement and do not seem functional in a real-life scenario. The *'Puppy'* utterance appears to suffer fewer fluctuations, however it is still plagued by a number of unwanted actions such as blinking (AU45) and gaze movements (AU61+62).

3 Data processing and feature extraction

To normalise the face scans, we employ the technique proposed by Chang *et al.* [3]: the nose root position and surface normal are computed using curvature analysis, then the 3D heads are registered through translations, rotations and scaling. Further refinement is achieved using the Iterative Closest Point algorithm [3], as depicted in Figure 1(a).

To assess the impact of unwanted facial actions (e.g. blinking) on the recognition performance, we wish to compare the results when analysing the lips only versus analysing the entire face. To this end, the lip region needs to be segmented. This can be achieved as shown in Figure 1(b). First, we use a thin-plate-spline process to warp a 3D lip template

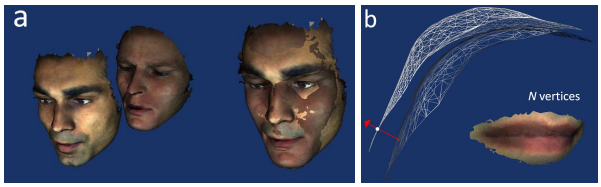

Figure 1: a) 3D alignment of video frames. b) Segmentation of the 3D lip subregion.

to the lip region of the face scan (let us call this region the target mesh). This operation although not very precise gives an adequate initial registration. We then project each vertex of the warped template along its normal towards the target mesh. Repeating this operation for all video frames, we obtain a new sequence of lip shapes in full correspondence (i.e. identical number of vertices and vertex topology).

3.1 Acquisition of artificial 2D data

Although representing faces in the 3D space is more accurate than in 2D, one should bear in mind a number of inconveniences inherent to 3D data analysis, for example the high cost of 3D data capture devices, as well as intensive processing and high data storage requirements. Therefore, we should assess whether there is a need to use 3D data, or facial dynamics read from 2D video provides already sufficient information for identity recognition. The challenge for such a comparative study is how to collect 2D dynamic data with identical frame rate and recording conditions than our 3D data. To this end, we artificially generate 2D data by projecting the 3D data onto a set of planes as shown in Figure 8. We chose multiple planes in order to simulate acquisition of 2D data from different view points, with and without occlusions. This test will be conducted for the fixed head situation only. For non-fixed head situations, there is already much evidence proving that 2D models cannot cope with head pose variations, while 3D models can overcome this problem [2, 3].

3.2 Feature extraction

The feature extraction is achieved by building a PCA shape model in similar fashion to the Active Shape Model [4] in 2D, and the Morphable Model [2] in 3D. In order to reduce the dimensionality of the feature space, we apply the same protocol to both 2D and 3D processes as follows: we retain only 90% of the variations, which involves keeping the p highest Eigenvectors. Thus, any shape $\overrightarrow{\mathbf{x}}_k$ can be approximated as:

$$\overrightarrow{\mathbf{x}}_k \approx |\overrightarrow{\mathbf{x}}| + \Phi \overrightarrow{\mathbf{v}}_k \tag{1}$$

where $|\overrightarrow{\mathbf{x}}|$ is the mean shape, Φ is the matrix of p Eigenvectors, and $\overrightarrow{\mathbf{v}}_k$ is the vector of the shape parameters. Inverting this equation, we can extract the shape variations of a sequence of D shapes as follows:

$$\overrightarrow{\mathbf{v}}_k \approx \Phi^T (\overrightarrow{\mathbf{x}}_k - |\overrightarrow{\mathbf{x}}|), \quad k \in \{1, ..., D\} \tag{2}$$

Figure 3(a) shows the trajectories of the facial dynamics extracted from a 3D video sequence of a *'Puppy'* utterance. The first three modes of variation (MoV) are depicted.

4 Similarity measures for dynamic signals

In this section, we will examine a number of methods to quantify the similarity S between two dynamic signals. In practice, it is more convenient to measure the distance D between the curves which represent these dynamic signals. S can then be defined as $S = \frac{1}{1+D}$.

4.1 Baselines for assessment of similarity measures

In this study, we use the Fréchet distance and Correlation Coefficients as baselines to assess the performances of further similarity measures because these formers are among the most classic techniques. However, such linear methods perform rather poorly when it comes to comparing dynamic signals which exhibit comparable global patterns but vary in time or speed, which is commonly the case for behavioural biometric features [5].

4.2 Dynamic Time Warping (DTW)

DTW is commonly used in speech recognition, and has achieved very good results when applied to signature authentication [5, 12]. The idea behind DTW consists of first computing the best alignment between the two signals by warping their time axes, before measuring their similarity. Details of the algorithm implementation can be found in [15].

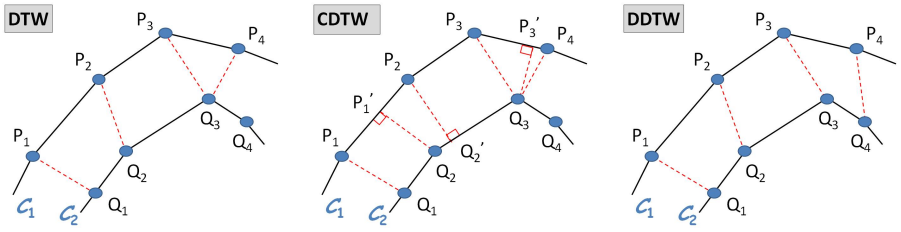

Figure 2: Curve matching using Dynamic Time Warping and its derived forms Continuous DTW (CDTW) and Derivative DTW (DTW).

Suppose we have two feature vectors $\overrightarrow{\mathbf{v}}_1$ and $\overrightarrow{\mathbf{v}}_2$ of lengths n and m, respectively depicted by curves C_1 and C_2 in Figure 2. Each point Q_j on C_2 is first warped to its closest point P_i of C_1, and the distance D between C_1 and C_2 is defined as the cumulated distance along the warping path. In other words, D is the sum of all pair-wise distances $d(P_i, Q_j)$, normalised by the path length. The complexity of DTW is $O(nm)$.

4.3 Continuous Dynamic Time Warping (CDTW)

The main limitation of DTW is that it maps two sequences using discrete points, thus its accuracy is sensitive to the sampling rate. Besides, the 'optimal' warping path found by DTW may not actually be the best. A method to improve DTW has been proposed by Munich *et al.* [12] and consists of extending the mapping process to the continuous domain. Figure 2 shows the result of CDTW where the data point Q_2 is mapped to a new point P_1' which is the projection of Q_2 onto the segment $\overline{P_1P_2}$. The subsequent similarity measurement remains similar to that of the classic DTW. The complexity of CDTW increases combinatorially and heuristic constraints are needed to make the problem tractable.

4.4 Derivative Dynamic Time Warping (DDTW)

One common weakness of DTW and CDTW is that they determine the warping path by considering distances between data points, regardless of the signal trends. Let us consider two data points which are close, but one belongs to a rising slope and the other to a falling slope, for example points P_4 and Q_3 in Figure 2. DTW and CDTW consider a mapping between P_4 and Q_3 optimal, although it would be inaccurate to map a rising trend to a falling trend. A new approach presented by Keogh *et al.* [10] proposes to compare the first order derivatives $\frac{d\vec{\mathbf{v}}_1}{dt}$ and $\frac{d\vec{\mathbf{v}}_2}{dt}$ instead of comparing the original signals. The resulted curve mapping is illustrated in Figure 2. The complexity of the algorithm is $O(nm)$.

4.5 Weighted Hybrid Derivative Dynamic Time Warping (WDTW)

Although the idea of considering the signal derivatives is interesting, our preliminary tests have shown that DDTW is very noise sensitive. Thus, we propose an improvement as follows. First, the signal itself contains useful information that should not be ignored, and smoothing this one helps stabilise the process. Secondly, incorporating further derivatives provides better knowledge, for example the first derivative gives information on speed and the second derivative on accelerations and decelerations. Thus, let us consider the quantity $[\vec{\mathbf{v}}, \frac{d\vec{\mathbf{v}}}{dt}, \frac{d^2\vec{\mathbf{v}}}{dt^2}]$ and compute the warping path using the hybrid distance:

$$d = w_0 * d_0 + w_1 * d_1 + w_2 * d_2 \tag{3}$$

where d_0 is the distance between the data points, d_i is the difference between the i^{th} derivatives, w_0,w_1 and w_2 are weights. Introducing weights allows the derivatives to play a role in the pattern recognition process in case their magnitudes are lower than that of the signal. However, since derivatives are noise sensitive, too high weight values will affect the performance of the algorithm. Thus, when choosing w_1 and w_2, we should take into account both the Signal-to-Noise ratio and the difference of magnitudes between the signal and its derivatives. In this study, we choose $w_0 = 1$, $w_1 = w_2 = 2$. The complexity of the algorithm is $O(nm)$.

4.6 Hidden-Markov Models (HMMs)

HMM was first introduced by Baker [1] and has become a fundamental technology underlying most of speech recognition systems nowadays, superseding the DTW based techniques. It is also the method exclusively used in previous works on lipreading [7, 11]. To apply this method to our problem, let us imagine that the facial dynamic patterns are produced by a stochastic process and can be modeled by an HMM. Thus, one HMM of the *'Puppy'* utterance is formed for each subject and stored in a database. When the recognition system receives the facial dynamic pattern of an unknown person, an observation probability is calculated for each of the HMMs in the database to determine the likelihood that the received pattern has been generated by this HMM. The greatest observation probability determines the closest HMM model.

In this work, we build one left-to-right HMM for each subject, and let the number of states equal the number of phonemes $Q = 2$ as suggested by Rabiner [14]. The first *'Puppy'* utterance of the recording is used for model training in similar fashion to the work of Luettin *et al.* [11]. The choices of model parameters (e.g. the number of states

and Gaussian mixtures GM) usually depend on the nature of the training data, and stem from experimentation. We have observed that for the 3D dynamic data used in this study, the system performs well with $Q = 2$ and $GM = 2$. The performance of the recognition process remains unchanged as we use greater numbers of states and/or Gaussian Mixtures.

5 Results and Discussions

Figure 3(a) shows the dynamics of the first three modes of variation (MoV) of a *'Puppy'* utterance. Tests conducted on 36 participants show that 90% of the shape variations are contained in the first 12 MoV. The intra-subject variations can be observed in Figure 3(b) where utterances of the same subject have been recorded over several months to account for different physical and emotional conditions. Although the performance is not perfectly reproducible, the dynamics exhibit familiar patterns, e.g. same trends of the onsets and offsets, similar durations and signal magnitudes, comparable proportions of the patterns *'Pup'* and *'py'* within the overall signal. In contrast, facial dynamics captured from different persons display highly distinctive patterns as depicted in Figure 3(c).

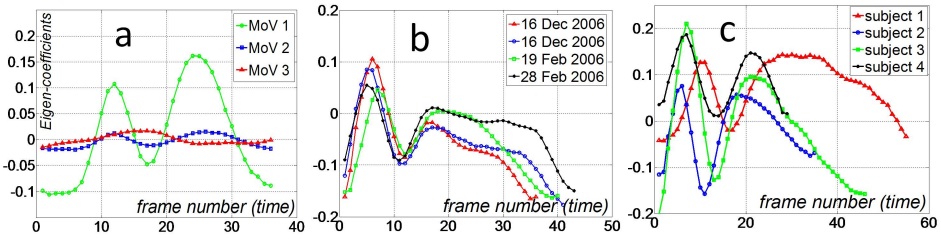

Figure 3: a) Three first modes of variation of a *'Puppy'* utterance b) Dynamics of the same subject over several months, 1^{st} MoV. c) Dynamics of different subjects, 1^{st} MoV.

5.1 Performance assessment of similarity measures

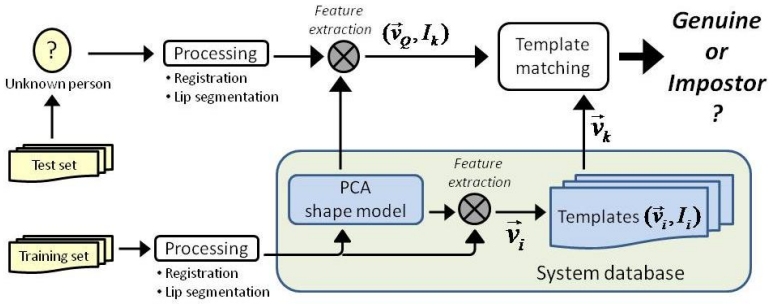

Figure 4: Architecture of the recognition system. Given an biometric feature vector v_Q of an unknown person and a claimed identity I_k, determine if the person is a genuine user or an impostor. Typically, v_Q is matched against v_k, the biometric template of I_k.

A face recognition prototype using facial dynamics is built is depicted in Figure 4, and the ROC curves are shown in Figure 5. In this experiment, the recognition is based on the first mode of variation only. We compute the False-Reject-Rate (FRR) and the False-Accept-Rate (FAR) accordingly to the definitions given in [13]. The Fréchet distance, Correlation Coefficients and DDTW perform very poorly in accordance with our prediction, as explained in section 4. DTW, CDTW and WDTW show very comparable performances, with a slight advantage for WDTW. CDTW does not noticeably improve the system performance because our fast video camera already provides a fine sampling rate. HMM does not exhibit the performance we expected (i.e. better than DTW), however the result obtained in our experiment is comparable to that of Luettin *et al.* [11], i.e. a verification rate of about 75% when using shape variations only.

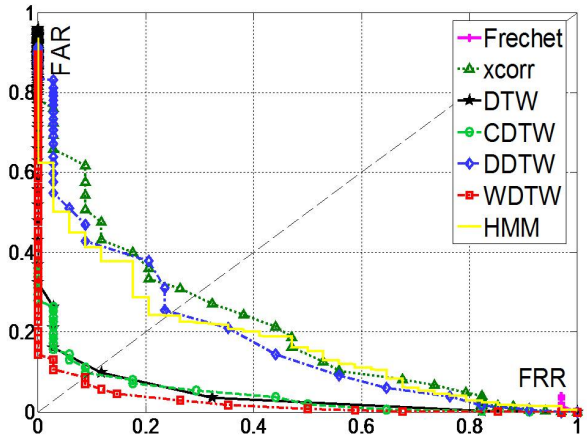

Figure 5: ROC curves of the identification system using several similarity measures.

5.2 Number of modes of variations used

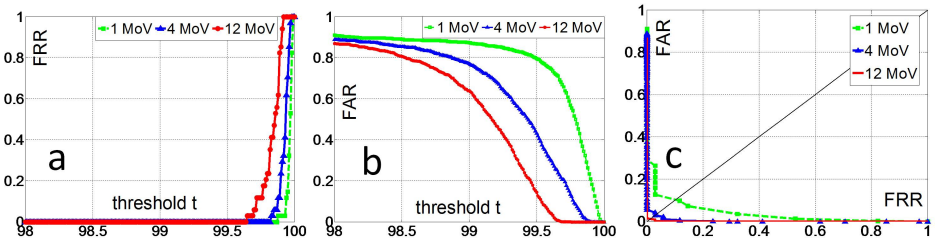

Figure 6: a) FRR and b) FAR for different numbers of MoV used c) The performance increases with the number of MoV used. Best score: EER=1.6% (12 MoV, WDTW).

Intuitively, the more modes of variation we take into account, the more discriminative the biometric feature becomes, thus the FAR decreases. However, since the higher-order modes of variations are also plagued by the intra-subject variations, the FRR increases. Overall, the performance of the system improves as seen in Figure 6(c).

5.3 Influence of unwanted facial actions

In Section 3, we predicted that - due to unwanted facial actions such as blinking - analysing the facial dynamics of the entire face (holistic) might degrade the performance of the recognition system. To verify this hypothesis, we compare the holistic approach to the analysis of the lip motions only. The results are shown in Figure 7. We observe that the FRR is higher when using the face than when using lips only because the system may issue a false reject when attempting to match the dynamics of the same person with and without blinking, for example. However, the lip motions alone are obviously less distinctive than the entire facial dynamics, thus the FAR becomes worse when we ignore part of the face. Overall, the performance improves when using the lip motions alone.

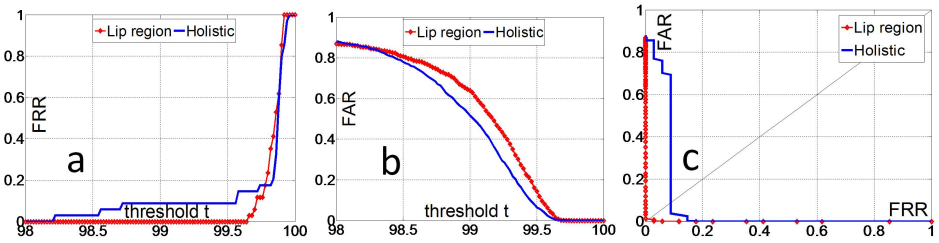

Figure 7: a) FRR and FAR of lip motions vs. holistic analyses b) ROC curves of lip motions vs. holistic analyses. Results computed with the WDTW algorithm.

5.4 Comparison 2D versus 3D dynamic data

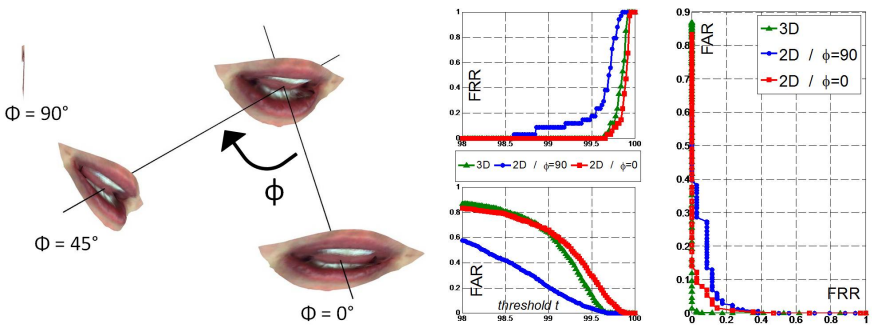

Figure 8: Comparison of the face recognition performances when using 3D dynamic data compared to using 2D dynamic data, for several viewpoints with and without occlusion.

Figure 8 shows the results of a comparative evaluation using 3D dynamic data versus 2D dynamic data for 2 viewpoints: $\Phi = 0$ is the 2D frontal view of the head where there is no occlusion, and $\Phi = 90$ corresponds to the profile view where occlusion occurs. Comparison of the ROC curves shows that 3D recognition yields better result than any 2D situation, which can be explained by multiple reasons. First, we can observe that the feature extraction techniques are not exactly comparable since all vertices are used to train the 3D shape model (≈ 1000 vertices for a lip shape), while the variations in 2D are

captured by building an Active Shape Model [4] using only tens of landmarks per frame. The precision of the landmark placement also plays a role. Secondly, there is a significant loss of information, not only due to the missing dimension but also due to occlusions. While FRR values are comparable between 3D and the frontal view, the profile view presents worse FRR because there is too few information in order to recognise the subject. However, FAR appears to be much better for the profile view than for the other cases. We are not sure how we can interpret this observation.

6 Conclusion and Future Work

This paper reports the results of a feasibility study which aims to establish whether facial gestures can be used as a behavioural biometric. Our first experiments, carried out on a database of 36 participants, indicate that the dynamics of even *very short* facial actions contain sufficient information for identity recognition. It has been also observed that there exists a hierarchy in the reproducibility and distinctiveness of facial gestures. For example, while short spoken words such as the word *'Puppy'* allow to accurately identify the speaker, non-verbal expressions such as the FACS-based facial actions [6] appear non-functional for use in biometric identification.

In order to quantitatively assess the performance of facial dynamics as a behavioural biometric, we investigate a number of suitable feature extraction techniques and pattern matching algorithms, including the Fréchet distance, Correlation Coefficients, Hidden-Markov Models, Dynamic Time Warping (DTW) and its derived forms, and build a prototype of an identification system using facial dynamics. The best performance is obtained using our algorithm WDTW, achieving an Equal Error Rate EER=1.6%. To our surprise, the HMM-based recognition did not yield the high performance we expected. One possible explanation to this is that we train the HMM models with very short data sequences, and DTW-based algorithms seem to perform much better in such a configuration.

Our preliminary results provide supportive evidences that facial gesture is a suitable biometric feature. However, we still need to conduct tests on a much larger database on the one hand, and carry out further experimentations on the other hand in order to assess the strength and weakness of using facial dynamics for person identification. Thus, we wish in particular to identify and compare a more complete set of spoken words and study the impact of spoofing. Besides, we are also examining if feature extraction methods such as Independent Component Analysis and Linear Discriminant Analysis can improve the performance of our system. Finally, we aim to combine facial gesture with voice recognition and passwording, and examine whether such a multi-modal identification system can cope with the stringent requirements for a high security application.

Acknowledgement: *The authors wish to thank Pr S. Richmond and Mr H. Popat of the School of Dentistry, Cardiff University for sharing their face database.*

References

[1] J.K. Baker. *Stochastic modeling for automatic speech understanding.* in Speech Recognition, R. Reddy, ed, New York: Academic Press, 1975.

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

[3] K. Chang, K. Bowyer, and P. Flynn. *Multiple Nose Region Matching for 3D Face Recognition under Varying Facial Expression.* IEEE Trans. on Pattern Analysis and Machine Intelligence 28(10):1695-1700, 2006.

[4] T. Cootes, G. Edwards, and C. J. Taylor. *Active appearance models.* IEEE Trans. Pattern Analysis and Machine Intelligence, 23(6):681685, 2001.

[5] A. Efrat, Q. Fan, and S. Venkatasubramanian. *Curve Matching, Time Warping, and Light Fields: New Algorithms for Computing Similarity between Curves.* Journal of Mathematical Imaging and Vision., 2007.

[6] P Ekman and W. Friesen. *The Facial Action Coding System (FACS): A Technique for the Measurement of Facial Action.* Consulting Psychologists., 1978.

[7] M. I. Faraj and J. Bign. *Audio-visual person authentication using lip-motion from orientation maps.* Pattern Recognition Letters 28(11), pp. 1368-1382, 2007.

[8] A. Jain, A. Ross, and S. Prabhakar. *An Introduction to Biometric Recognition*. IEEE Transactions on circuits and systems for video technology, Vol. 14, No. 1, 2004.

[9] D Johnston, D. Millett, and A. Ayoub. *Are Facial Expressions Reproducible?* Cleft Palate-Craniofacial Journal., 2003.

[10] E.J. Keogh and M.J. Pazzani. *Derivative dynamic time warping.* First SIAM International Conference on Data Mining, 2001.

[11] J. Luettin, N.A. Thacher, and S.W. Beet. *Speaker identification by lipreading*. International Conference on Spoken Language Proceedings, pp. 62-64, 1996.

[12] M.E. Munich and P. Perona. *Continuous Dynamic Time Warping for Translation-Invariant Curve Alignment with Applications to Signature Verification.* The Proceedings of the Seventh IEEE International Conference on Computer Vision, 1999.

[13] P. J. Phillips, A. Martin, C. L. Wilson, and M. Przybocki. *An Introduction to Evaluating Biometric Systems.* IEEE Computer (Special issue on biometrics), pp. 56-63, 2000.

[14] L. R. Rabiner. *A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition.* Proc. of the IEEE, 77 (2), pp. 257286, 1989.

[15] H. Sakoe and S. Chiba. *Dynamic programming algorithm optimization for spoken word recognition.* IEEE Trans. Acoustics, Speech, and Signal Processing, 1978.

[16] M. Turk and A. Pentland. *Eigenfaces for Recognition*. Cognitive Neuroscience, vol. 3, no. 1, pp. 71-86, 1991.