3D Facial Actions for Biometric Applications

Lanthao Benedikt[1], D. Cosker[2],
[1]School of Computer Science,
Cardiff University,
5 The Parade, Roath,
Cardiff,
CF24 3AA,
United Kingdom

Paul L. Rosin[1] and David Marshall[1]
[2]School of Computer Science,
University of Bath,
Bath
BA2 7AY,
United Kingdom

1 Introduction

The face is the primary means for humans to recognize each other in everyday life and therefore represents the most natural choice of a biometric. Although the human face is commonly used as a physiological biometric, very little work has been done to exploit the idiosyncrasies of facial motions for person identification by computer. Early works such as Eigenfaces [1] and methods based on Support Vector Machines [2] have inspired a number of 2D face recognition solutions which are now employed in various commercial and forensic applications (e.g. Neven Vision, Viisage, Cognitec) [34]. Despite such wide acceptance, the face is commonly considered a weak biometric because it can display many different appearances due to rigid motions (e.g. head-poses), non-rigid deformations (e.g. face expressions), and also because of its sensitivity to illumination, aging effect and artifices such as make-up and facial hair [3].

With recent advances in 3D imaging technologies has made 3D facial imaging both static and (more recently) dynamic 3D video. In terms of facial analysis 3D data simplifies some processes — such as head-pose and illumination problems [4]. However, dealing with facial expression variability still remains an open issue. While Chang et al. [5] suggest to use expression-invariant facial regions for identity recognition (e.g. the region around the nose), another solution is to employ statistical model-based algorithms such as the 2D Active Appearance Model [6] (suitably extended to 3D) and the 3D Morphable Model [7] which can accurately interpret unseen facial poses thanks to a prior learning process [20].

More recent research trends have started to consider facial expressions as an additional source of information rather than a problem to overcome. Pioneering works have reported very promising results in the contexts of speaker

identification by lip-reading [8], and speech/facial features[9] and lip motion[10]. However little emphasis has been placed on examining in depth the characteristics of facial dynamics and their suitability for biometric applications, several questions remain open:

- Are facial actions stable over long time intervals?
- How much do they vary with the subject's emotional and physical conditions?
- Are they sufficiently discriminative across a large population?
- Can we arbitrarily employ any facial expressions for identity recognition or is there a hierarchy in their biometric power?

We have begun to study such questions: Previous studies [26, 27, 28] have investigated the *uniqueness* and *permanence* of facial actions to determine whether these can be used as a behavioral biometric. Experiments were carried out using 3D video data of participants performing a set of very short verbal and non-verbal facial actions. The data has been collected over long time intervals to assess the variability of the subjects' emotional and physical conditions. Quantitative evaluations are performed for both the identification and the verification problems, the results indicate that emotional expressions (e.g. smile and disgust) are not sufficiently reliable for identity recognition in real-life situations, whereas speech-related facial movements show promising potential. This chapter presents an overview of this research, readers should consult the papers [26, 27, 28] for complete details.

The chapter is organized as follows. Section 2 reviews the existing public databases and explains our choices for a new set of very short verbal and non-verbal facial actions; Section 3 describes in details the data recording conditions and apparatus. Our experiments employ exclusively naïve users and seek to model as closely as possible a real-life scenario; Section 4 outlines the architecture of an identity recognition system using facial actions, the data processing tasks and an accurate feature extraction method are described, then an efficient pattern recognition algorithm derived from Dynamic Time Warping [23] is proposed; Section 5 assesses the use of facial actions in biometrics considering both the face identification and the face verification problems, in light of which improvements are proposed.

2 What Facial Actions Make a Suitable Biometric

Is it possible that any facial actions can be considered for person recognition? In principle this may be the case. However, choosing those which exhibit high biometric discriminative power is analagous to choosing strong computer login passwords over weak ones. It is also preferable to use *very short* facial actions

to reduce the processing effort so that genuine users can gain access quickly to the secure services.

The majority of related works have been carried out by researchers working on speech analysis [8, 9, 10], who usually employ popular face databases for lipreading and speech synthesis such as the M2VTS [11] and the XM2VTSDB [12] databases which include audio-video data of continuously uttered digits from '0' to '9', and spoken phrases such as *'Joe took father's green shoe bench out'*. Also commonly used is the DAVID database [13] where 9 participants wearing blue lipstick utter isolated digits. These databases have been designed for speech research purposes and are not suitable for biometrics. In fact, the use of physical markers is inconvenient for a real-life application such as biometrics; and furthermore, although long phrases are necessary for speech analysis, they require intensive processing effort, especially when 3D dynamic data is employed. Although one might consider using only short segments of the long phrases, such 'cut out' syllables are inevitably plagued by co-articulation effects, which unnecessarily complicates the problem at this early stage. For these reasons, we decided to collect a new set of *very short* and *isolated* facial actions.

2.1 Verbal Facial Actions

With regard to speech-related facial motions, one might be tempted to see a direct link with research on speech analysis where the distinctiveness of lip motions has been studied to a certain extent [32][37]. However, lipreading and biometrics are two different problems. While lipreading aims to recognize phonemes from lip shapes (visemes) and must be speaker-independent, we seek on the contrary to recognize the visemic dissimilarities across speakers uttering the same phoneme. Another field where there has also been a great deal of interest in characterizing facial motions is Orthodontics and Craniofacial research. However, these works often use a very small number of subjects (between 5 and 30) and examine limited speech postures [17], which is inconclusive for biometrics.

Our objective here is to assess: (1) the repeatability of viseme production over time for any speaker, and (2) the distinctiveness of lip motions across speakers. To this end, we will examine a set of words chosen among the visemes of the English language as depicted in Table 1:

Consonants — The *point of articulation* plays a great role in the strength of a consonant. While bilabial consonants (involving the lips e.g. /p/, /b/, /m/) and labiodental consonants (upper teeth and lower lip e.g. /f/, /v/) are informative because their visible variations can be easily captured by the camera, consonants involving internal speech organs such as the palate and the tongue (e.g. /t/, /k/, /r/, etc) are expected to be poor because their variations are hidden.

Vowels — Vowels typically form the nuclei of syllables, while consonants form the onsets (precede the nuclei) and the coda (follow the nuclei). A codaless syllable of the form V, CV, CCV - where V stands for vowel and C

Viseme	Phoneme	Viseme	Phoneme
/p/	P	**/k/**	K
	B		G
	M		N
	EM		L
/f/	F		NX
	V		HH
/t/	T		Y
	D		EL
	S		EN
	Z	**/iy/**	IY
	TH		IH
	DH	**/aa/**	AA
	DX	**/ah/**	AH
/w/	W		AX
	WH		AY
	R	**/er/**	ER
/ch/	CH	**/ao/**	AO
	JH		OY
	SH		IX
	ZH		OW
/ey/	EH	**/uh/**	UH
	EY		UW
	AE	**/sp/**	SIL
	AW		SP

Table 1: English language Phoneme to Viseme Mapping [36].

for consonant - is called an open syllable, and a syllable that has a coda e.g. VC, CVC, CVCC is called a closed syllable. The vowel of an open syllable is typically long, whereas that of a closed syllable is usually short. Two adjacent vowels in a syllable of the form CVVC usually behave as a long vowel. The r-controlled (vowel followed by 'r') and the l-e-controlled (consonant-l-e at end of syllable) are neither long nor short, but allow trailing as in a long vowel [35].

Our comparative evaluation has been performed on different types of syllables, using the following words: ***'puppy'*** (CVC/CV, short vowel/long vowel), ***'baby'*** (CV/CV, long vowel/long vowel), ***'mushroom'*** (CVCC/CVVC, short vowel/long vowel), ***'password'*** (CVCC/CVC, irregular case: long vowel/r-controlled vowel), ***'ice cream'*** (VCV/CCVVC, long vowel/long vowel), ***'bubble'*** (CVC/CV, short vowel/l-e-controlled), ***'cardiff'*** (CVC/CVC, r-controlled vowel/short vowel), ***'bob'*** (CVC, short vowel), and ***rope*** (CVC, irregular case: long vowel).

2.2 Non-verbal Facial Actions

Unlike visual speech which involves mainly the lip motions, emotional expressions also involve movements of the forehead, the eye-brows, the eyes, the nose, *etc.* Thus, in order to accurately capture the idiosyncrasies of a person's facial dynamics, the entire face needs to be analyzed. However, such a holistic approach may be computationally prohibitive. Therefore it is useful to determine which facial regions are the most important to convey identity, so that we can analyze only the most informative part of the face. This question has already been investigated in perceptual studies, but these works were carried out on 2D still images rather than on 3D dynamic data, and there have been so far no convergent conclusions. While Brunelli&Poggio [14] suggested that the mouth is the most discriminative facial feature across individuals, more recent studies indicate that it is rather the eye-brows, followed by the eyes, the mouth, and the nose [15].

In our study, we examined a set of facial actions involving muscles in various facial regions. On the one hand we study the basic emotions (e.g. smile, disgust); and on the other hand we investigate a number of facial action units (AU) based on the Facial Action Coding System (FACS) [16], e.g. brow raiser (AU1+2), brow lowerer (AU4), upper lid raiser (AU5), nose wrinkler (AU9), and lip corner puller (AU12). Interestingly, research in Orthodontics and Craniofacial has found evidence that facial actions involving maximal muscle stretches are more reproducible compared to moderate expressions [17]. Therefore, we focus on facial expressions of maximum intensity only.

3 Data Acquisition

To collect data of facial actions for our research, two 3D video cameras operating at 48 frames per second are employed (see Fig. 1). Although the cameras have been purchased from the same provider (3dMDFace^{TM} Dynamic System), they are of different generations, hence the 3D mesh densities and the texture qualities they produce are noticeably different. The recording sessions are carried out in two laboratories with different ambient conditions (day light or dark room), and there is no strict control of the head position; small head movements are allowed as long as we ensure an ear-to-ear coverage of the face. No physical markers are used. Such settings resemble a real-life scenario where faces are also expected to be collected from different data capture devices in different recording conditions. This way, we can evaluate how well our recognition algorithm performs on data of variable qualities which has been collected in moderately constrained environments (see Fig. 2).

The participants are staff and students at the Schools of Computer Science, Engineering and Dentistry at Cardiff University, UK. All are naïve users who have not been trained beforehand; in particular, none are familiar with the FACS Coding System [16]. The participants are asked to perform a number of basic emotions (smile, disgust expression) and a small set of Action Units as

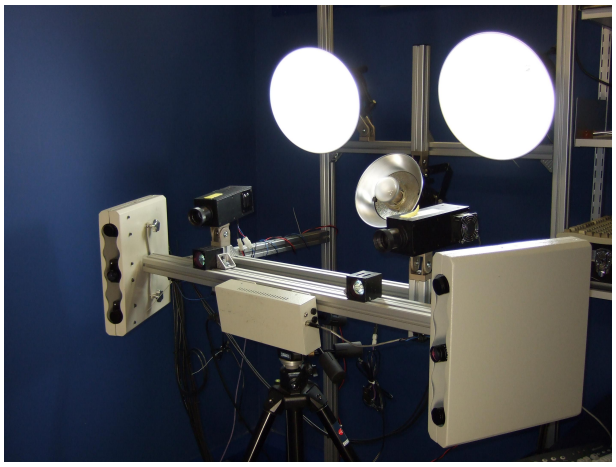


Figure 1: The 3dMDFace^{TM} Dynamic System.

described in Section 2.2, targeting a maximal muscle stretch; on the other hand, they are also required to utter a set of *isolated words* as described in Section 2.1, speaking in a normal and relaxed way.

Figure 2: Stills from two video sequences of the same subject uttering the word 'puppy', collected in two recording sessions with different settings to test how well the recognition algorithm performs on data of variable qualities.

The recording sessions are scheduled over large time intervals (over a few weeks or months and in some cases, over more than two years) to assess the variability of the subjects' emotional and physical conditions. Due to difficulties to collect 3D dynamic data on a large scale, our database is unfortunately nonuniform at the moment. We have been able to collect data from 94 participants (61 males and 33 females, 70 are natural English speakers) uttering the word 'puppy' over more than two years in some cases. However, only 15 participants were used to study other utterances (e.g. as 'password', 'mushroom', etc), and the repetitions have been recorded over a period of one month maximum. The smile dynamic has been studied for about 50 subjects, and smaller tests are conducted on FACS AUs. In reality, we very quickly abandoned the investigation of AUs because during the recording sessions, it became clear that

it was very difficult for naïve users to produce accurate AUs, let alone to repeat exactly the same performance several times.

4 Facial Identity Recognition

Many techniques have been proposed in previous studies for extracting facial dynamic features. Of the approaches which do not require the use of physical markers, there are for example the work of Pamudurthy et al. [19] which aims to track the motions of skin pores, and the work of Faraj et al. [10] which uses the velocity of lip motions for speaker recognition. These methods are interesting and novel, nevertheless they still require more research. In this study, we adopt another feature extraction approach relying on the well-established model-based algorithms used in static face recognition e.g. the Active Appearance Model [6] and the Morphable Model [7]. Readers are referred the papers [26, 27, 28] for complete details on the methods described here .The data pre-processing and feature extraction steps are now briefly outlined.

4.1 Face Normalization

One non-trivial preliminary task in face recognition is to normalize the scans so that they can be compared in the same coordinate frame (i.e. similar head pose and size). The normalization is achieved using a nose-matching technique inspired from a method proposed by Chang et al. [5], the result is illustrated in Fig. 3. The nose root is found using 3D curvature analysis, then the surface normal at the nose root and the eye-direction are computed. These quantities determine the adequate translations, rotations and scaling to align the faces. Further refinements are achieved using the Iterative Closest Point matching algorithm [38], applied to the most stable regions around the nose root.

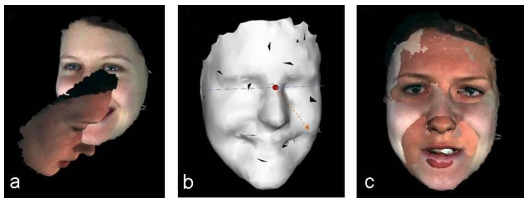


Figure 3: Rigid object 3D registration: (a) Two raw face scans presenting different head poses and sizes need to be aligned; (b) the nose root is located, together with the eye directions and surface normals; (c) the alignment is achieved through rotations, translations and scaling, then refined by applying the Iterative Closest Point algorithm to the nose region.

In practice, we use one video frame as the reference frame, and register all frames in the database to the latter. The choice of the reference frame is arbitrary, as long as we employ one and only one reference for the entire database.

4.2 Frame Correspondence and Segmentation

One principal requirement in our feature extraction approach is that all face scans must be in full correspondence, i.e. they must have the same number of vertices and identical vertex topology. However, our 3D data capture technology treats each scan as an independent problem, resulting in all meshes being uncorrelated. Therefore, one additional processing step is needed in order to bring all frames into correspondence, which can be achieved as follows.

The idea consists of choosing one face to be *the* reference frame for the entire face database, and use a thin-plate-spline warping process to deform this one to each of the frames in the database. This produces a new database where all faces have the same vertex topology as the reference frame, but the facial expression of each frame is identical to that of the corresponding frame in the original database.

The purpose of the warping process is two-fold. First, it brings all the frames into correspondence; and second, it also permits the segmentation of the region of interest (ROI). For example, by using a lip shape as a reference template and deforming this one to the lip region of the target frames, we can produce a sequence of lip shapes in full correspondence as shown in Fig. 4.

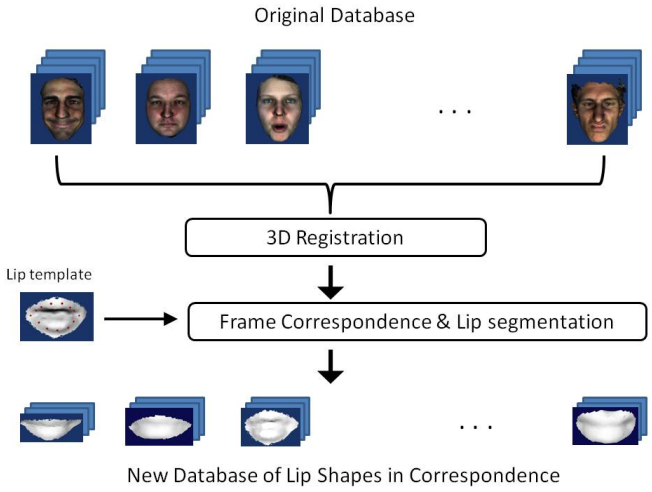


Figure 4: The warping process also permits the segmentation of the Region of Interest when using different facial templates. Here the lip segmentation is realized by using a lip template.

4.3 Facial Dynamics: Representation and Extraction

After the above pre-processing steps, the face database is now divided into training set (gallery) and test set (probe). The gallery includes only one video sequence per subject due to limited data availability, the probe contains several

repetitions for each subject. The gallery and probe are entirely distinct data segments, the probe includes subjects enrolled in the gallery, as well as 'unseen' persons. In this study, we typically build one model for each facial expression, which includes all subjects enrolled in the gallery.

Our statistical model of combined 3D shape and texture is built in similar fashion to the 2D Active Appearance Model [6], but the shape model includes the xyz coordinates of all vertices as in the Morphable Model [7]. Fig. 5 depicts the temporal variations of the first Eigencoefficient and the corresponding face expressions of a subject performing the FACS Action Unit 9 (nose wrinkler).

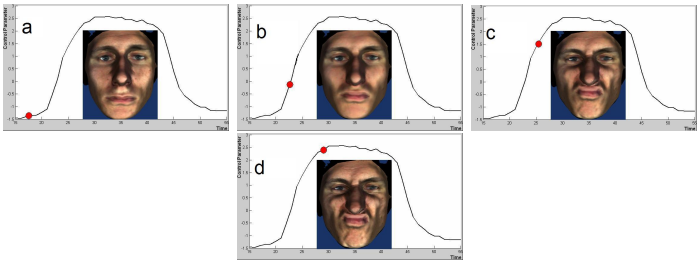


Figure 5: A subject performing a FACS Action Unit 9 (nose wrinkler) and corresponding temporal evolution of the first Eigencoefficient.

The facial dynamics can be extracted from the variations of the 3D shape alone, independently of the texture information. Therefore, for the remainder of this study, we build a simple 3D Active Shape Model (3D ASM) instead of the complete Active Appearance Model. For a more efficient face recognition (i.e. dimensionality reduction of the feature space and noise removal), we retain only 90% of the shape variations. Fig. 6 shows the lip dynamics of four subjects saying 'puppy'. The temporal variations of the first three Eigencoefficients $(v_k^1, v_k^2, v_k^3, k \in \{1, ..., N\})$ are depicted.

4.4 Matching Facial Dynamics

Well-established algorithms such as Hidden Markov Models (HMM) and Gaussian Mixture Models are the preferred pattern matching techniques employed in the previous works[8][10]. However, it is unclear whether these statistical approaches would perform efficiently in the present context where we use only very short data sequences, and do not have many repetitions per subject to train an accurate HMM.

For this reason, in a related work [26], we survey a number of pattern matching techniques commonly used in behavioral biometrics (e.g. voice, signature authentication), and evaluate which of these performs the best when applied to facial dynamic matching. Well known methods such as the Fréchet distance [21], Correlation Coefficients [22], Dynamic Time Warping (DTW) [23], Continuous DTW (CDTW) [24], Derivative DTW (DDTW) [25], and HMM [8] have been

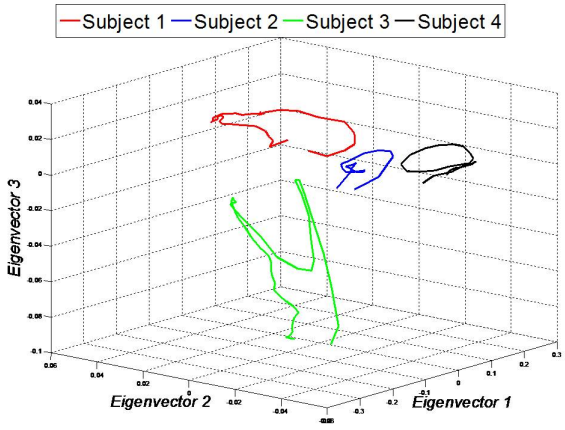
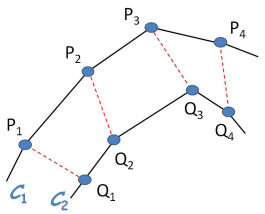


Figure 6: 'Puppy' dynamics of several subjects in the feature space.

examined, in light of which we have proposed an improved algorithm Weighted Derivative DTW (WDTW) [26] as follows.

Suppose we have two feature vectors which are respectively depicted by two sequences of discrete data points C_1 and C_2 as shown in Fig. 7. For each point P_i of C_1, we compute its closest point Q_j on C_2 with respect to the distance $d^i = w_0 \times d_0^i + w_1 \times d_1^i + w_2 \times d_2^i$, where d_0^i is the Euclidean distance between P_i and Q_j, d_1^i is the difference between the local first derivatives, and d_2^i the difference between the local second derivatives; w_0, w_1 and w_2 are weights. The distance D_{WDTW} between C_1 and C_2 is defined as the sum of all pair-wise distances d^i, normalized by the sequence length N. Thus, the similarity between two feature vectors can be computed as $S = 1/D_{WDTW}$.



$$d^i = w_0 \times d_0^i + w_1 \times d_1^i + w_2 \times d_2^i$$

$$D_{WDTW} = \frac{1}{N} \sum_{i=1}^{N} d^i$$

$$S = \frac{1}{D_{WDTW}}$$

Figure 7: Weighted Derivative DTW algorithm (WDTW).

The overall facial dynamics matching system is summarise in Fig 8

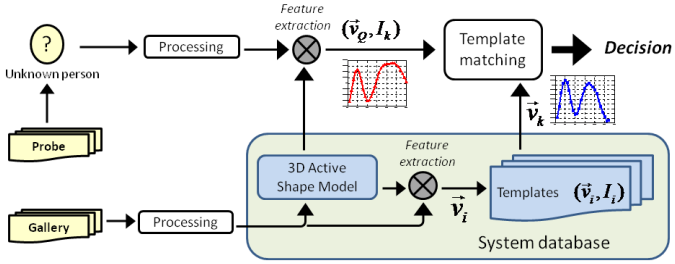


Figure 8: Architecture of the face dynamics matching prototype. Given an biometric feature vector v_Q of an unknown person and a claimed identity I_k, determine if the person is a genuine user or an impostor. Typically, v_Q is matched against v_k, the biometric template of I_k.

5 Results

We have investigated the use of facial actions for person recognition utilising WDTW which performs the best in the present context [26]. A summary of the comparative evaluation of pattern matching algorithms has been presented in [26, 27].

5.1 The Identification Problem

The block diagram of the face verification system as seen in Fig. 8 is still valid, with the difference that there is no claimed identity I_k and the probe feature vector $\overrightarrow{\mathbf{v}}_Q$ is matched against all templates $\overrightarrow{\mathbf{v}}_i$ in the database.

The system can operate in two modes. *Watch list:* are you in my database? and *basic identification:* you are in my database, can I find you? In both cases, the identification utilises a *similarity* measure, S, which can be marked as true if it is above some threshold, t. In our experiment we use the similarity between two feature vectors $S = 1/D_{WDTW}$. Table 2 shows a subset of the matching results: the lines correspond to the probes, which are matched against the gallery templates listed in the columns. The system is capable to recognize the correct user since the similarity between utterances from the same subject is always greater than that between different subjects. In practice, the correct answer does not always correspond to the best match, therefore the performance of the system is measured by plotting the Cumulative Match Curve (CMC) [29] as shown in Fig. 9. The percentage of correct identifications found in the m best matches is calculated for several values of the rank m.

When the system is used in the *Watch list* mode, a threshold needs to be determined such that users who are unknown to the system are rejected (e.g. Laura, Melanie in Table 2). For example, threshold $t = 2.00$ can be a suitable choice.

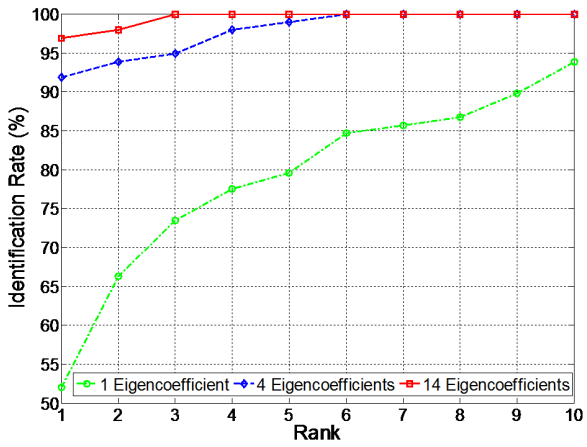


Figure 9: Cumulative Match Curve of the identification System: percentage of correct answers found in the top m best match, for several values of rank m.

Table 2: Similarity between data sequences $S = 1/D_{WDTW}$. Columns: reference templates. Lines: Unknown users.

	Abi	Avril	Bahvna	Chris	Emily	George	Hash	Jamie	James	Kim	Luke	Matt	Vedran	Vitaly
Abi	**15.78**	4.15	5.19	4.78	3.21	1.95	1.43	1.82	2.58	1.36	1.95	2.21	3.17	3.26
Avril	5.18	**14.05**	3.89	7.06	3.96	2.29	1.67	1.74	2.66	2.04	2.38	2.34	3.34	2.96
Bahvna	4.61	2.64	**12.87**	3.43	3.28	1.44	1.97	1.40	2.43	1.09	1.39	1.44	5.60	2.04
Chris	5.22	5.22	3.99	**13.86**	4.13	2.37	1.97	1.78	2.59	2.15	2.31	2.92	6.74	3.90
Emily	2.72	2.43	2.14	2.73	**8.87**	1.22	0.82	0.90	1.36	1.13	1.15	0.95	2.86	1.45
George	1.22	1.34	1.15	1.54	1.08	**5.76**	2.23	1.27	1.04	1.39	1.14	3.34	1.35	2.38
Hash	1.61	1.50	2.38	1.43	0.99	2.12	**22.32**	1.79	2.01	0.95	1.69	3.23	2.46	2.10
Jamie	1.57	1.37	1.32	1.62	0.98	1.28	1.72	**7.49**	1.29	0.85	1.55	2.75	1.56	1.51
James	2.17	2.32	2.21	1.77	1.64	1.22	1.53	1.42	**14.15**	0.90	2.23	2.09	2.18	1.42
Kim	1.00	1.26	0.97	1.72	1.18	1.30	1.04	0.83	0.89	**8.24**	0.92	1.20	1.20	1.45
Luke	2.39	2.84	1.85	2.23	1.47	1.50	1.72	1.94	2.46	1.25	**12.06**	3.29	2.20	1.89
Matt	1.78	1.76	1.49	1.62	1.03	2.05	3.36	1.93	2.01	1.03	2.48	**16.26**	1.66	2.23
Vedran	3.70	3.10	4.96	4.34	3.23	2.19	2.93	1.63	2.80	1.42	2.33	2.91	**23.51**	3.10
Vitaly	2.90	2.42	2.26	3.14	2.00	2.52	2.02	1.48	1.61	1.59	1.63	3.26	2.57	**14.63**
Laura	0.74	0.94	0.82	1.19	1.05	0.31	0.40	0.46	0.62	0.50	0.61	0.45	0.94	0.47
Melanie	1.12	1.08	1.00	1.35	1.52	1.65	0.81	0.67	0.83	1.76	0.62	0.94	1.07	1.68

5.2 Permanence and Uniqueness of Facial Actions

The discriminatory power of any biometric feature depends essentially on its permanence (i.e. the intra-subject variations should be small) and its uniqueness (i.e. the inter-subject variations should be significant). Facial dynamic is a behavioral trait, and as such, it is not expected to be strictly reproducible over time, due to the variability of the subject's emotional and physical conditions. Our next experiment aims to measure to which extent facial dynamics can be considered permanent, and to assess whether the inter-subject distinctiveness is sufficiently large compared to the intra-subject fluctuations.

5.2.1 Variations over Long Time Intervals

Six participants, code-named H, J, V, C, G and L have been recorded uttering the word 'puppy' several times. All participants are natural English speakers with the exception of V. Recordings of H and G were taken over more than two years, J were recorded over three months and over one month for C. V was recorded on the same day, with only 15 minutes between the two recording sessions and with three repetitions per session. L is used as an unseen subject.

The first utterance of five subjects H1, J1, V1, C1 and G1 are employed to build a 3D ASM which is then used to estimate the feature parameters of all the utterances. The pair-wise distances of the facial dynamics are computed and compared [28].

Results show that the reproducibility of facial actions is very person-dependent. For example, H's and G's utterances appear more stable over two years compared to V's within the same day. Also, there is a greater similarity between utterances taken in the same session e.g. {H1 and H2}, {H5 and H6}, {J1 and J2}, {J3 and J4}, {V1,V2 and V3}, {V4,V5 and V6}, etc. The distinctiveness across speakers is more significant than the intra-subject variations, thus the clusters are accurately formed. L is an unseen subject who presents a very important dissimilarity to all of the subjects in the gallery. These observations indicate that facial dynamics of even very short facial actions provides sufficient discriminatory power to be used for person recognition.

5.2.2 Different Speech Postures

Fig. 10 shows the dynamics of four different subjects uttering the word 'puppy' several times. A careful examination shows that the closed syllable (short vowel) /pup/ appears more stable compared to the open syllable (long vowel) /-py/. In particular, the onsets of the syllable /pup/ are nearly identical. This observation holds for all the 94 subjects recorded in our database.

As shown in Fig. 11, the repetitions of the word 'password' (CVCC/CVC, long vowel/r-controlled vowel) seem to indicate that closed syllables are repeatable over time. This tendency can also be observed in the dynamics of other utterances. In other expreriments citec27b the closed syllables such as /diff/ in 'cardiff', /bub/ in 'bubble' and /bob/ appear reproducible across the repetitions, while the open syllables (long vowels) such as /-by/ in 'baby', and the

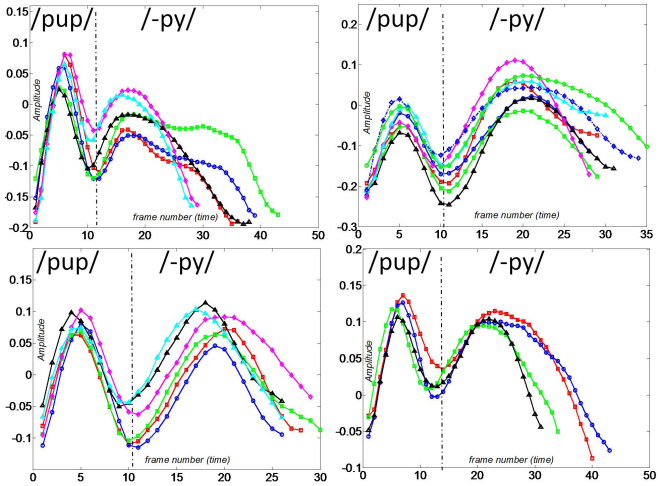


Figure 10: Reproducibility of the 'puppy' utterances from four subjects recorded over several month intervals. Variations of the first Eigencoefficient is shown, which corresponds to the lip opening. The closed syllable /pup/ appears more stable compared to the open syllable /-py/. However, the coda of the /pup/ syllable seems affected by co-articulation.

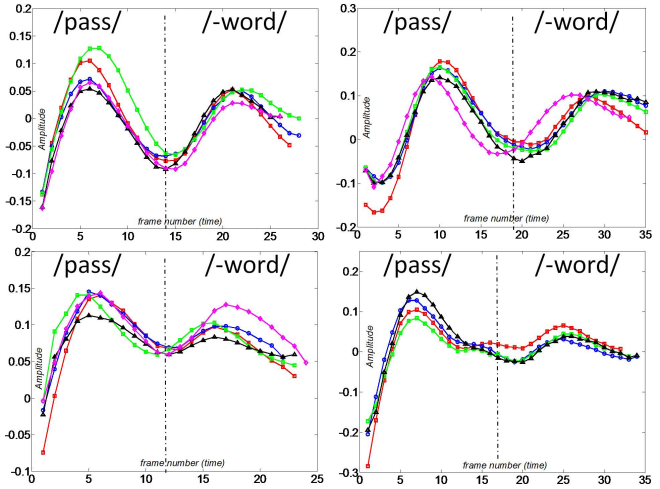


Figure 11: Dynamics of the 'password' utterances of four subjects recorded over several month intervals. Variations of the first Eigencoefficient which corresponds to the lip opening.

two syllables in 'mushroom' and in 'ice cream' present fluctuations in their dynamics, with regard to both the durations and the trends of the trajectories. The r-controlled vowel /car/ in 'cardiff', and the l-e-controlled vowel in 'bubble' also show significant fluctuations.

These observations indicate that there is a hierarchy in the reproducibility of syllables, which means that some types of speech-related facial movements are more suitable for biometric compared to others. However, our experiments are not sufficient to establish any formal rule at the moment. We are collecting data in order to investigate this issue further.

5.3 Non-verbal Facial Actions

Figure 12: Three video sequences of a naïve user performing an AU12-standardized maximal smile. The data was recorded over three months to assess the variability of the subject's emotional and physical conditions. Significant variations and inaccuracies can be observed, e.g. lip opening, eye-brow movements, facial asymmetries. The expression appear very unstable.

Fig. 12 shows three sequences of the same subject performing a maximal smile, the videos were captured over several months, and Fig. 13 shows the temporal variations of the three lead Eigencoefficients corresponding to the smile sequences. Considering the large variety of smile patterns that a person can produce, we wish to standardize the performance and required the subject to perform an AU12 maximal smile (closed lips, no significant facial movements in other facial regions than the lips). Although clear instructions have been given, significant intra-subject variations and inaccuracies can still be noticed across the repetitions, which result in the large fluctuations of the trajectories of the Eigencoefficients.

Examinations of the smile dynamics from 50 subjects show similar intra-subject variations and many occurrences of unwanted and non-repeatable actions such as blinking, eye-brow movements, facial asymmetries. For these reasons, the holistic feature extraction does not seem adequate (see Section **??**). However, the local feature extraction approach also shows very poor recognition result because the lip motions alone are not sufficiently discriminative in the example of the smile. Other experiments have been carried out using the

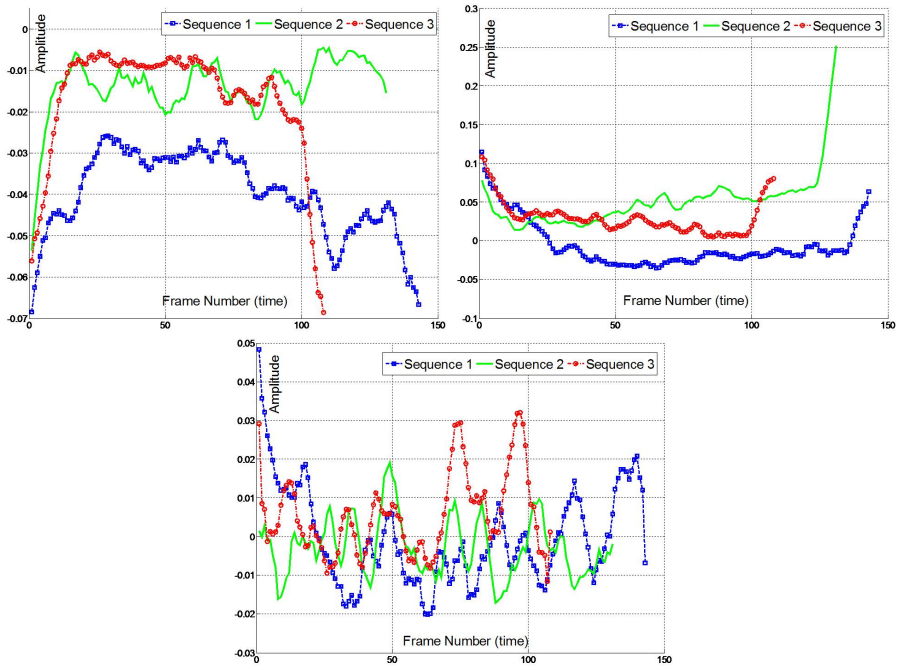


Figure 13: Smile dynamics corresponding to the three sequences shown in Fig. 12. Variations of the three lead Eigencoefficients. In this experiment, we adopt the holistic feature extraction approach in order to capture the idiosyncrasies of the entire face.

disgust expression, where similar problems have been observed. For these reasons, we believe that non-verbal facial expressions are not suitable for biometric applications.

Finally, regarding the use of FACS Action Units for person identification, we have observed during the recording sessions that it is very difficult for naïve users to produce accurate AUs, let alone to repeat identical performances. Fig. 5 shows a subject performing AU9 (nose wrinkler). Perceptual evaluation done by a FACS-trained psychologist has found this sequence to include not only AU9, but also other AUs. This raises the question whether AU-based face expressions are suitable in a real-life scenario where users are not familiar with the FACS coding system.

6 Conclusions and Future Work

We have investigated the feasibility of using facial actions for biometric identification. Although previous work have been carried out on using lip motions for speaker recognition, there was very little emphasis on studying in depth the permanence and distinctiveness of facial motions. Furthermore, the problem has never been evaluated in an environment resembling a real-life scenario. We have reviewed the face databases and the recognition methods used in previous work, in light of which we have proposed several improvements.

First, we explain our choice of a new and meaningful set of *very short* verbal and non-verbal facial actions, which allow comparison of the biometric power of different types of facial movements. The evaluation is carried out on 3D dynamic data which has been recorded over long time intervals to assess the variability of the subjects' emotional and physical conditions. The use of very short facial actions is essential in order to minimize the processing effort, which is very important for a real-time application such as face recognition.

A face recognition system has been implemented using facial dynamics and proposed an accurate pattern matching algorithm derived from Dynamic Time Warping. The results indicate that even *very short* verbal facial actions (e.g. utterance of a two-syllable word) provide sufficient information for person identification. However, while speech-related facial motions show good potential, emotional expressions (e.g. smile and disgust) and the FACS Action Units appear very unstable, especially when naïve users are employed. This indicates that non-verbal expressions may not be suitable for biometrics.

In the particular case of speech-related facial motions, we observe that there exists a hierarchy in the reproducibility of different types of utterances. While short vowels seem very repeatable, even over long time periods, long vowels show significant fluctuations. . If we can establish formal rules to identify which types of facial motions are the most reproducible and discriminative across speakers, this will help the users to select strong passwords and the recognition performance will be improved.

7 Acknowledgements

The authors would like to acknowledge the volunteer participation of all the members of staff and students of the Schools of Dentistry, Engineering and Computer Science at Cardiff University, UK; and in particular, we would like to thank Mr James MacKenzie for his valuable help to collect the face data used in this research.

References

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

[2] P. J. Phillips. "Support vector machines applied to face recognition", *in Proc of the Conf on Advances in Neural Information Processing Systems* , vol. II, pp. 803-809; 1999.

[3] A. K. Jain, A. Ross and S. Pankanti, "Biometrics: A Tool For Information Security", *IEEE Transactions on Information Forensics and Security*, vol. 1, no. 2, pp. 125-143; 2006.

[4] K. W. Bowyer, K. Chang, and P. J. 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. 115; 2006.

[5] 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*, vol. 28, no. 10, pp. 1695-1700; 2006.

[6] G. Edwards, T. Cootes, and C. J. Taylor. "Face Recognition Using Active Appearance Models", *in Proc. of the European Conf on Computer Vision*, vol. 2, pp. 581-595; 1998.

[7] V. Blanz and T. Vetter. "Face recognition based on fitting a 3D morphable model". *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 25, no. 9, pp. 1063-1074; 2003.

[8] J. Luettin, N.A. Thacker, and S.W. Beet. "Speaker identification by lipreading", *in Proc of the Intl Conf on Spoken Language Proceedings*, vol. 1, pp. 62-64; 1996.

[9] M. J. Roach, J. D. Brand and J. S. Mason. "Acoustic and Facial Features for Speaker Recognition", *in Proc of the Intl Conf on Pattern Recognition*, vol III, pp. 258-261; 2000.

[10] M. I. Faraj, and J. Bigun. "Audio-visual person authentication using lip-motion from orientation maps", *Pattern Recognition Letters*, vol. 28, no. 11, pp. 1368-1382; 2007.

[11] S. Pigeon, and L. Vandendrope. "The M2VTS Multimodal Face Database", *in Proc. of the First Intl Conf. Audio and Video-Based Biometric Person Authentication*; 1997.

[12] K. Messer, J. Matas, J. Kittler, J. Luettin, and G. Maitre. "XM2VTSDB: The extended M2VTS database", *in Proc of the Second Intl Conf on Audio and Video-based Biometric Person Authentication*, Springer Verlag, pp. 72-77; 1999.

[13] C. C. Chibelushi, S Gandon, J. S. Mason, F. Deravi, and D Johnston. "Design Issues for a Digital Integrated Audio-Visual Database", *in IEE Colloquium on Integrated Audio-Visual Processing for Recognition, Synthesis and Communication*, pp. 7/1-7/7; 1996.

[14] R. Brunelli, and T. Poggio. "Face recognition: Features versus templates", *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 15, no. 10, pp. 1042-1052; 1993.

[15] P. Sinha, B. Balas, Y. Ostrovsky, and R. Russell. "Face Recognition by Humans: Nineteen Results All Computer Vision Researchers Should Know About", *Proceedings of the IEEE*, vol. 94, Issue 11, pp. 1948-1960; 2006.

[16] P Ekman and W. Friesen. "The Facial Action Coding System: A Technique for the Measurement of Facial Action", *Manual for the Facial Action Coding System. Consulting Psychologists Press.*; 1978.

[17] D. Johnston , D. T. Millett and A. F. Ayoub. "Are Facial Expressions Reproducible?", *Cleft Palate-Craniofacial Journal.* vol. 40, no. 3, pp. 291-296; 2003.

[18] D. Fidaleo and M. Trivedi. "Manifold analysis of facial actions for face recognition", *in Proc of the 2003 ACM SIGMM workshop on Biometrics methods and applications*, pp. 65-69; 2003.

[19] S. Pamudurthy, E. Guan, K. Mueller, and M. Rafailovich. "Dynamic approach for face recognition using digital image skin correlation", *in Audio and Video-Based Biometric Person Authentication. Lecture Notes in Computer Science*, vol. 3546, pp. 1010-1018; 2005.

[20] X. Lu. "3D Face Recognition across Pose and Expression", *PhD thesis, Michigan State University*; 2006.

[21] A. Efrat, Q. Fan and S. Venkatasubramanian. "Curve Matching, Time Warping, and Light Fields: New Algorithms for Computing Similarity between Curves", *in Journal of Mathematical Imaging and Vision*, vol. 27, no. 3, pp. 203-216.; 2007.

[22] N. Dhananjaya. "Correlation-Based Similarity Between Signals for Speaker Verification with Limited Amount of Speech Data", *Multimedia Content Representation, Classification and Security. Lecture Notes in Computer Science*, vol. 4105, pp. 17-25; 2006.

[23] H. Sakoe and S. Chiba. "Dynamic programming algorithm optimization for spoken word recognition", *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. 26, issue 1, pp. 43-49; 1978.

[24] M.E. Munich and P. Perona. "Continuous Dynamic Time Warping for Translation-Invariant Curve Alignment with Applications to Signature Verification", *in Proc. of the Seventh IEEE Intl Conf on Computer Vision*, pp. 108-115; 1999.

[25] E.J. Keogh and M.J. Pazzani. "Derivative dynamic time warping", *in Proc of the First SIAM Intl Conf on Data Mining*; 2001.

[26] L. Benedikt, D. Cosker, P. L. Rosin and D. Marshall. "Facial Dynamics in Biometric Identification", *in Proc of the Bristish Machine Vision Conference (BMVC)* vol. 2, pp. 235-241; 2008.

[27] L. Benedikt, D. Cosker, P. L. Rosin and D. Marshall. "Facial Dynamics in Biometric Identification", *in Proc of the IEEE Second International Conference on Biometrics: Theory, Applications and Systems (BTAS)*, pp. 1-6; 2008.

[28] L. Benedikt, D. Cosker, P. L. Rosin and D. Marshall. "Assessing the Uniqueness and Permanence of Facial Actions for Use in Biometric Applications", *To Appear IEEE Transactions on Systems, Man and Cybernetics - Part A*, 2009.

[29] R. M. Bolle, J. H. Connell, S. Pankanti, and N. K. Ratha. "The Guide to Biometrics", *Springer 2004. ISBN: 0-387-40089-3*; 2004.

[30] T. Hastie and W. Stuetzle. "Principal curves", *Journal of the American Statistical Association*, vol. 84, pp. 502-516; 1989.

[31] 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, pp. 4-20; 2004.

[32] S. A. Lesner, and P. B. Kricos. "Visual Vowels and Diphthongs Perception across speakers". *in Journal of the Academy of Rehabilitative Audiology*, vol. 14, pp. 252-258; 1981.

[33] J. S. Mason, and J. D. Brand. "The Role of Dynamics in Visual Speech Biometrics". *in Proc of the IEEE Intl. Conf on Acoustics, Speech, and Signal Processing*, vol. 4, pp. 4076-4079; 2002.

[34] P. J. Phillips, W. T. Scruggs, A. J. OToole, P. J. Flynn, K. W. Bowyer, C. L. Schott, and M. Sharpe. "Face Recognition Vendor Test (FRVT) 2006". *http://www.frvt.org/FRVT2006*

[35] J. Laver. "Principles of Phonetics". *Cambridge University Press*; 1994.

[36] P. Lucey, T. Martin, and S. Sridharan. "Confusability of Phonemes Grouped According to their Viseme Classes in Noisy Environments". *in Proc of the Australian Int. Conf. on Speech Science and Technology*, pp. 265-270; 2004.

[37] W. M. Weikum. "Visual Language Discrimination". *PhD thesis, The University of British Colombia, Vancouver*, 2008.

[38] P. Besl and N. McKay. "A Method for Registration of 3-D Shapes". *IEEE Trans. Pattern Analysis and Machine Intelligence*, 14:239256, 1992.