(Excerpt from the PhD Thesis entitled 'Using 3D Facial Motion for Biometric Identification', 11/2009)

PhD Thesis - Chapter 2

Facial Motion: a novel biometric?

Author: Lanthao Benedikt

Supervisors: Dr. David Marshall, Dr. Paul Rosin

April 5, 2010

Cardiff University - School of Computer Science

Address: Queen's Buildings, 5 The Parade, Cardiff CF24 3AA, United Kingdom

Website: http://www.cs.cf.ac.uk

Thesis Abstract

In the current context of heightened security, expanded research effort is needed for exploring novel biometric modalities, while continuing to improve existing solutions. The purpose of this research is to investigate whether the 3D spatiotemporal dynamics of facial expressions can be used for person identification, and to compare the advantages of this novel approach to classic face recognition using static mugshots. The working hypotheses are formulated as follows:

> ***Hypothesis 1:*** *Human facial movements are stable over time and sufficiently distinctive across individuals to be used as a biometric identifier. However, there exists a hierarchy in the reproducibility and uniqueness of facial actions, such that some facial expressions are more suitable for person recognition compared to others.*

> ***Hypothesis 2:*** *There exists a motion signature associated with moving faces, which describes not only the behavioural traits of a person, but also reveals physical idiosyncrasies that becomes visible only when the face is in motion. This identity cue is independent from the physical information conveyed by static faces, such that the temporal dynamics of facial expressions contribute to improve identity perception.*

The purpose of the present Chapter is to conduct a literature review on what has been done previously in related fields with regard to the use of facial motions in biometrics, in light of which we examine how further progress can be made.

1 The role of motion in face perception

1.1 Categorising facial signals

The face is by far the most important organ for interpersonal communication in our social life. It is capable of conveying emotions, cognitive states, identity, and furthermore, it also houses an important part of the speech-production apparatus. In an extensive study published in 2005, Pantic suggested that signals produced by the human face can be classified into four categories [1]:

- *Static signals:* permanent features of the face, e.g. the bone structure, the overall proportions of the face, the soft tissue, the skin property which determines the light reflection, and thus, affects the face appearance.

- *Slow signals:* changes which occur gradually over time e.g. wrinkles and loss of the skin elasticity. Considerable growth of the skeletal structures can usually be observed in infants and young adults, along with an increase in both the muscle and fatty tissues. Later, there is little change in the bone structure, although the cartilage often continues to grow, especially in men, affecting in particular the shape of the nose [2]. These aging effects cause the static signals to be non permanent over large periods of time.

- *Rapid signals:* temporary neuromuscular activities which cause either visible changes in the face appearance, or in the head position and orientation with respect to the viewer. These signals include both nonrigid motions (e.g. speech, facial expressions) and rigid motions (e.g. head nodding, shaking). Rapid signals can be voluntary e.g. speech, or involuntary e.g. blinking, blushing.

- *Artificial signals:* artifices e.g. eye-glasses, cosmetics and facial hair.

The static facial signals are traditionally employed for identity recognition. The most popular algorithms include methods that extract facial geometrical information for recognition [3–5], or those which project the raw face images into a low-dimensional subspace and use the projection coefficients for matching, e.g. Eigenfaces [6], Fisherfaces [7], and ICA-based algorithms [8]. There also exists an emerging trend, called skin texture analysis, that exploits the visual details of the skin [9].

Regardless of the methods employed, the slow, rapid and artificial facial signals are traditionally regarded as problems that degrade the recognition performance and need to be overcome [10–13]. For example, extensive effort has been made to develop expression-invariant algorithms that consist of matching facial features around the nose region that do not deform too much with facial expressions [14]. This approach may be challenging since the human face in real-world is a living animated object rather than a static one. Furthermore, the nose region alone may not be sufficiently discriminative when the biometric database is large.

A new research trend has recently emerged, suggesting that the slow signals and the artificial facial signals can contribute to improve identity recognition. For example, slow facial signals can be exploited for age estimation, while artificial signals facilitate gender recognition [1]. As far as the rapid signals are concerned, only a few studies in computer science have investigated their use for identity recognition [13,15–18]. These pioneer works mainly focussed on developing algorithms while ignoring some fundamental questions: is facial motion a viable biometric feature? Is it stable and unique to each individual? Does facial motion convey an additional identity cue that is different to that of a static face? The objective of this chapter is to gather evidence and findings from related research fields (e.g. psychology, neuroscience, articulatory phonetics) in order to gain a better understanding of the usefulness of facial motion for application in biometric identification.

1.2 Three theories from perceptual psychology

Research in cognitive neuropsychology has long established that motion plays an important role in many object recognition tasks [19–21]. However, the relationship between static and dynamic perceptions remains unclear. In the particular case of face recognition, it has been proven that the visual kinetics facilitate face recognition under sub-optimal viewing conditions, e.g. when faces are degraded by negation [22], black and white thresholding [23], and pixelation [24]. If a majority of researchers believe that facial motion improves identity perception, a small number of experimental studies disagree. At least three theories have been proposed:

- There exists *a motion signature per se*: motion reveals the idiosyncratic patterns of faces, these constitute a reliable cue for the identity if repeated regularly [25]. Besides, facial movements are believed to carry rich information about gender [26, 27], age [28], and to some extent, identity [22–24, 27, 29].
- The *structure-from-motion* concept: motion provides multiple viewpoints of the head, and several facial poses. As a result, this facilitates identity perception, but the temporal dynamic itself does not have any significant contribution [30, 31].
- The benefits of motion is null for recognition of unfamiliar faces, which led some researchers to question whether facial dynamics are truly informative [31, 32]. For example, Snow et al. [30] have found that speech-related facial movements distract the observer's attention from the identity of a face, thus degrading the recognition performance [30, 32].

To date, the exact role of dynamics in face recognition is still a subject of debate. While some studies report a motion benefit [20, 31, 33, 34], others find no benefit of motion [32, 35, 36]. A closer look shows that the experiment settings differ significantly between these studies, e.g. cooperative or non-cooperative subjects, the facial stimuli, the types of data presented to observers. For example, Snow et al. [30] employ realistic 2D videos of heads involved in rigid motions (Figure 1), whereas Hill & Johnston [27] project facial nonrigid motions of actors onto a synthetic 3D head models (Figure 2). In another scenario, Knappmeyer et al. [34] train participants to discriminate two synthetic faces that were animated with different nonrigid facial motion patterns (Figure 3). All these works employ cooperative subjects under optimal viewing conditions, while results reported by Christie et al. [32] and Bruce et al. [35, 36] employ surveillance videos of non-cooperative subjects. Such disparities may explain why conclusions differ. It emerges however from these studies that different types of motions affect identity perception in different ways.

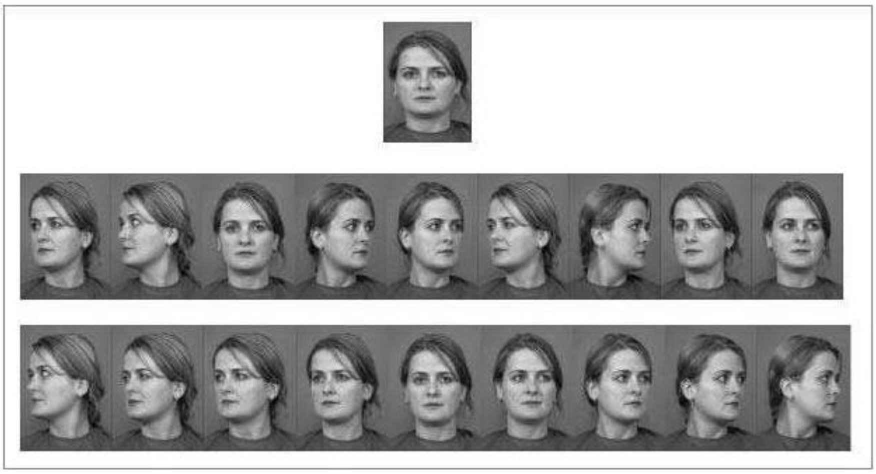

Figure 1: Stimuli for a study comparing static and dynamic face perceptions by humans presented by Snow et al. [30]. Similar recognition rates are observed whether the recognition is based on one static face (row 1), a random set of static faces (row 2), or a true rigid motion sequence (row 3), displayed for identical exposition time. This proves that rigid facial motion does not have any impact on identity recognition.

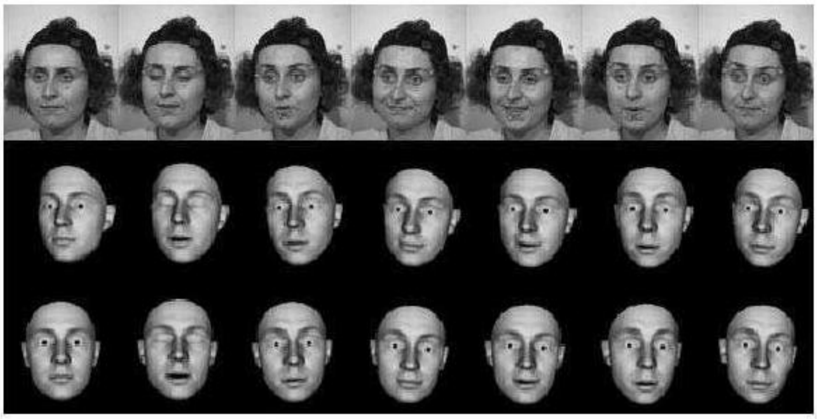

Figure 2: Facial stimulus used in Hill and Johnston [27]. Facial motions of a human actor are transferred to a synthetic head. Observers are asked to recognise both gender and identity from motion information alone. Rigid head motions appear useful for identity recognition, while nonrigid motion allows gender recognition.

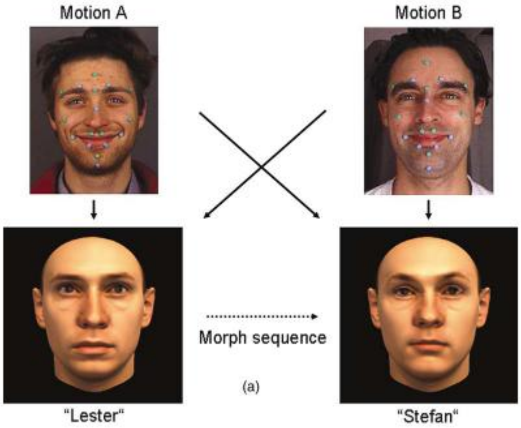

Figure 3: Stimuli used in Knappmeyer et al. [34]. During a training phase observers were familiarised with two moving faces Stefan and Lester, one animated with Motion A and the other animated with Motion B. At test, each face of a morph sequence between Stefan and Lester are combined with Motion A or Motion B. Observers had to decide whether the animated faces looked more like Stefan or Lester. The study found that nonrigid motion biases observers' perception of identity.

It has often been an issue to determine which types of motions - rigid or nonrigid - are the most beneficial for identity recognition. For example, experiments have considered rigid head movements such as nodding [20,25,34], facial expressions [20, 34], blinking rate and speaking patterns such as small mouth movements and upward smile during speech [20]. Here again, there are two different opinions. While Hill & Johnston [27] found only weak benefits of nonrigid motions compared to robust effects of rigid head motions, Knappmeyer et al. [34] found on the contrary strong effects of nonrigid motions. However, unlike Hill & Johnston who employ speech-related movements, Knappmeyer et al. used expressive facial actions (Figure 3).

In summary, examination of related works in perceptual studies above underlines a number of main issues. First, there is no standardised methodology for the study of facial motion in identity perception, making it difficult to compare results and gain reliable knowledge of the problem. One might also express reservation concerning the use of either realistic video, or motion capture to animate synthetic faces. Since both facial forms and facial motions are presented to observers, it is unclear to which extent the physical aspect of the face - artificial or real - influences identity perception. Therefore, one cannot know for sure that the observers' decision making is based solely on motion. Secondly, the experiments carried out in these studies typically employ small numbers of observers, between 13 and 29 in Knappmeyer et al. [34], 49 in Lander et al. [33], 8 healthy observers and 1 prosopagnosic patient in Steede et al. [20]. Since humans possess different abilities to recognise faces, unless experiments are carried out on a statistically significant number of participants, results and interpretations need to be considered with extreme caution. Three indications emerge nevertheless from these studies: (1) There has been so far more evidence indicating that motion is beneficial for identity perception. However, this does not necessarily exclude the other two theories, as these may play complementary roles, and apply to different phases in the process of perceiving, encoding, and remembering a face in different conditions of the recognition task [31], (2) Different types of facial motions seem to affect identity perception in different ways, (3) Familiarity of faces plays a role.

1.3 Neurological basis of face perception

A quantitative method to assess the role of motion in face recognition stems from cognitive neuroscience. In real-life, the way in which the brain simultaneously processes multiple facial stimuli (e.g. facial form, motion, emotion, age, gender) has been extensively studied by means of either Functional Magnetic Resonance Imaging (fMRI) or Positron Emission Tomography (PET) [37–40]. Figure 4 shows a comparison of brain activities across three recognition tasks.

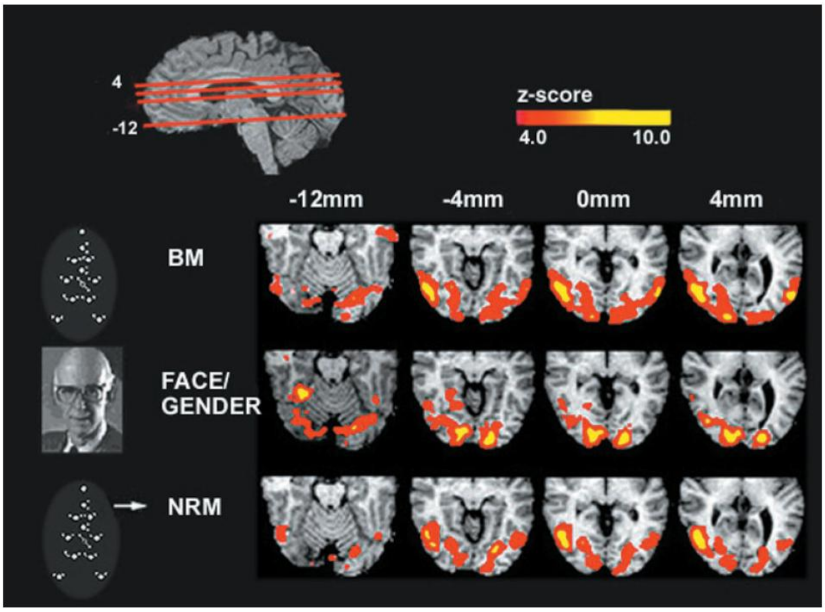

Figure 4: Comparison of brain activities across three recognition tasks. (1) BM: biological motions of a subject wearing reflective markers, (2) Face/Gender recognition, (3) NRM: nonrigid body motions [21].

In general, the recognition process of any object activates areas within the temporal cortices of the brain. For complex objects such as faces, the recognition takes place in several places including the inferior temporal gyrus (ITG), the superior temporal sulcus (STS), the ventral striatum, the inferior convexity, the fusiform gyrus (FG), and the amygdala. Unknown objects are compared to mnemonic representations of known objects stored in memory, and the decision making is achieved through various responses of neurons located in the latter parts of the ventral stream [21, 39, 41].

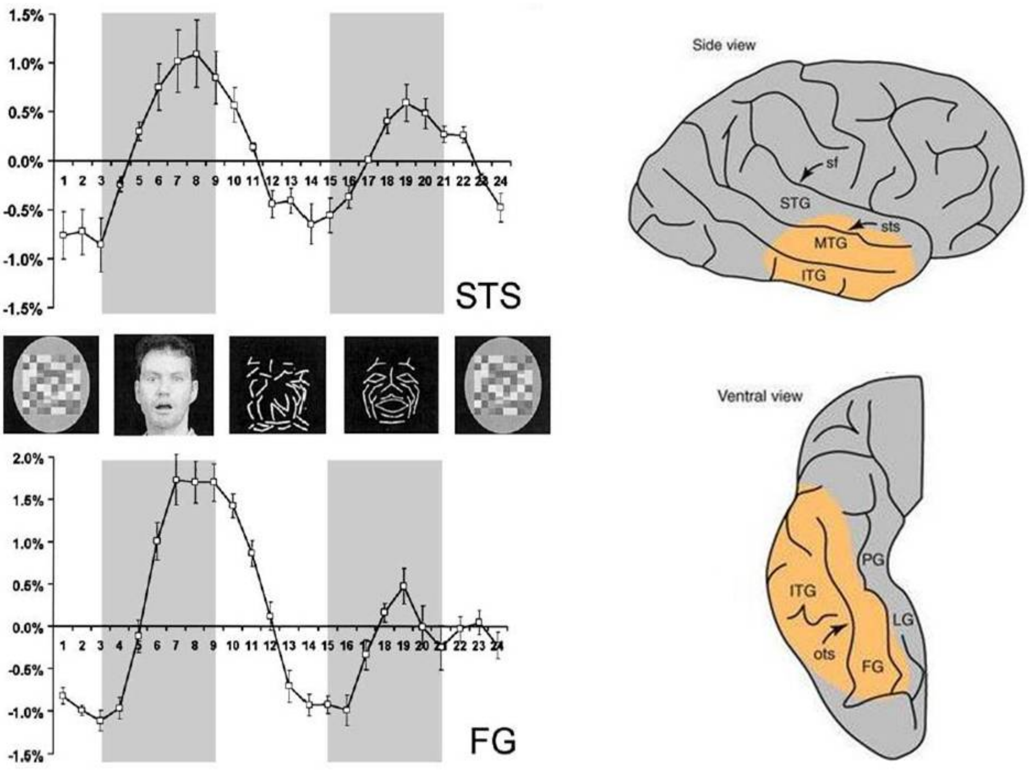

Figure 5: Brain activities in response to different types of stimuli in Puce et al. [41]. FG: fusiform gyrus. ITG: inferior temporal gyrus. LG: lingual gyrus. MTG: middle temporal gyrus. STG: superior temporal gyrus. OTS: occipito-temporal sulcus. STS: superior temporal sulcus.

Although the STS and the FG are commonly thought to be involved in static face recognition, it is unclear whether these regions integrate both form and motion, and to what degree. To answer this question, Puce et al. [41] proposed to investigate brain activation in healthy subjects when presented with either natural or visually impoverished line drawn facial motion displays. They employed alternative mouth opening and closing movements repeated 7 times over a 12-second period. There were no significant rigid head movements. The responses of the STS and the FG are shown in Figure 5. The timing and location of neural activity elicited to both natural and impoverished images of facial motion indicate that both motion types evoked responses in the same cortical region at around the same latency. However, the brain activity induced by realistic facial motion is significantly stronger compared to that induced by drawn line motion.

These results partially contradict earlier work carried out by Haxby et al. [39,40], which suggested that static and dynamic facial stimuli elicited different cognitive mechanisms. According to Haxby, the moving components are processed in the superior temporal sulcus region (STS) of the dorsal visual stream, whereas the static components are processed in the fusiform face area. This theory was later extended by O'Toole et al. [25] who proposed a comprehensive model of face perception, as shown in Figure 6.

There also exists an emerging research trend for assessing the role of motion in face perception, which consists of studying prosopagnosia, an impairment in which a person is unable to recognise familiar faces. To date, only two studies have investigated whether a person who is impaired at static face recognition can use facial motion as a cue to identity. While the prosopagnosic patient studied in Steede et al. [20] performed successfully on all tasks relative to using moving faces, indicating that he could use facial motion to access identity, Lander et al. [42] reported that

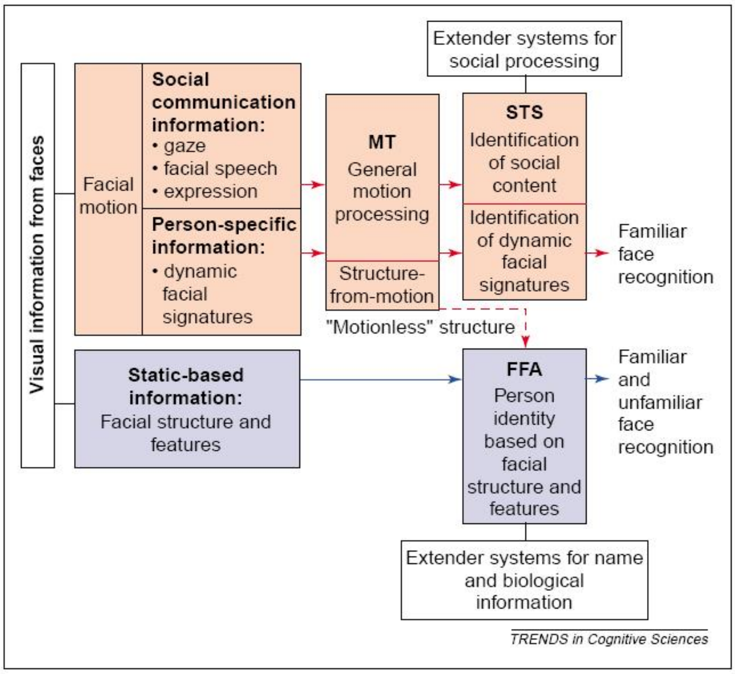
Visual information from faces
Facial motion
Social communication information:
• gaze
• facial speech
• expression
Person-specific information:
• dynamic facial signatures
Static-based information:
Facial structure and features
MT
General motion processing
Structure-from-motion
"Motionless" structure
Extender systems for social processing
STS
Identification of social content
Identification of dynamic facial signatures
Familiar face recognition
FFA
Person identity based on facial structure and features
Familiar and unfamiliar face recognition
Extender systems for name and biological information
TRENDS in Cognitive Sciences

Figure 6: Model proposed by O'Toole et al. [25] for mapping face recognition onto the neural system. MT: middle temporal visual area. FFA: fusiform face area. STS: superior temporal sulcus.

their prosopagnosic patient was unable to use facial motions to explicitly recognise individuals. However, they observed that the patient was significantly better at matching moving faces which exhibit similar patterns, suggesting that motion plays a role to some extent.

Finally, the study of emotion recognition in faces has also received a great deal of interest. It has long been established that the amygdala - a group of neurons located deep within the medial temporal lobes of the brain - is responsible for assigning affective significance to faces. Damage to the fusiform gyrus and the amygdala results in impaired face recognition [43,44]. fMRI scans of healthy subjects show that part of the fusiform gyrus was significantly activated during face recognition [37,38], while fMRI scans of patients with autism and Asperger syndrome show a failure to activate the fusiform face area during face processing [44].

2 Underlying anatomy of facial actions

In proposing facial motion as the basis for a new class of biometrics, it remains to ascertain that it is viable, i.e. it must be permanent and unique to each individual. This aspect has been largely ignored in previous studies [15–18], or merely assumed [45], relying on the fact that if the anatomical structure of a static face is unique, then this uniqueness should be reflected in the dynamic patterns inferred from facial expressions. However, it is unclear whether the uniqueness of static faces itself has ever been formally proven [46]. It is possible that this belief was suggested as a premise long time ago, dating back to at least the Roman Empire, and empirically established over time. Pliny the Elder (23-79 A.D.), in Naturalis Historia, wrote:

> *"The human features and countenance, although composed of but ten parts or little more, are so fashioned that among so many thousands of men there are no two in existence who cannot be distinguished from one another"*

2.1 Uniqueness across individuals

It has been long known to anthropologists that human facial musculature can vary greatly across individuals [47]. In 1931, Huber observed that each ethnicity had its own modes of expression, and attributed this peculiarity to the different arrangements of the facial musculature of each race [48]. For example, the risorius muscle (see Figure 7), which is involved in various expressions e.g. extreme fear and smile, is generally absent in people of Melanesian ancestry. Other recent studies have also reported similar muscle arrangement disparity in subjects within the same ethnic group [49]. Of the core muscles which are present in every individual, structural vari-

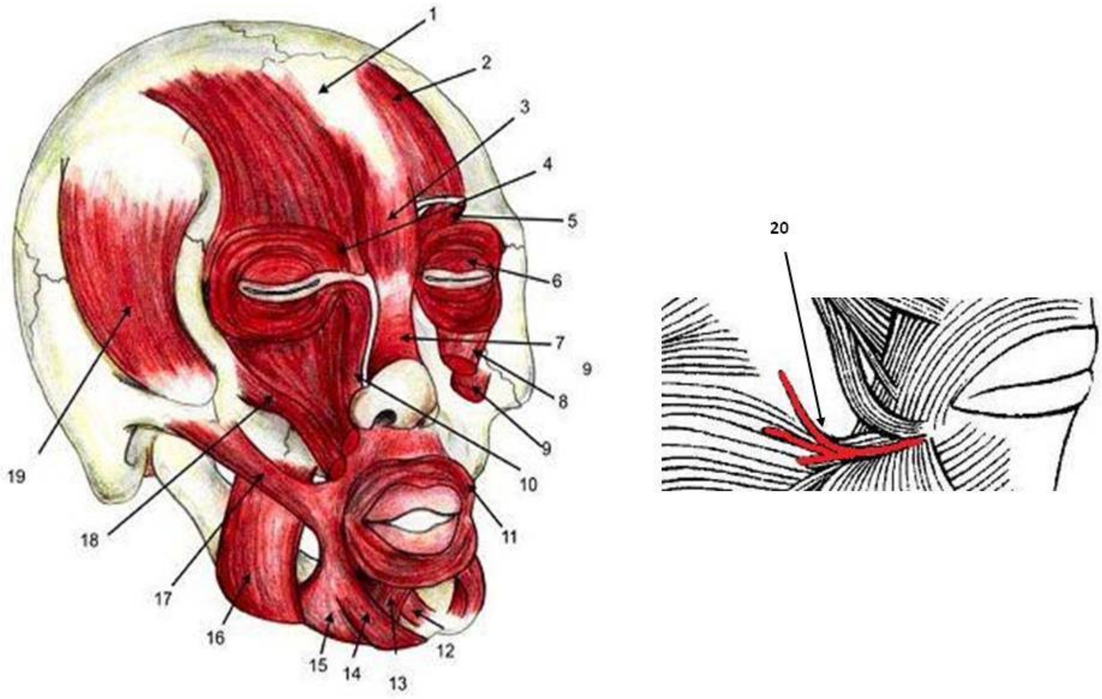

Figure 7: Facial muscles: 1) Galea aponeurotica, 2) Frontalis, 3) Procerus, 4) Depressor supercilii, 5) Corrugator supercilii, 6) Orbicularis oculi, 7) Nasalis, 8) Levator labii superioris, 9) Levator anguli oris, 10) Levator labii superioris alaeque nasi, 11) Orbicularis oris, 12) Mentalis, 13) Depressor labii inferioris, 14) Depressor anguli oris, 15) Platysma, 16) Masseter, 17) Zygomaticus major, 18) Zygomaticus minor, 19) Temporalis, 20) Risorius.

ations can usually be observed in the way they are connected to the bone structure and the soft tissue, and also in their asymmetries. For example, in most individuals, the platysma muscle inserts to the skin over the inferior part of the mandible, but in some individuals, it inserts in the lateral cheek, causing a furrow [50]. Dimples, which are caused by the presence of a bifid zygomaticus major muscle, are present in 17 of 50 subjects studied in Pessa et al. [49,51].

These observations are interesting inasmuch as they allow to explain why some patterns are unique to certain persons, e.g. dimples and ability to raise one eyebrow. This has led Waller et al. [52] to formulate the hypothesis that humans only need a core set of 5 facial muscles to produce basic facial expressions (e.g. anger, happiness, surprise, etc). These involve a combination of basic facial actions which are universally recognizable across individuals of any ethnic groups, age, and gender. For some individuals, however, more muscles can be involved in the production of certain facial expressions, allowing them to exhibit idiosyncrasies. In light of these findings, it appears plausible that moving faces exhibit higher distinctiveness compared to static faces, because certain underlying idiosyncrasies become visible only when the muscles are activated, e.g single or double cheek dimples, muscle folds, and furrows.

2.2 Stability over time

To date, there is not a great deal of published information on the stability of facial actions over time. The only literature stems from either craniofacial research, in the context of dental surgery planning [53–55], or from research on emotions [56,57] based on the Facial Action Coding System (FACS) developed by Ekman et al. [58].

The first systematic study on facial expression reproducibility in craniofacial research was reported by Johnston et al. [53], as shown in Figure 8. The authors

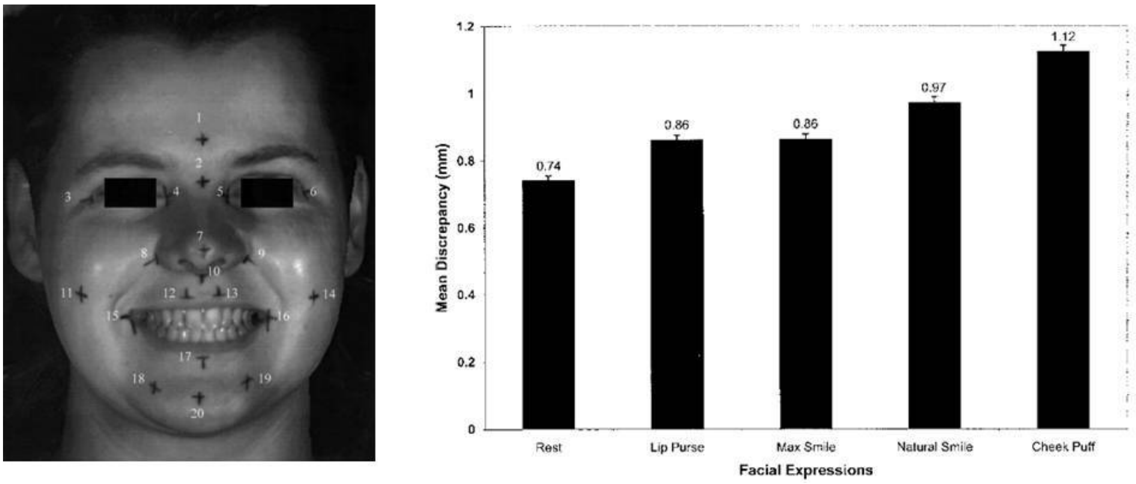

Figure 8: Assessment of facial motion reproducibility in Johnston et al. [53]. Landmarks are manually placed on the subject face, and their displacements across recording sessions are measured.

examined the stability of facial motion in 30 healthy Caucasians over two weeks, using expressions such as lip purse, maximal smile, natural smile, cheek puff, and the rest position where participants were asked to speak short words in a relaxed way. The most interesting aspect that emerges from this study is that there exists a hierarchy in the reproducibility of facial actions, with rest position being significantly more reproducible than lip purse, maximal smile, natural smile, and finally cheek puff. The main limitation of this study is that it uses the displacements of the landmarks manually drawn on faces as a measure to assess the motion stability. Since it is difficult to repeatedly place the landmarks exactly at the same positions across different recording sessions, especially on the cheeks and chin, the assessment might be plagued by landmark placement error. In a more recent study, Popat et al. [54,55] proposed a markerless 3D data analysis method to assess the stability of facial movements in 22 healthy subjects uttering the word 'puppy' within a 10-second time interval. Reproducibility was measured as the percentage point deviation between two corresponding frames. High inter and intra-subject variations were observed across repetitions, with the viseme ‘pup’ appearing less stable than the viseme ’-py’.

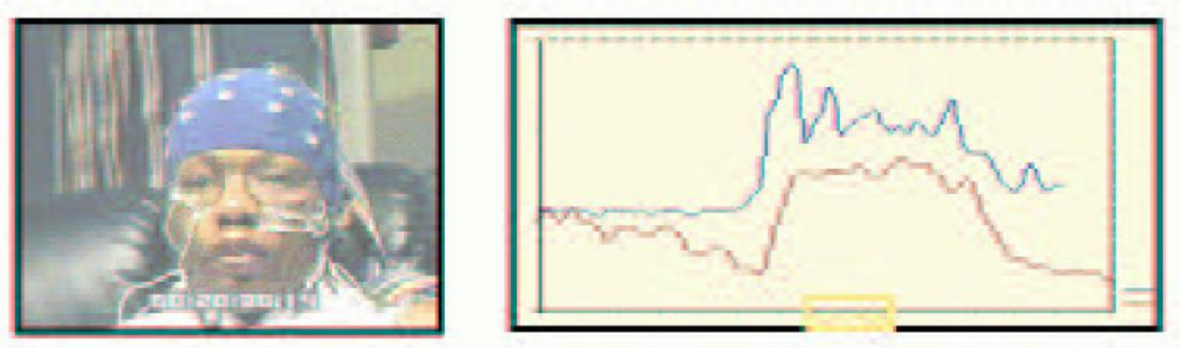

Figure 9: Assessment of face expression reproducibility by Ekman and associates [57]. EMG of the zygomaticus major (blue) and the corresponding temporal variations of lip corner motion (red) in a subject smiling.

To this date, Ekman and his associates [56, 57] have been the only ones to investigate the reproducibility of facial actions specifically for use in identity recognition. Their study, however, is limited to two emotional expressions, happiness and pain. In a first study, 65 young and healthy adults are observed while watching a 5-minute film clip intended to elicit genuine emotions, e.g. smiles. Two methods are used to measure facial activities: electromyography (EMG) of the zygomaticus major muscle, and the polar coordinates of the lip corners, as shown in Figure 9. In a second study, 85 middle-aged to older adults known to have a history of heart disease are observed over a 4-month interval in a clinical interview setting. Facial expressions are manually coded using the FACS coding system [58], then the Pearson’s correlation coefficients are used to compare facial performances [57]. Although high recognition error rates of $\approx 50\%$ have been found, both studies conclude that individual differences in facial expression are stable over time and these can be used for person recognition.

Very little work - if any - has been done so far to study the reproducibility of speech-related facial actions for person identification. The purpose of the next section is to review the literature in articulatory phonetics in order to collect useful information related to both the uniqueness and stability of speech-related facial movements.

2.3 Articulatory phonetics

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 [59,60]. However, lipreading and biometrics are two different problems. While lipreading aims to recognise phonemes from lip shapes (visemes) and must be speaker-independent, biometrics seeks on the contrary to recognise the visemic dissimilarities across speakers uttering the same phoneme.

Research on visual speech, which aims to establish correspondence between phonemes and visemes, focusses only on the lip region. A viseme is a static lip pose associated with a particular sound. The lip dynamic during the transition from one viseme to another is computed through morphing [62, 63]. Table 1 shows a phoneme-to-viseme mapping for English language [61]. This methodology sets a good basis for the synthesis of intelligible and realistic visual speech, but does not provide any insight into the recognition of a person's speaking patterns. Speech production is a far more complex process, and a detailed examination of the vocal track is necessary to understand the idiosyncrasies and the reproducibility of related facial movements.

Figure 10 shows the major speech organs and the principal points of articulation in the vocal track. Speech postures can involve intra-oral points of articulation - e.g. the dental (/t/,/d/) and the velar (/k/,/g/) - or extra-oral points of articulation - e.g.

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

Table 1: Phoneme to viseme mapping for English language [61].

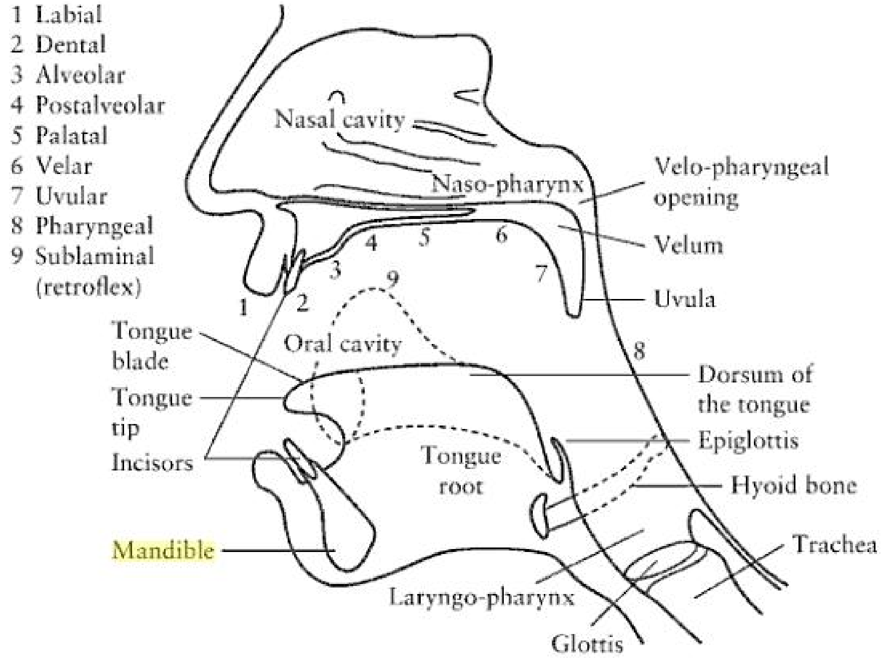

Figure 10: Major speech organs and points of articulation [64].

Speech postures	**Duration (ms)**	**$\approx$ Number of video frames**
unrounded to rounded	50 - 100	3 - 5
protrusion to un-protruded	200 - 300	10 - 15
single reversal	$\geq$ 200	$\geq$ 10
duration of stoppage	40 - 150	2 - 7

Table 2: Examples of speech postures and their typical durations. The corresponding number of frames is calculated for a camera operating at 48fps.

the bilabial (/p/,/b/,/m/) and the labiodental (/f/,/v/). The majority of consonants require physical contacts between two articulators, momentarily blocking the airflow through the vocal track. On the contrary, vowels and some particular consonants such as /w/, do not involve any blockage of the airstream [64].

A common characteristic event is one where two articulators make contact and then release. Several measures can be employed to describe the timing of these movements. Westbury et al. [65] proposed to measure the time taken to perform a single reversal in which the structure moves away from contact and returns to make contact again. The time taken by an articulator to manoeuver from a constricted position to a specific position for a vowel and back again to a constricted position is usually slightly greater than 200ms, except for vowels which require small lip apertures. Movements of the lower lip, the tongue blade and the tongue dorsum are linked to that of the mandible which operates at a period of $\approx$ 150ms. Therefore, alternative opening and closing movements of the lip require at least 150 to 200ms. Kenneth et al. [66] reported that the time taken to manoeuver the lips from an unrounded to a rounded configuration is typically from 50 to 100ms. Perkell et al. [67] estimated the total time required from the beginning of a protrusion movement to a non-protruded configuration for a vowel like /u/ appears to be in the range of 200 to 300ms. Examples of basic speech units and their timing are shown in Table 2.

While the durations of stop consonants (i.e. involving blockage of the vocal track) are constrained by the timing of speech kinetics, utterances of vowels and non-stop consonants (/w/,/r/,/sh/) can be prolonged as long as one's breath lasts, like when *humming a tune.* For nasal (/m/,/n/), the airflow blockage does take place in the oral track, but the velum is lowered to let the air through the nose cavity, allowing the sound to prolong. These findings are consistent with the hypothesis that there exists a hierarchy in the reproducibility of facial actions. Consonants that involve strong contacts between two speech articulators are constrained by timing of the speech kinetics, while vowels and consonants that allow trailing are more likely to be subject to variations, depending on the speech contexts, co-articulations, speech-rate, and the speaker's hesitations.

Hypotheses that explain the uniqueness of nonverbal facial expressions also apply to speech-related facial movements, i.e. the disparities of facial musculatures and their participation in facial expressions across individuals [52]. The Orbicularis oris is the major muscle associated with lip movements. It operates in various combinations with other muscles to yield a considerable range of lip configurations [64]:

- Vowels such as /aw/ (e.g. talk, hoard) have lip rounding and protrusion as part of their articulatory configuration. This requires the Orbicularis oris and the Mentalis, with the latter raising and protruding the lower lip when contracted.

- Precise, rapid closure and release of the lips, as required in bilabial (/p/,/b/), involve the Orbicularis oris to close and hold the lips together, and the levator and depressor muscles to open the lips rapidly at release. Raising of the upper lip is controlled by the Zigomaticus Minor, and the two Levator Labii Superior muscles. Lowering of the lower lip is controlled by the Depressor Labii Inferior.

- Lateral movements of the mouth corners is controlled by the Buccinator and the Risorius which spread the lips, and the Zygomaticus Major which draws the mouth angle back and upwards. For labiodental (/f/,/v/), the Orbicularis oris pulls the lower lip inwards and presses it against the upper teeth, while the Buccinator, Risorius and Zygomaticus Major retract the mouth angles to spread the lips. Visemes such as /ee/ (e.g. heed), require the lips to be spread, also involve the Buccinator, Risorius and Zygomatic Major muscles.

The mandible is capable of movements in vertical, longitudinal and lateral directions. It acts both as a moving articulator an anchor point for a number of muscles which affect and are affected by its movements. Many internal organs of the vocal track are not visible under normal circumstances, they do nevertheless cause visible facial appearance variations. For example, the intra-oral pressure when air is injected into the oral cavity prior to (/p/,/b/) causes a slight cheek puff.

Finally, it is important to notice that language is a social learning process which is influenced by feedbacks from the environment, and may change over time. For native English speakers, language is acquired during childhood, the pronunciation of words is adjusted to the local accent, and reaches a certain stability in adulthood. For non-native speakers, the learning process of a foreign language is also influenced by feedbacks from the environment, combined with particular pronunciations that characterise the accents inherited from the mother tongues. Local dialects also plays a great role. For example, in a study of speech perception, Ladefoged [68] found that 27 out of 30 speakers of Californian English used an inter-dental /θ/ in which the tip of the tongue was protruded between the teeth in words such as ('think', 'thin'), whereas the majority of Southern British speakers used a dental /θ/ without tongue protrusion. Many sociolinguistic differences can be found, for example, 'either' one says /ee-th-er/, and others say /ahy-th-er/, the word 'tomato' is pronounced by some as /tuh-mey-toh/, whereas others say /tuh-mah-toh/. All these peculiarities result in different observable facial deformations caused by the movements of different sets of articulators, making speech-related facial motions highly distinctive across individuals.

3 Automatic machine recognition

Little work has been done so far to investigate the automation of facial motion recognition by machine vision techniques. Of the few attempts, there exist two main trends. The first trend stems from facial expression recognition for behavioural studies, and the second arises from research on lipreading / speech analysis.

3.1 Methods inspired by facial expression recognition

The majority of implementations today are based on algorithms that are borrowed directly from research on facial expression recognition. They employ emotional facial expressions (e.g. smile, surprise, anger) and adopt a *binary gesture* approach, using two images of a person, one neutral face and one at the apex of some expression. Landmarks are placed on key facial features in both images, then the displacements of the corresponding landmark points are measured as an indication of facial movements and used for recognition. The classification employs Hidden Markov Models (HMM) to estimate the AU parameters [69, 70], or other methods such as PCA, ICA, Gabor wavelet, and local Eigen-features [71]. Extensive reviews of different methods have been conducted by Fidaleo in 2003 [72] and Collins in 2006 [73]. The most recent work is that of Tulyakov et al. in 2007, who tested a semi-automated system using 46 sad, 38 happy expression images of 3 persons excerpted from Big Brother footage, with a result of 40% Equal Error Rate using an Eigen-features approach [74].

3.2 Speaker recognition by lipreading

In 1996, Luettin et al. [15] observed clear idiosyncrasies across speakers while studying lipreading, and proposed to use this peculiarity for speaker identification. Audio-video data of 12 subjects uttering digits from ‘1’ to ‘4’ was employed in their test, as shown in Figure 11. Two repetitions were recorded for each subject, the first one was used to train an Active Shape Model [75], and the second one for testing. Using HMM for classification, they reported 72.9% correct matches using shape alone, 89.6% using intensity alone, and 91.7% for combined shape and intensity.

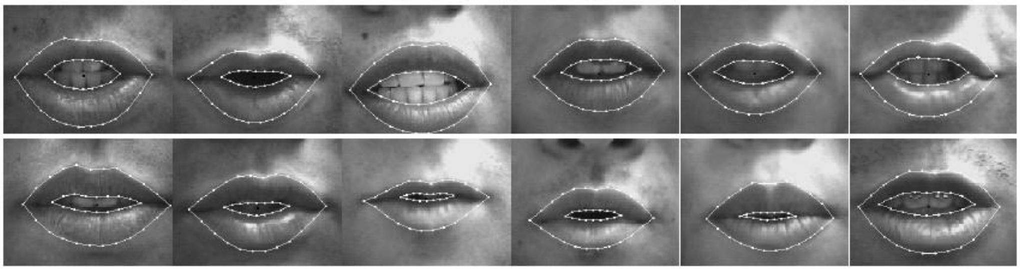

Figure 11: Example images of subjects uttering digits from 1 to 4 and lip tracking results in Luettin et al. [15].

A number of similar studies followed Luettin's work with comparable results [16], of which the most recent is the work of Faraj et al. [18] where the authors reported a recognition rate of 98% on a database of 300 subjects, using both video and audio data in the recognition process. All these systems share two particularities: (1) they employ long speech sequences of uttered digits from '0' to '9', and (2) they focus on the lip motions while excluding all other parts of the face. More recently, in 2009, Tistarelli et al. [76] tested a holistic approach on a database of 21 participants uttering digits from '1' to '10', with 5 recordings from each subject. Using a Pseudo-Hierarchical HMM, they reported a performance of 6% Equal Error Rate.

3.3 Standard face databases

At the moment, the main barrier to experimental study of facial motions for identity recognition is the lack of appropriate face data. Of the previous works, the majority employ standard face databases which have been collected for other research purposes, for instance, the 2D Cohn-Kanade [69] database of FACS Action Units [58] that was aimed for emotion recognition. More recently, a new face database was released by the University of Binghamton, USA which includes 3D video sequences of 101 subjects performing six basic emotional expressions e.g. angry, disgust, fear, happiness, sadness, and surprise [77], as shown in Figure 12.

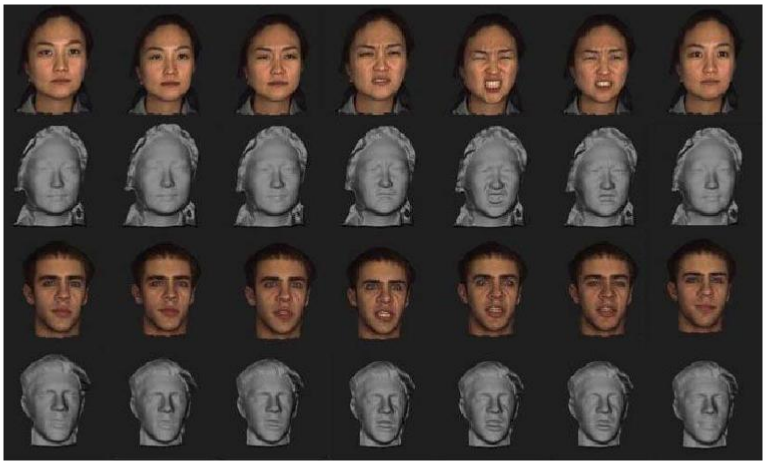

Figure 12: The face database collected at the University of Binghamton which include 3D video data of 101 subjects performing 6 basic emotions [77].

The major limitation of these databases is that they are aimed for facial expression recognition, and do not include repetitions over time. Furthermore, the participants are usually trained by psychologists to produce only 'valid' facial expressions. This type of highly controlled lab-condition data is not suitable for the testing of applications such as biometrics where, in real-life situations, naïve users may interact with the system in the most unpredictable way.

Other popular face databases are those employed in speech research, such as the M2VTS [78] and the XM2VTSDB [79] databases, which include audio-video sequences of continuously uttered digits from '0' to '9', and spoken phrases such as *'Joe took fathers green shoe bench out'*. Also commonly used is the DAVID database [80] where 9 participants wearing blue lipstick utter isolated digits, as shown in Figure 13(a). There exist also customised data sets such as that recorded by Fidaleo et al. [17] as shown in Figure 13(b). All these face databases present a number of limitations that make them unsuitable for use in biometrics. In fact, the use of make-up and devices that involve physical contacts is inconvenient in a real-world situation because this is too invasive and inevitably raises hygiene concerns when deployed in public places e.g. busy airports and ATM machines.

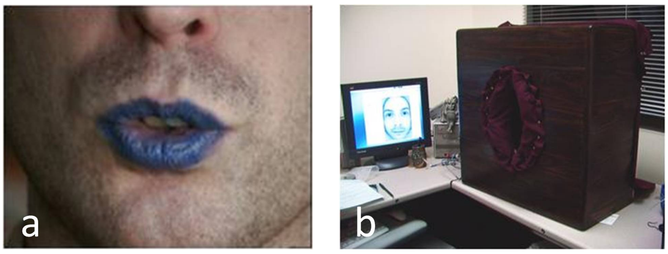

Figure 13: Left: data capture for the DAVID database [80]. Right: a wooden box is used to control the head position in Fidaleo et al. [17].

As far as the choice of speech is concerned, uttered digits from ‘0’ to ‘9’ are not the most adequate for the study of facial motions in biometrics. In fact, although any type of speech can be considered for identity recognition, choosing words which exhibit high biometric power is similar to choosing strong computer login passwords over weak ones. Words such as bilabial articulations (/p/,/b/,/m/) are viseme rich because the variations occur at the visible parts of the face and are easy to capture by a video camera. Therefore, these speech postures should be preferred to intra-oral articulations e.g. (/t/,/n/,/s/). Spoken words such as *‘Mississippi’* used in craniofacial research [53] are also not suitable for biometrics because of the limited variety of visemes involved. It is well known for password-based computer login that repetitive patterns make weak passwords, this also applies to the present case. Furthermore, although long phrases are necessary for speech analysis, they require intensive processing effort, especially in 3D. 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.

References

[1] M. Pantic, “Face for Interface,” Book Chapter. Encyclopedia of Multimedia Technology and Networking **vol. 1,** pp. 308–314 (2005).

[2] T. J. Hutton, B. F. Buxton, P. Hammond, and H. W. W. Potts, “Estimating average growth trajectories in shape-space using kernel smoothing,” IEEE Transactions on Medical Imaging **vol. 22(6),** pp. 747–753 (2003).

[3] T. Kanade, “Picture Processing by Computer Complex and Recognition of Human Faces,” PhD Thesis, University of Kyoto (1973).

[4] R. Brunelli and T. Poggio, “Face recognition: Features versus templates,” IEEE Transactions on Pattern Analysis and Machine Intelligence **vol. 15(10),** pp. 1042–1052 (1993).

[5] L. Wiskott, J. M. Fellous, N. Krueuger, and C. von der Malsburg, “Face Recognition by Elastic Bunch Graph Matching,” Chapter 11 in Intelligent Biometric Techniques in Fingerprint and Face Recognition pp. 355–396 (1995).

[6] M. Turk and A. Pentland, “Eigenfaces for recognition,” Cognitive Neurosciences **vol. 3(1),** pp. 71–86 (1991).

[7] P. Belhumeur, J. Hespanha, and D. Kriegman, “Eigenfaces vs. fisherfaces: recognition using class specific linear projection,” IEEE Transactions on Pattern Analysis and Machine Intelligence **vol. 19(7),** pp. 711–720 (1997).

[8] M. S. Bartlett, J. R. Movellan, and T. J. Sejnowski, “Face Recognition by Independent Component Analysis,” IEEE Transactions on Neural Networks **vol. 13(6),** pp. 1450–1464 (2002).

[9] S. Pamudurthy, E. Guan, K. Mueller, and M. Rafailovich, “Dynamic approach for face recognition using digital image skin correlation,” Audio and Video-Based Biometric Person Authentication **vol. 3546,** pp. 1010–1018 (2005).

[10] W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld, “Face Recognition: A Literature Survey,” Pattern Recognition **vol. 35(4),** pp. 399–458 (2003).

[11] K. Chang, K. Bowyer, and P. Flynn, “Adaptive rigid multi-region selection for handling expression variation in 3D face recognition,” IEEE Workshop on Face Recognition Grand Challenge Experiments pp. 157–157 (2005).

[12] 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(1),** pp. 358–361 (2006).

[13] X. Lu, “3D Face Recognition across Pose and Expression,” PhD Thesis. Michigan State University (2006).

[14] K. Chang, K. Bowyer, and P. Flynn, “Multiple Nose Region Matching for 3D Face Recognition under Varying Facial Expression,” IEEE Transactions on Pattern Analysis and Machine Intelligence **vol. 28(10),** pp. 1695–1700 (2006).

[15] J. Luettin, N. A. Thacher, and S. W. Beet, “Speaker identification by lipreading,” in Proc. of the International Conference on Spoken Language Proceedings pp. 62–64 (1996).

[16] J. D. Brand, J. S. D. Mason, and S. Colomb, “Visual Speech: A Physiological or Behavioural Biometric?,” Audio and Video-Based Biometric Person Authentication. Lecture Notes in Computer Science. Springer, Berlin (2001).

[17] D. A. Fidaleo and M. Trivedi, "Manifold analysis of facial gestures for face recognition," ACM SIGMM Multimedia Biometrics Methods and Application Workshop (2003).

[18] M. I. Faraj and J. Bigun, "Audio-visual person authentication using lip-motion from orientation maps," Pattern Recognition Letters **vol. 28(11),** pp. 1368–13824 (2007).

[19] K. G. Munhall, P. Servos, A. Santi, and M. A. Goodale, "Dynamic visual speech perception in a patient with visual form agnosia," Neuroreport **vol. 13(14),** pp. 1793–1796 (2002).

[20] L. L. Steede, J. J. Tree, and G. J. Hole, "I can't recognize your face but I can recognize its movement," Cognitive Neuropsychology **vol. 24(4),** pp. 451–466 (2007).

[21] L. Vaina, J. Solomon, S. Chowdhury, P. Sinha, and J. Belliveau, "Functional neuroanatomy of biological motion perception in humans," in Proc. of the National Academy of Sciences of the United States of America **vol. 98(20),** pp. 11656–11661 (2001).

[22] B. Knight and A. Johnston, "The role of movement in face recognition," Visual Cognition pp. 264–273 (1997).

[23] K. Lander, F. Christie, and V. Bruce, "The role of movement in the recognition of famous faces," Memory and Cognition **vol. 27,** pp. 974–985 (1999).

[24] K. Lander and L. Chuang, "Why are moving faces easier to recognize?," Visual cognition **vol. 12(3),** pp. 429–442 (2005).

[25] A. J. O'Toole, D. A. R. Dana, and H. Abdi, "Recognizing moving faces: A psychological and neural synthesis," Trends in cognitive sciences **vol. 6(6),** pp. 261–266 (2002).

[26] D. S. Berry, "Child and adult sensitivity to gender information in patterns of facial motion," Ecological Psychology **vol. 3(4),** pp. 348–366 (1991).

[27] H. Hill and A. Johnston, "Categorizing sex and identity from the biological motion of faces," Current Biology **vol. 11,** pp. 880–885 (2001).

[28] D. S. Berry, "What can a moving face tell us?," Journal of Personality and Social Psychology **vol. 58(6),** pp. 1004–1014 (1990).

[29] V. Bruce and T. Valentine, "When a nod's as good as a wink: The role of dynamic information in facial recognition," Practical aspects of memory: Current research and issues pp. 169–174 (1988).

[30] S. Snow, G. Lannen, A. O'Toole, and H. Abdi, "Memory for moving faces: Effects of rigid and non-rigid motion," Journal of Vision **vol. 2(7),** abstract 600 (2002).

[31] D. A. Roark, S. E. Barrett, M. J. Spence, H. Abdi, and A. J. O'Toole, "Psychological and Neural Perspectives on the Role of Motion in Face Recognition," Behavioral and Cognitive Neuroscience Reviews **vol. 2(1),** pp. 15–46 (2003).

[32] F. Christie and V. Bruce, "The role of dynamic information in the recognition of unfamiliar faces," Memory and Cognition **vol. 26,** pp. 780–790 (1998).

[33] K. Lander and V. Bruce, "The role of motion in learning new faces," Visual cognition **vol. 10(8),** pp. 897–912 (2003).

[34] B. Knappmeyer, I. M. Thornton, and H. H. Bulthoff, “The use of facial motion and facial form during the processing of identity,” Trends in cognitive sciences **vol. 43,** pp. 1921–1936 (2003).

[35] V. Bruce, Z. Henderson, K. Greenwood, P. Hancock, A. Burton, and P. Miller, “Verification of face identities from images captured on video,” Journal of Experimental Psychology **vol. 5,** pp. 339–360 (1999).

[36] V. Bruce, Z. Henderson, C. Newman, and A. Burton, “Matching identities of familiar and unfamiliar faces caught on CCTV images,” Journal of Experimental Psychology **vol. 7,** pp. 207–218 (2001).

[37] N. Kanwisher, J. McDermott, and M. Chun, “The fusiform face area: a module in human extrastriate cortex specialized for face perception.,” Journal of Neuroscience **vol. 17,** pp. 4302–4311 (1997).

[38] G. McCarthy, A. Puce, J. Gore, and T. Allison, “Face specific processing in the human fusiform gyrus,” Journal of Cognitive Neuroscience **vol. 8,** pp. 605–610 (1997).

[39] J. V. Haxby, E. A. Hoffman, and M. Gobbini, “The distributed human neural system for face perception,” Trends in cognitive sciences **vol. 4,** pp. 223–233 (2000).

[40] J. V. Haxby, E. A. Hoffman, and M. Gobbini, “Human neural systems for face recognition and social communication,” Biological Psychiatry **vol. 51,** pp. 59–67 (2002).

[41] A. Puce, A. Syngeniotis, J. C. Thompson, D. F. Abbott, K. J. Wheaton, and U. Castiello, “The human temporal lobe integrates facial form and motion:

evidence from fMRI and ERP studies," Neuro-image **vol. 19,** pp. 861–869 (2003).

[42] K. Lander, G. Humphreys, and V. Bruce, "Exploring the role of motion in prosopagnosia: Recognizing, learning and matching faces," Neurocase **vol. 10(6),** pp. 462–470 (2004).

[43] A. Damasio, J. Damasio, and G. V. Hoesen, "Prosopagnosia: anatomic basis and behavioral mechanisms," Neurology **vol. 32,** pp. 331–341 (1982).

[44] R. Schultz, I. Gauthier, A. Klin, R. Fulbright, A. Anderson, F. Volkmar, P. Skudlarski, C. Lacadie, D. Cohen, and J. Gore, "Abnormal ventral temporal cortical activity during face discrimination among individuals with autism and Asperger syndrome," Archives of General Psychiatry **vol. 57,** pp. 331–340 (2000).

[45] E. Y. Zhang, S. J. Kundu, D. B. Goldgof, S. Sarkar, and L. V. Tsap, "Elastic Face, An Anatomy-Based Biometrics Beyond Visible Cue," in Proc. of International Conference on Pattern Recognition **vol. 2,** pp. 19–22 (2004).

[46] A. K. Jain, A. Ross, and S. Pankanti, "Biometrics: A tool for information security," IEEE Transactions on Information Forensics and Security **vol. 1(2),** pp. 125–143 (June 2006).

[47] G. Duchenne, "The mechanism of human facial expression," New York: Cambridge University Press (1859).

[48] E. Huber, "Evolution of the Facial Musculature and Facial Expression," Baltimore: Johns Hopkins Press; London: Oxford University Press (1931).

[49] J. Pessa, V. Zadoo, E. J. Adrian, C. Yuan, J. Aydelotte, and J. Garza, "Variability of the midfacial muscles: analysis of 50 hemifacial cadaver dissections," Plastic Reconstruction Surgery **vol. 102,** pp. 1888–1893 (1998).

[50] K. L. Schmidt and J. F. Cohn, "Human facial expressions as adaptations: Evolutionary questions in facial expression," American Journal of Physical Anthropology **vol. 116,** pp. 03–24 (2001).

[51] J. Pessa, V. Zadoo, P. Garza, E. J. Adrian, A. Dewitt, and J. Garza, "Double or bifid zygomaticus major muscle: anatomy, incidence, and clinical correlation," Clinical Anatomy **vol. 11,** pp. 310–313 (2003).

[52] B. M. Waller, J. J. C. JJ, and A. M. Burrows, "Selection for universal facial emotion," Emotion **vol. 8(3),** pp. 435–439 (2008).

[53] D. Johnston, D. T. Millett, and A. F. Ayoub, "Are Facial Expressions Reproducible?," Cleft Palate-Craniofacial Journal **vol. 40(3),** pp. 291–296 (2003).

[54] H. Popat, S. Richmond, R. Playle, D. Marshall, P. Rosin, and D. Cosker, "Three-dimensional motion analysis - an exploratory study. Part 1. Reproducibility of facial movement," Orthodontics and Craniofacial Research **vol. 11,** pp. 216–223 (2008).

[55] H. Popat, S. Richmond, R. Playle, D. Marshall, P. Rosin, and D. Cosker, "Three-dimensional motion analysis - an exploratory study. Part 2. Reproducibility of facial movement," Orthodontics and Craniofacial Research **vol. 11,** pp. 224–228 (2008).

[56] K. Schmidt and J. Cohn, "Dynamics of facial expression: Normative characteristics and individual differences," IEEE International Conference on Multimedia and Expo pp. 728–731 (2001).

[57] J. F. Cohn, K. Schmidt, R. Gross, and P. Ekman, “Individual Differences in Facial Expression: Stability over Time, Relation to Self-Reported Emotion, and Ability to Inform Person Identification,” in Proc of the International Conference on Multimodal User Interfaces **vol. 116,** pp. 491–498 (2002).

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

[59] 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).

[60] W. M. Weikum, “Visual Language Discrimination,” PhD thesis, The University of British Colombia, Vancouver (2008).

[61] P. Lucey, T. Martin, and S. Sridharan, “Confusability of Phonemes Grouped According to their Viseme Classes in Noisy Environments,” in Proc. of the 10th Australian International Conference on Speech Science and Technology pp. 265–270 (2004).

[62] N. Chomsky and M. Halle, “The Sound Pattern of English,” Harper and Row, New York (1968).

[63] T. Chen, “Audio-visual integration in multimodal communication,” in Proc. of the IEEE pp. 837–852 (1998).

[64] C. Yallop, J. Fletcher, and J. Clark, “An Introduction to Phonetics and Phonology. 3rd Revised Edition,” Blackwell Publishing Ltd (2006).

[65] J. R. Westbury and J. Dembowski, “Contextual influences on stop consonant articulatory postures in connected speech.,” in Proceedings of the XIVth International Congress of Phonetic Sciences pp. 2419–2422 (1999).

[66] K. N. Stevens, “Acoustic Phonetics,” MIT Press, Cambridge, MA, USA. (2000).

[67] J. S. Perkell, “Coarticulation strategies: Preliminary implications of a detailed analysis of lower lip protrusion movements.,” Speech Communication **vol. 5(1),** pp. 47–68 (1986).

[68] P. Ladefoged, “A Course in Phonetics,” Heinle, division of Thomson Learning; International edition (2005).

[69] J. Cohn, A. Zlochower, J. Lien, T. Kanade, and Y. Wu, “Automated face coding: A computer-vision based method of facial expression,” Psychophysiology **vol. 35(1),** pp. 35–43 (1999).

[70] J. Lien, T. Kanade, J. Cohn, and C. Li, “Automated facial expression recognition based on facs action units.,” In Proceedings of the 3rd International Conference on Automatic Face and Gesture Recognition pp. 390–395 (1998).

[71] G. Donato, M. S. Bartlett, J. C. Hager, P. Ekman, and T. J. Sejnowski, “Classifying facial actions,” IEEE Transactions on Pattern Analysis and Machine Intelligence **vol. 21,** pp. 974–989 (1999).

[72] D. A. Fidaleo, “G-Folds: an appearance-based model of facial gestures for performance driven facial animation,” PhD Thesis. University of Southern California. (2003).

[73] T. Collins, “Facial Dynamics for Identity Recognition. PhD Thesis Proposal, supervisor Pr. B. Fisher,” Institute for Perception, Action and Behaviour. School of Informatics. University of Edinburgh (2006).

[74] S. Tulyakov, T. Slowe, Z. Zhi, and V. Govindaraju, “Facial Expression Biometrics Using Tracker Displacement Features,” in Proc of IEEE Conference on Computer Vision and Pattern Recognition pp. 1–5 (2007).

[75] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham, "Active Shape Models - Their Training and Application," Computer Vision, Graphics and Image Understanding **vol. 1(61),** pp. 38–59 (1995).

[76] M. Tistarelli, M. Bicego, and E. Grosso, "Dynamic face recognition: from Human to Machine Vision," in Journal of Image and Vision Computing **27(3),** 222–232 (2009).

[77] L. Yin, X. Chen, Y. Sun, T. Worm, and M. Reale, "A High-Resolution 3D Dynamic Facial Expression Database," in Proc. of the 8th International Conference on Automatic Face and Gesture Recognition (FGR08) pp. 1–6 (2008).

[78] S. Pigeon and L. Vandendrope, "The M2VTS Multimodal Face Database," in Proc. of the First International Conference on Audio and Video-Based Biometric Person Authentication **vol. 1206,** pp. 403–409 (1997).

[79] 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).

[80] C. C. Chibelushi, S. Gandon, J. S. Mason, F. Deravi, and D. Johnston, "Design Issues for a Digital Integrated Audio-Visual Database," Colloquium on Integrated Audio-Visual Processing for Recognition, Synthesis and Communication pp. 1–7 (1996).