PHOTOREALISTIC RETRIEVAL OF OCCLUDED FACIAL INFORMATION USING A PERFORMANCE-DRIVEN FACE MODEL

by

Fatos Berisha

A thesis submitted in partial fulfillment of the requirements for the degree of

Doctorate of Philosophy

Department of Psychology, UCL

2009

Date

____

UNIVERSITY COLLEGE LONDON

ABSTRACT

PHOTOREALISTIC RETRIEVAL OF OCCLUDED FACIAL INFORMATION USING A PERFORMANCE-DRIVEN FACE MODEL

by Fatos Berisha

Facial occlusions can cause both human observers and computer algorithms to fail in a variety of important tasks such as facial action analysis and expression classification. This is because the missing information is not reconstructed accurately enough for the purpose of the task in hand. Most current computer methods that are used to tackle this problem implement complex three-dimensional polygonal face models that are generally time-consuming to produce and unsuitable for photorealistic reconstruction of missing facial features and behaviour.

In this thesis, an image-based approach is adopted to solve the occlusion problem. A dynamic computer model of the face is used to retrieve the occluded facial information from the driver faces. The model consists of a set of orthogonal basis actions obtained by application of principal component analysis (PCA) on image changes and motion fields extracted from a sequence of natural facial motion (Cowe 2003). Examples of occlusion affected facial behaviour can then be projected onto the model to compute coefficients of the basis actions and thus produce photorealistic performance-driven animations.

Visual inspection shows that the PCA face model recovers aspects of expressions in those areas occluded in the driver sequence, but the

expression is generally muted. To further investigate this finding, a database of test sequences affected by a considerable set of artificial and natural occlusions is created. A number of suitable metrics is developed to measure the accuracy of the reconstructions. Regions of the face that are most important for performance-driven mimicry and that seem to carry the best information about global facial configurations are revealed using *Bubbles*, thus in effect identifying facial areas that are most sensitive to occlusions.

Recovery of occluded facial information is enhanced by applying an appropriate scaling factor to the respective coefficients of the basis actions obtained by PCA. This method improves the reconstruction of the facial actions emanating from the occluded areas of the face. However, due to the fact that PCA produces bases that encode composite, correlated actions, such an enhancement also tends to affect actions in non-occluded areas of the face. To avoid this, more localised controls for facial actions are produced using independent component analysis (ICA). Simple projection of the data onto an ICA model is not viable due to the non-orthogonality of the extracted bases. Thus occlusion-affected mimicry is first generated using the PCA model and then enhanced by accordingly manipulating the independent components that are subsequently extracted from the mimicry. This combination of methods yields significant improvements and results in photorealistic reconstructions of occluded facial actions.

Table of Contents

List of figures

Acknowledgments

The author wishes to thank his PhD supervisors Profs Alan Johnston and Peter W. McOwan. It has been a real privilege working with them and it is *their* tireless application, generosity, guidance and encouragement that made the completion of this thesis possible.

Publications and Demos

PUBLICATIONS

Manuscript in preparation

Berisha, F., Johnston, A. & McOwan, P. Identifying regions that carry the best information about global facial configurations. (To be submitted to *Proceedings of the Royal Society of London B*).

POSTER PRESENTATIONS

2007

Berisha, F., Johnston, A., & McOwan, P. Spatial location of critical facial motion information for PCA-based performance-driven mimicry. *Vision Sciences Society 7th Annual Conference*, 11 – 16 May 2007, Sarasota, FL., USA.

2006

Berisha, F. 'Lip-reading' the masked face: a PCA approach to solving facial occlusion problems in natural motion sequences. *UCL Postgraduate Conference at Cumberland Lodge*, 24 – 25 April 2006, Great Windsor Park, UK.

2005

Cowe, G., Johnston, A., & Berisha, F. Principal components of natural facial motion. *EPSRC Doctoral Training Centres Workshop*, 5 – 6 September 2005, Birmingham, UK.

Refereed Conference Abstracts

2007	Berisha, F., Johnston, A., & McOwan, P. Spatial location of critical facial motion information for PCA-based performance-driven mimicry [Abstract]. *Journal of Vision*, 7(9):495, 495a, http://journalofvision.org/7/9/495/, doi:10.1167/7.9.495.

Science Exhibitions/Demos

2007	The perception deception: maths-made optical illusions: How can mathematics help us understand how our brain works? *Royal Society Annual Summer Science Exhibition*, 2 – 5 July 2007, London, UK.

Glossaries

Glossary of Abbreviations

2D	Two dimensional
3D	Three dimensional
AU	Action Unit (one of the 50 constituent muscles/muscle groups that make up FACS)
CSG	Constructive Solid Geometry
EMG	Electromyography
FACS	Facial Action Coding System (Ekman & Friesen 1978)
FACS+	Essa and Pentland's modification of FACS (Essa & Pentland 1997)
ICA	Independent Components Analysis (Bell & Sejnowski 1995)
McGM	Multi-channel Gradient Model (Johnston et al 1999)
MRI	Magnetic Resonance Imaging
PCA	Principal Components Analysis

Glossary of Terms

Avatar	Computer-generated virtual face model
Correspondence problem	Given an image, the task of locating each pixel's new location in a subsequent image
Dot-tracking	The dynamic tracking of markers physically attached to an actor
Drive	To manipulate an avatar through an actor's movements
Eigenfaces	The principal components extracted from a set of images vectorised by their pixel-wise intensities
Face space	A multi-dimensional space in which axes define facial characteristics or movements
Flow-field	An array of motion vectors
Hard-coded	Manually implemented
Hard-wired	Determined by neurological or physiological mechanisms
Hollow face illusion	A visual illusion in which a concave facial surface appears to be convex
Iconic changes	Image changes due to the appearance or occlusion of features

Inversion effect The well-documented impairment in processing of faces when viewed upside-down (Valentine 1988; Yin 1969; Yin 1970b)

Laser scanning The acquisition of 3D surface information using lasers

Lighting changes Image changes due to lighting

Morphing The smooth transition between two images with a combination of warping and blending

Optic flow See 2.3.2

Performance-driven animation The process in which an actor’s motion is tracked and then transferred onto a computer-generated model, to imitate movement

Photogrammetry The extraction of three dimensional information from calibrated photographs

Photo-realistic A sufficient level of realism such that a synthetic image could feasibly be mistaken for a photograph

Pixel Picture element. The smallest resolvable rectangular area of an image on screen or in memory

Pixel-wise intensity vectorisation A vectorisation resulting from a list of the image pixel grey levels

Polygonal mesh A computer-generated surface consisting of polygonal (usually triangular) elements

Prosopagnosia A deficit in which all faces appear unfamiliar (Bodamer 1947)

Rendering The projection of a three dimensional scene onto a 2D image plane

Shape The information describing how the features of a facial image differ in their position from a standard

Shape-free See Texture

Snake A deformable contour model (Terzopoulos & Waters 1993)

Texture The image information once a face has been warped onto a standard shape

Thatcher illusion An visual illusion in which a face altered by inverting the eye and mouth regions is not perceived to be abnormal when the whole altered image is inverted (Thompson 1980)

Videoconference A conference linking participants at different locations using telecommunications and video technology

Voxel Volume element. The smallest resolvable box-shaped part of a three dimensional space

Warp To distort an image with a flow field, such that each image location is repositioned at its destination

Glossary of Mathematical Terms and Symbols

w	Image width
h	Image height
$\mathbf{x}_i$	Training vectors
$\mathbf{X}$	Matrix constructed with $\mathbf{x}_i$ s as its columns
N	Dimensionality of vectorisation
M	Number of training vectors
P	The number of vectors in a basis set
$\boldsymbol{\mu}$	Mean of training vectors
$\boldsymbol{\varphi}_i$	Mean-centred training vector ($\boldsymbol{\varphi}_i = \mathbf{x}_i - \boldsymbol{\mu}$)
$\boldsymbol{\Phi}$	Matrix constructed with $\boldsymbol{\varphi}_i$ s as its columns
$\mathbf{b}_i$	The i^{th} principal component of $\mathbf{X}$
$\mathbf{B}$	Matrix constructed with $\mathbf{b}_i$ s as its columns

$[\mathbf{U},\mathbf{V}]$	A vector flow field where $\mathbf{U}$ is the matrix of x components and $\mathbf{V}$ is the matrix of y components
$\mathbf{y}_i$	Driving vectors
$\boldsymbol{\psi}_i$	Mean-centred driving vector ($\boldsymbol{\psi}_i = \mathbf{y}_i - \boldsymbol{\mu}_{drive}$)
$\boldsymbol{\Psi}$	Matrix constructed with $\boldsymbol{\psi}_i$ s as its columns
$\mathbf{c}_i$	Coefficient vectors (weights on basis vectors)
$\mathbf{z}_i$	Output vectors

Chapter 1

INTRODUCTION

Faces are organs of communication *par excellence* that display an astonishing array of social signals. These signals can be detected and interpreted effortlessly and with good precision by another human being and convey important information about the bearer, such as their age, their sex or their identity. Facial expressions and gestures also inform us about their emotional state, and verbal & non-verbal communication is supported by our perception of facial motion. Nevertheless, the apparent ease with which human subjects or even computer models perform tasks like facial action analysis and expression classification quickly disappears when some visual facial information is missing due to occlusions of the face.

This thesis concerns the accurate retrieval of such occluded visual facial information using a photo-realistic, moveable computer model of the face that synthesises human faces and their movements from standard video footage examples. The thesis sets out with an overview of the psychology of face perception and leads to the subject of computer modelling of the face. Thereafter, the automatic and example-based computer-generated facial model (Cowe 2003) is introduced and its performance in retrieving occluded facial information is tested and analyzed. After addressing the weaknesses of the model and locating important areas of the face for performance-driven mimicry, several enhancements are developed resulting in photo-realistic reconstructions of occluded facial actions.

Biological structure of the face

Through a combination of evolutionary and genetic influences, the human face just as that of many animals is a symmetric structure with eyes placed horizontally above a single centrally placed nose and mouth. Our skulls and foreheads have evolved to the current size to house the brain; the shapes of the jaws, teeth were influenced by our ancestors' diets while the sizes and positions of our nasal cavity and eyes complement our species' other predatory traits.

While the hard and soft tissues of the face produce the individual variations in appearance important for identification and categorisation, it is the movements of the face which give it this ability to transmit a range of social signals. Emotional expressions are a result of movements of facial skin and connective tissue, or *fascia*, caused by the contraction of one or more of the 44 bilaterally symmetrical facial muscles, of which only four are attached to and move skeletal structures (e.g. the jaw) in mastication, while the others innervated by the fascial nerve operate to arrange facial features in meaningful or functionally useful configurations (Rinn 1984). This is not to say, however, that the facial muscle configuration has evolved thus specifically to facilitate facial expression:

> "there are no grounds, as far as I can discover, for believing that any muscle has been developed or even modified exclusively for the sake of expression" (Darwin 1872 - page 355).

According to Darwin and others since, our specific human expressive movements are seen as remnants of behavioural responses to emotionally arousing events.

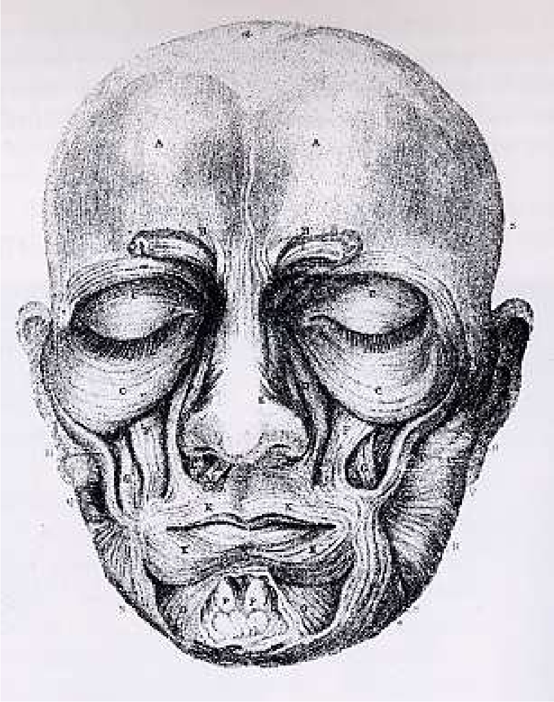


Figure 1.1 - Illustration of the muscles of the face, by Sir Charles Bell

The structure of the face thus makes it an ideal medium or system for transmitting a wide range of different signals and messages, capable of tremendous flexibility and specificity. This system conveys information using four classes of signals (Ekman & Friesen 1975):

1) static facial signals: permanent features of the face that contribute to facial appearance, like the bony structure and soft tissue masses
2) slow facial signals: changes in facial appearance over time, like wrinkles, texture, etc.
3) artificial signals: exogenously determined features, such as eyeglasses, and cosmetics
4) rapid facial signals: phasic changes on neuromuscular activity that lead to visible changes in facial appearance.

These four types of signals and different combinations thereof contribute to perceptions of facial identity, expression, sex, etc. This type of classification formed the basis of early attempts to record and encode individual facial expressions in a comparable and interpretable manner, namely the *Facial Action Coding Scheme* (FACS) (Ekman & Friesen 1978). Such a system however is highly unlikely to be employed in biological vision, since FACS offers no temporal encoding for an inherently dynamic stimulus such as the face. FACS parameters were later extended by adding temporal information via optic flow data, giving rise to FACS+ (Essa & Pentland 1997).

Biological face processing

Given this diversity of signals and information that can be extracted from a face, we might expect face perception to be accomplished by a system with multiple components (Bruce & Young 1986 - see fig.1.2 below).

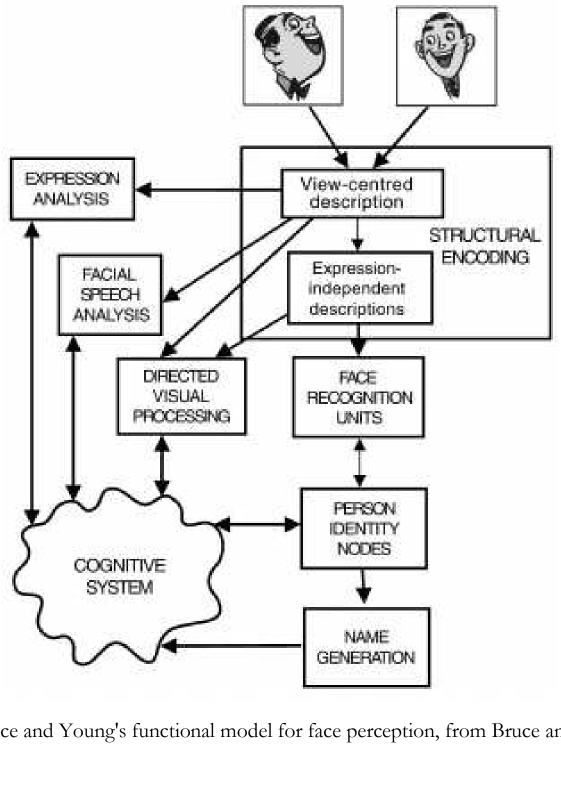


Figure 1.2 - Bruce and Young's functional model for face perception, from Bruce and Young (1986)

Indeed, evidence from neuropsychology suggests that dissociable neural systems exist for different face processing tasks. For example in the recognition of individual faces, an early event-related potential (ERP) N170 response to faces was recorded, whether those faces were familiar or unfamiliar, probably originating in ventral occipito-temporal cortex, but later ERPs occurring between 250ms and 500ms after stimulus presentation seem to be sensitive to face familiarity (Bentin et al 1996; Bentin et al 1999). Other distinct components of the face processing system were found to be responsible for the discrimination of emotional expressions (Schyns et al 2007) and the discrimination of the direction of overt attention, i.e. gaze (Puce et al 2000). But recent work seems to suggest that as much as three functionally dissociated neural mechanisms are involved in general face processing, namely one for the task of detection, one for configural analysis and one for recognition (Anaki et al 2007; Flevaris et al 2008).

So it seems that cognitive neuroscientists have made great advances in identifying indicators of neural substrates involved in extracting the different types of information conveyed by faces. But does this mean that each of these neural substrates is a face-specific processing component that together with the others forms a system which is itself face-specific or does this system (and all its implied components) only discriminate between similar exemplars of the same category, faces being prototypical stimuli but not the exclusive ones? This question was asked by Thierry et al (2007ab; 2007ba), pointing out that the face stimuli that elicit N170 were nearly always presented in full-frontal view while the other stimuli are more perceptually variable, leading to uncontrolled inter-stimulus perceptual variance (ISPV). Their findings seem to call into question the face selectivity of the N170 and establish ISPV as a critical factor to control in experiments relying on multi-trial averaging. However, their claims have been refuted and substantially weakened by subsequent studies showing amongst other things that control of ISPV does not abolish the N170 face effect (Bentin et al 2007).

Face-specific area

The debate about the spatial loci of face perception is just as active as the one described in the previous section. It focuses especially on the early components of face recognition centred on the mid-fusiform gyrus, called the *fusiform face area* (FFA) (Kanwisher et al 1997).

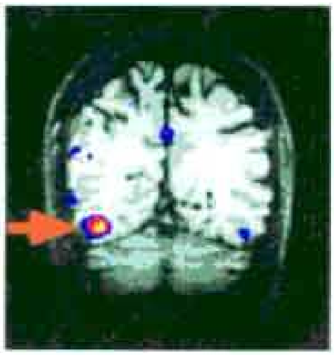


Figure 1.3 - Arrow points at fusiform face area (FFA)

Tong et al. (2000) show that this area is activated by a wide variety of face stimuli (including cartoon faces and cat faces) compared to other non-face objects. They argue that this region is selectively involved in some aspect of the perceptual analysis of faces such as the detection of a face in an image or the structural encoding of the information necessary for face recognition. Indeed, patients with lesions in the said area do display a deficit – *prosopagnosia* - where all faces appear unfamiliar to them, even in cases failing to recognise their own face (Bodamer 1947).

By contrast, Gauthier et al. (2000) argue that the FFA is more active when subjects make subordinate-level classifications than basic-level classifications (e.g. classifying a particular canine as a beagle rather than a dog). This finding is at odds with the findings of Kanwisher et al. and Tong et al. which were showing that the response in the FFA was at least twice as strong when discriminating between faces as compared with within-class discriminations between hands, houses and backs of human heads. One explanation given by Kanwisher is that the discrepancy might arise due to Gauthier et al. using a different technique to identify the FFA from that originally proposed by Kanwisher et al (1997), thus inadvertently looking in the wrong place. If this is really the reason for the discrepancy, then a solution might be found very soon, bearing in mind the fast advances made in fMRI technology, and Grill-Spector, Kanwisher and Chun. (2004) do produce further proof that the FFA indeed is face specific. As things stand, this debate still actively continues.

Face-specific mechanisms

The problem of specificity doesn't concern only the FFA. It is also thought that the brain has special mechanisms devoted to the sole purpose of processing faces. A single-case study of a 16-year old boy named Adam (Farah et al 2000) who became prosopagnosic following bilateral infarction in the occipital lobes at one day of age, reveals that Adam could recognise objects much better than faces. This seems to suggest that there exists some innate face-specific mechanism whose function can't be assumed by other structures, despite plenty of time and opportunity for it to happen.

However, Gauthier and Logothetis (2000) again argue that through extensive training specialised mechanisms can be acquired whose characteristics will resemble, and may even overlap or be identical with, those used to recognise faces. As evidence, they cite Gauthier and Tarr's work on recognising "greebles" which, like faces, share a common spatial configuration (Gauthier & Tarr 2002; Gauthier et al 1998).

Recently presented key new evidence from McKone, Kanwisher et al (2007) from multiple approaches – behavioural studies, neuropsychology, brain imaging and monkey single-unit recording – argues strongly in support of face specificity over expertise. Future work using fMRI in monkeys should allow for a more precise survey of a large cortical territory and at a variety of stages during the acquisition of extensive expertise with novel objects. In combination with studies of the deficits that result from brain damage, this work will be crucial in resolving the debate about the specificity of visual recognition systems. They may also help answer the even more fundamental question: what might a face 'template' look like and how would it perform (both computationally and neurally) holistic processing.

Analytic vs Holistic face processing

A popular hypothesis is that object recognition is *analytic* and part-based whereas face recognition is *holistic* and configural.

Inverting faces, for example, has a disproportionately large detrimental effect on recognition than for most other objects. When testing recognition with pictures of faces against houses, planes and schematic men-in-motion, Yin found recognition of faces to be superior when upright, but when the pictures were inverted, performance on faces was degraded far worse than for any of the other stimuli (Bradshaw et al 1980; Phelps & Roberts 1994; Valentine & Bruce 1986; Yin 1969; Yin 1970a). Similar results have also been found when comparing faces to houses and words (Farah et al 1998). These findings seem to support the above hypothesis because inversion impairs the perception of the spatial configuration among features on which face recognition depends more than identification of the features themselves, which would suffice for much of object recognition. They also establish inversion as a marker of face-specific processes and a tool for investigating what makes face-recognition special.

Others have postulated that this effect could occur simply because of our much greater exposure to upright faces, and in some cases reproduced the effect with other subjects and objects, like with experts vs non-experts tested in recognising inverted images of pedigree dogs, the experts suffering a greater impairing effect (Diamond & Carey 1986). Experiments on the recognition of other-race faces seem to contradict this view, since inversion of Black faces caused Caucasian subjects to display a greater impairment in recognition of Black faces rather than Caucasian faces (Valentine & Bruce 1986). It seems this result intimates that other-race faces are encoded in a somewhat less efficient manner, and inversion then hinders their decoding even more.

What is intriguing in all this is that Tong et al. (2000) found only a slight reduction in activation of the FFA when examining the effects of face inversion, the response to inverted faces still remaining much higher than the response to objects, which seems to suggest that even inverted faces are not treated as objects by the FFA. A solution to this quandary was suggested by Moscovitch & Moscovitch (2000). They propose that the object system forms a representation of the face based on information congruent with its operating characteristics, which it then transfers to the FFA for further processing. The FFA, in turn, sends its output to more anterior regions for identification. Thus, even inverted faces should activate the FFA, though not as strongly, and at a delay, compared to upright faces, as shown later experimentally by Haxby et al. (1999). So the holistic processing paradigm does seem to hold here. Alternatively, as Maurer et al. (2002) claim, more than one configural processing action may take place, the holistic one being only one of them:

> "… first order relations that define faces (i.e. two eyes above a nose and mouth), holistic processing (glueing the features together into a gestalt), and processing second-order relations (i.e. the spacing among features)."

Powerful configural effects have been found with faces using psychophysical tests like the classic Thatcher illusion (Thompson 1980) which rather strikingly demonstrates our insensitivity to spatial relationship in inverted faces. Here, the eyes and mouth in a photo depicting Margaret

Thatcher's smiling face were cut out and turned upside down, and when the result is viewed with the entire face in its usual orientation the face appears to have a grotesque expression, but shown upside-down it is difficult to see that there is anything at all abnormal about the face. Young and Hay (1986) provided another piece of experimental evidence that we do not process features independently from each other, using composite faces created by adjoining the top half of one famous individual's face with the bottom half of another's. Here people were quite good in identifying isolated top or bottom halves seen on their own, but when joined together it became more and more difficult the more the halves became well-aligned. These and other experiments provide evidence that suggests quite strongly that the visual system does not store a face just in terms of its individual features, but rather as a more general, holistic configuration where spatial inter-relations play a crucial role in recognition.

Facial representation invariance

Recognition can be seriously impaired by changes in view, lighting, size and other aspects of the perceived face. Does this mean that regardless of the proven ability and effectiveness of our visual system to identify familiar objects when shown in different positions or orientations, the representation of faces is not invariant to such changes?

The psychophysical evidence seems to suggest that indeed a viewpoint or lighting invariant representation of faces is not present in our visual system. In the case of view, Hill, Schyns et al. (1997) demonstrated how subjects performed poorly in a recognition task where viewing conditions were altered.

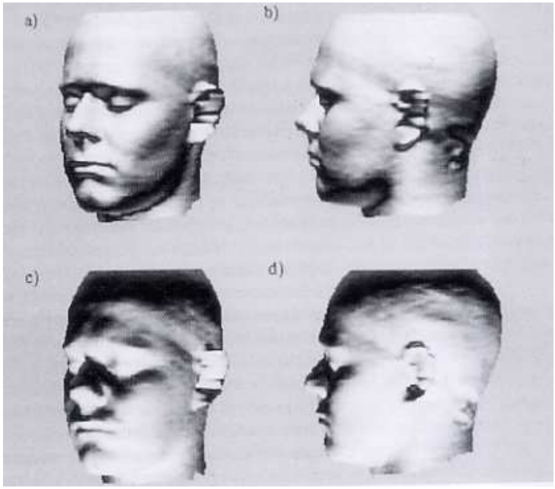


Figure 1.4 - Even in same viewpoint, it's hard to tell it's the same person due to lighting changes! (taken from Hill and Bruce 1997)

Although all views were equally well recognised when they all had been learned, they were shown to be surprisingly poor at generalising to novel views when given a single view of a face, with performance decreasing as the difference in viewing angle increased. Further studies seem to confirm this viewpoint dependence in the high-level encoding of facial identity (Benton et al 2006; Fang & He 2005), but others warn that face recognition reaction time and accuracy costs that are attributed purely to viewpoint changes could also be affected by the information that is typically unavailable in the experimental stimuli (normally 2D), rather than being solely a result of the underlying neural representation of facial identity (Burke et al 2007).

A similar impairment in recognition was recorded with recognition tasks under different lighting conditions. In same-or-different comparison tasks with pairs of laser-scanned heads presented from varying views and under varying lighting conditions, Hill and Bruce found that variations in lighting posed difficulties as great as variations in view (Hill & Bruce 1996 - see figure 1.4). An advantage for illumination from above was found, with better performance in a matching task under this condition, again most likely due to higher exposure to the illumination from above lighting condition.

Intriguingly, by illuminating faces from below and then inverting, Johnston et al. (1990) showed that the face inversion effect could be significantly reduced, a reduction also noted with contrast-negated faces, when lit from below (Liu et al 1999). And while this may point to some kind of surface based code for faces, the hollow-face experiments by Hill and Bruce (1993; 1994) show that familiarity with the three-dimensional structure of the face still seems to play some part in the process. Recently some experiments looked at size invariance as well. Lee et al. (2006) show that size changes up to four-fold had no effect on face discrimination and recognition, while viewpoint changes were again confirmed to be detrimental to recognition.

Neurophysiological investigations of the macaque brain have uncovered cells in the superior temporal sulcus tuned to specific facial orientations, particularly full-face and profile (Perret et al 1991; Perret et al 1985) and also such view dependent cells were found to be lighting and position invariant (Hietanen et al 1992). Added to the fact that Hasselmo (1989) a had already found cells that respond to all views of a face, we seem to have here plenty of indications that the brain has a two-dimensional image-based storage scheme for faces with a collection of views encoded separately in order to attain recognition from a variety of viewpoints (Wallis & Bulthoff 1999). Psychophysical evidence supports this hypothesis with demonstrations that, having learned two views of an object, subjects perform better when tested on views between them, rather than outside (Tarr & Pinker 1989). This can be explained in the view-based context by considering interpolated views to partially excite cells responsive to both learned views, whilst extrapolated views partially excite cells responsive to only one of the learned views. Wallis and Bülthoff propose that these invariant representations, based on individual views, can be learned by "experience through temporal coupling as well as physical similarity of views". They also conducted another type of experiment (Wallis & Bulthoff 2001) where viewing position and identity of a face were simultaneously altered, and here subjects treated the views as though they were of the same person.

Variance of features and motion of the face

We can ask, what are the principal sources of variance in the human face? To find this out psychologists often conduct experiments in which different sources of information are systematically concealed or enhanced, or in which different cues are put into conflict with one another. One such experiment was conducted by Fraser and Parker (1986), where by randomly flashing up individual features that made up a composite face and testing subjects on their ability to detect which was missing, they found that the most salient feature was the outline of the face, followed by eyes, mouth, then nose. Also, Shepherd et al. (1981) had found the principal sources of variation in a set of faces to be hairstyle, face shape and age.

To encode such changes in an efficient manner, a parameterised model of the face is needed, with factors related directly to the principal variations in faces. But the evidence above shows that, when not testing specific features, principal sources of variation appear to be more subtle and global, such as face shape and age. Thus it seems a better solution to consider a representation in which the face varies in terms of pseudo-features that affect the configuration of the face as a whole. And by considering faces to be parameterised by a set of features, regardless of their local or global nature, a *face space* can be created in which dimensions are composed of the parameters and the average of all faces lies at the centre. Leopold et al. (2001) recently demonstrated powerful after-effects in the context of the face-space paradigm. By adapting subjects to a particular face, they showed how recognition tasks for faces situated along that identity vector in face space were facilitated, whilst recognition was impaired for other faces. They also attempted to throw some light on the neural principles of encoding of face spaces by testing two different models (example-based and mean or norm-referenced models) using electrophysiological data from macaque area IT. They found that the majority of IT neurons might represent deviations from a norm or mean face, which is determined by an average over the distribution of typically occurring faces (Giese & Leopold 2004; Leopold et al 2006).

Regarding motion, while initially it was thought that motion in sequences doesn't improve recognition when compared to stills, it was shown using point light sources that certain objects can be recognised from their motion only (Johansson 1973). Bassili then found that naïve subjects were able to recognise the sequences (generated by filming blacked out faces and teeth with makeup and scattered white circular labels over the surface) as faces from the movement of the point light sources alone, leading him to postulate that facial motion was sufficient information for the recognition of an object as a face (Bassili 1978; 1979), with subjects even recognising emotions! Bruce and Valentine used this technique to investigate whether individuals could be recognised only on the basis of their facial motion (Bruce & Valentine 1988), and they found that above chance results were achieved by subjects in recognition of emotions and of individuals from a small set, but the performance was still very poor. However, with point light displays however a lot of the motion is lost, so Knight and Johnston created the stimuli by degrading image sequences through photographic negation, arguing that this maintains the full motion field, and their subjects did indeed find famous faces in moving sequences of negated images significantly easier to recognise than in stills, reinforcing the hypothesis that motion cues do provide useful information in face processing tasks (Knight & Johnston 1997). Furthermore, the view that a dynamic sequence simply provides more views of the face thus improving recognition was discredited by comparing performance on similarly degraded dynamic sequences to performance on the same frames simultaneously presented (Lander et al 1999). Here too subjects did better in recognising famous faces from moving sequences. Motion cues were also successfully used to test sex and identity judgements. When an androgynous 3D face was animated with facial movements of actors (Hill & Johnston 2001), subjects were able to successfully discriminate sex and identity. Motion is indeed shown to have an important role in facial recognition and categorisation by later studies by O'Toole et al. (2002) and Knappmeyer et al. (2003).

Chapter 1 summary

The human face is an astonishing organ of communication, able to transmit a wide spectrum of social signals and messages. Its structure is specialised and the signals it conveys can contribute to perceptions of the bearer's facial identity, expression, sex and more. The perception of these signals is achieved by a system with multiplicity of components processing separate signals, and there is ample evidence to suggest that both the component processes and the location of early processes of face recognition (FFA) are face-specific. Psychophysical and neuropsychological experiments also seem to suggest that while object recognition is analytic and part-based, face perception and recognition is holistic and configural. Looking at other psychophysical evidence, it appears that a viewpoint or lighting invariant representation of faces is not present in our visual system (although possibly size-invariant representations are), and that most importantly, we seem to store faces in a two-dimensional manner. Thus a two-dimensional, image-based approach could be very effective in encoding facial identity, expressions and sex. Candidate methods for computer-generating image-based models of faces, with movement dimensions extracted from the experience of the face in motion will be discussed in the next chapter of this thesis.

Chapter 2

COMPUTATIONAL MODELLING OF THE FACE

In the previous chapter, faces were shown to be complex, multidimensional, and informative objects capable of large deformations. As a result, in order to build a good quality computational model of the face it is important to choose a representation of the face that is powerful enough to realistically reproduce the full range of variation and movement that faces display, and generate facial animation capable of fooling the top expert system in face processing – our own visual system.

Three-dimensional representations

Artificial facial animation is mainly viewed from some two-dimensional projecting or reflecting surface, like a screen of some kind. However, the face is really a 3D structure, so some approaches have involved representing it in 3D. The scene is later projected from 3D to a 2D image for viewing, a process commonly known as *rendering*, by setting up virtual light sources and a viewpoint and tracing the path of light to that viewpoint.

Polygonal mesh representations of faces

The simplest 3D approach is probably a representation of the face surface by a set of polygons, usually triangles, connected at each vertex. This surface is known as a polygonal mesh. The use of flat polygons to represent smoothly varying surfaces inevitably leads to errors, which can be made arbitrarily small by increasing the number of polygons, at the cost of increasing the rendering time and storage requirements. Most graphics cards, however, now have fast, efficient, inbuilt polygon rendering routines especially for this purpose.

Polygonal models of the head for face animation purposes were introduced by Parke (1972). His polygonal meshes were derived rather crudely by first hand-painting a mesh on one side of a subject's face which was then photographed from frontal and profile viewpoints. Vertex coordinates were measured in 2D, the 3D co-ordinates were geometrically recovered, the mesh was then constructed and the faces of polygons were coloured. Later, Williams used a laser to scan a plaster cast of a human model's head (Williams 1990). The scanned data was in cylindrical co-ordinates. Photographs were taken of the model's head and painstakingly aligned and registered with the scanned data to map onto the computer head, while today it's possible to scan real heads with custom-made laser scanners, such as those produced by Cyberware^{TM} (www.cyberware.com), and simultaneously capture the localised texture map as the scanner rotates around the head, eliminating the need for the time-consuming alignment stage (see fig.2.1).

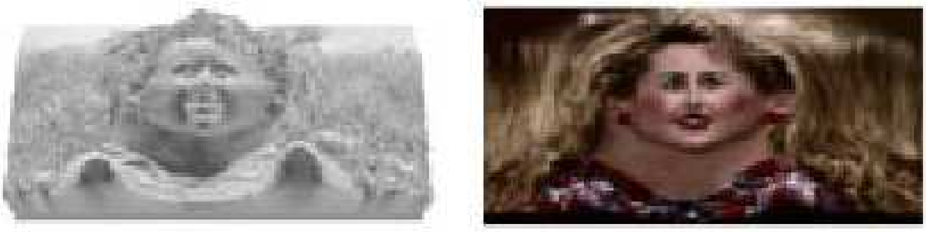


Figure 2.1 - Range data (left) and texture map (right) obtained with a Cyberware^{TM} laser scanner

Other surface representations include *implicit surfaces*, defined by a single equation, $f(x, y, z) = 0$. Any point satisfying that equation will be on the surface. A simple example of this would be a sphere of radius *r*, centred at *(a,b,c)* .

$$f(x,y,z) = (x-a)^2 + (y-b)^2 + (z-c)^2 - r^2 = 0$$

While useful with simple shapes, these equations become unwieldy for complex surfaces, such as the face, and require much more processing, hence such representations are not such a good choice for modelling of the head. Parametric surfaces are similar to implicit surfaces, but are defined instead by three functions of two parametric variables, typically based on cubic equations, with one for each spatial dimension, x, y and z (Forsey 1990). Parametric surface patches are much more efficient for approximating a curved surface than polygons, with far fewer needed to satisfy a particular error threshold and they do not suffer from polygonal edge effects. However, they are much more computationally expensive to process than the simpler polygon.

Volumetric & muscle-modelling representations

3D objects can also be represented by a volumetric approach, by combining building block primitives such as spheres, cylinders, cuboids, etc. These primitives can be deformed and merged to build elaborate three-dimensional structures, a process known as *constructive solid geometry* (CSG). Often CSG presents a model or surface that appears visually complex, but is actually little more than a set of cleverly combined or de-combined objects (fig.2.2). While this is perfectly acceptable for simple face models, difficulties arise when trying to model the detail required for realistic faces.

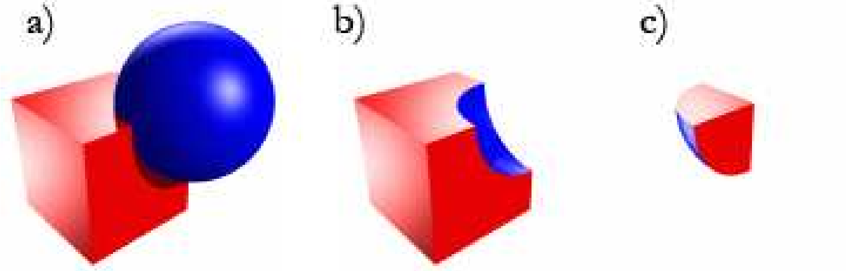


Figure 2.2 - Primitives can be combined into compound objects using set operations, like in examples a) boolean union, b) boolean difference or c) boolean intersection (image taken from Wikipedia)

But with aspirations to more accurately model the control of the face, research moved on to modelling facial muscles. This is now a common approach in facial animation and improves on Parke's first parameterised head model by simulating muscle actions rather than hard-wiring performable actions. Since muscle models are still parameterised, they can easily be controlled by adjusting a small number of parameters, and movement can be restricted to reasonable muscle actuations. The first such model was introduced by Platt and Badler (1981), where muscles acted as simple springs.

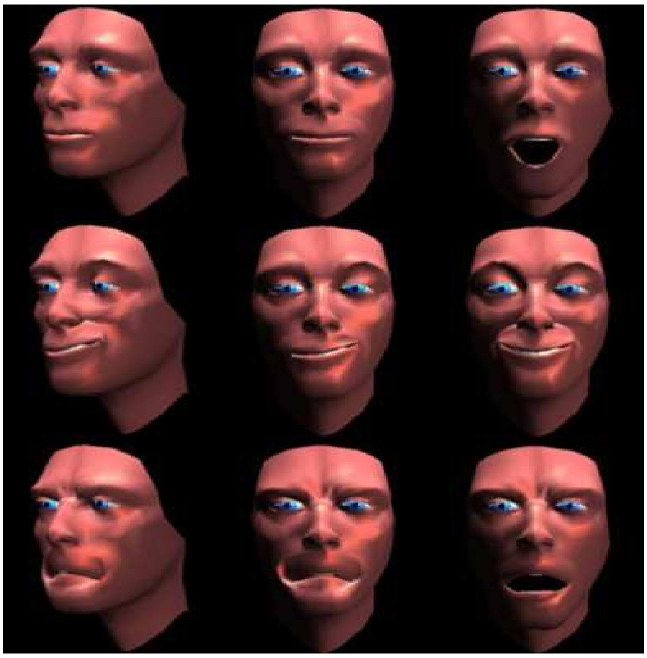


Figure 2.3 - Terzopoulos and Waters' model with dermal tissue analogue (taken from Terzopoulos & Waters 1993)

Waters further developed the muscle model (Waters 1987) by using a simplified model from research on facial muscles, based on the FACS system (Ekman & Friesen 1978). This scheme for coding facial movement describes movement in terms of 50 specified action units (AU's), each representing a muscle or a small group of muscles. Waters modelled ten of these to control a polygonal model, moving a particular muscle by moving its nodes of attachment maximally and the neighbouring points with diminishing strength as distance from the node increased. Terzopoulos and Waters enhanced Water's original model with human facial tissue modelled as a deformable lattice of point masses connected with biphasic elastic springs (Terzopoulos & Waters 1993). In analogy to real dermal tissue (fig.2.3), the biphasic springs allowed the synthetic surface tissue to initially readily extend under low strain up to some threshold (1st phase), and then exert rapidly increasing restoring forces beyond this (2nd phase).

Advanced 3D morphable face models

However, models such as that of Terzopoulos and Waters, where the parameters are highly subjective and uncorrelated, achieving photo-realistic face synthesis or animation is a hard task that requires substantial expert human intervention and with very limited results to show for it. Using context-free parameters derived by statistical modelling methods is one approach that began to emerge as a good candidate for achieving photo-realism in 3D face modelling.

Pighin et al. (1998) developed a system that allowed manual specification of correspondences across multiple images and then use vision techniques to compute 3D reconstructions. A 3D polygonal mesh model is then fitted to the reconstructed 3D points. The face models were highly realistic but also required a manually intensive procedure for their production.

Roy-Chowdhury and Chellappa (2003) introduced a technique of 3D reconstruction from short monocular sequences taking into account the statistical errors in reconstruction algorithms. They use stochastic information to fuse incomplete information from multiple views and this technique was applied to various applications including face modelling.

Blanz and Vetter (1999) also used a context-free parameter approach and this time came up with an automated technique for the synthesis of photo-realistic 3D faces. They use a linear combination of 200 3D face scans extracted via the previously mentioned Cyberware$^{\text{TM}}$ laser scanner to build a morphable face model. They first vectorised the geometry of the 3D face data into a shape-vector $\mathbf{S} = (X_1, Y_1, Z_1, ..., X_n, Y_n, Z_n)^T \in \text{¡}^{3n}$ containing all the *X, Y,* and *Z* coordinates of its *n* vertices. Similarly they represented the texture of the face by a texture vector $\mathbf{T} = (R_1, G_1, B_1, ..., R_n, G_n, B_n)^T \in \text{¡}^{3n}$ containing all the *R, G* and *B* colour values of the corresponding *n* vertices. An average of shape and texture of the face ($\bar{\mathbf{S}}$and $\bar{\mathbf{T}}$) was calculated and the main modes of variation in the dataset were computed as parameters using well-known technique for data compression called *Principal components analysis* (PCA). This gave an orthogonal coordinate system formed by the eigenvectors $\mathbf{S}_i$ and $\mathbf{T}_i$ for both the shape and texture models ($\mathbf{S}_{model}$ and $\mathbf{T}_{model}$):

$$\mathbf{S}_{model} = \bar{\mathbf{S}} + \sum_{i=1}^{m-1} \alpha_i s_i \text{ and } \mathbf{T}_{model} = \bar{\mathbf{T}} + \sum_{i=1}^{m-1} \beta_i t_i$$

where $\boldsymbol{\alpha}, \boldsymbol{\beta} \in \text{¡}^{m-1}$ represent the respective coefficients for the shape and texture eigenvectors or principal components (PCs).

They then came up with an algorithm that adjusts the weights of the linearly combined PCs for an optimal reconstruction of a new face either from a 2D image or a new 3D exemplar, with minimal manual initialisation required. To avoid generating an unlikely face a probability distribution is imposed on linear combination results. The reconstructions using their model are very realistic, almost reaching the quality of the laser scans themselves (fig 2.4).

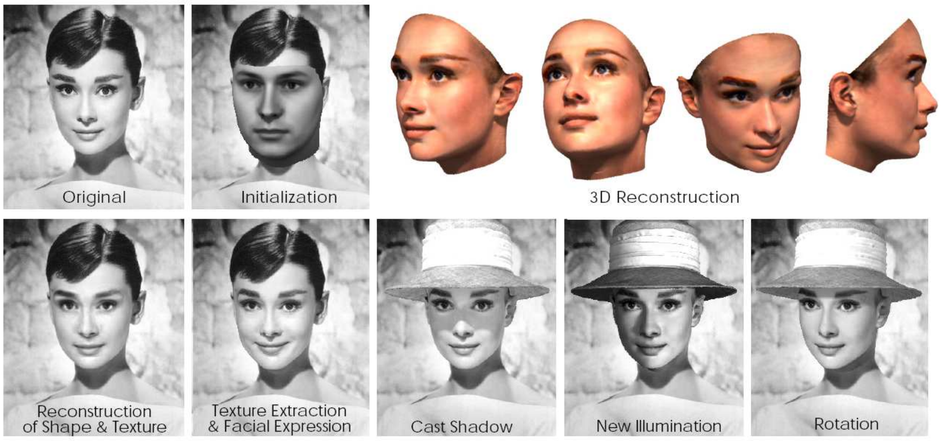


Figure 2.5 – After manual initialization, the algorithm automatically matches a coloured morphable to the 2D image of Audrey Hepburn. Rendering the inner part of the 3D face on top of the image, new shadows, facial expressions and poses can be generated using the model (taken from Blanz and Vetter 1999).

Apart from single pose generation of a 3D face from another 2D or 3D example face, Blanz and Vetter extended and used their model in a variety of other tasks, such as face recognition across poses and illuminations by fitting the above 3D morphable model (Blanz et al 2005; Blanz et al 2002; Blanz & Vetter 2003) and even photo-realistic animation that can be applied to any face shown in a single image or a video (Blanz et al 2003). The methodology of the 2003 paper provided the theoretical basis for the image-based performance driven mimicry model used in this thesis, which will be described in more detail in the next chapter.

The PCA representation of shape and texture information forms the basis of another generative model used for faces, the *Active Appearance Model* (AAM). It uses another type of iterative algorithm to learn the relationship between a training image and a synthesised model example and then generate an approximation within the model (Cootes et al 1998). It is worth mentioning that Cootes et al. face images marked at key points to outline the main features and achieve a good correspondence between the subject faces while Blanz et al. used optic flow techniques to extract a dense flow field matching the faces. The original AAM method was later extended into a 3D version (Jing et al 2004).

A two-dimensional facial representation

What all these high-res polygonal mesh surface and other 3D models have in common is that they all have to be generated either with the help of extremely skilled artists or with expensive equipment such as laser scanners, and the sheer number of vertices makes them effectively uncontrollable, needing complex hard-coded underlying muscle models to constrain and parameterise movements.

It seems that a more practical representation would be one that takes into account the evidence described in the previous section, which suggests that our visual system seems to process faces in a two-dimensional manner. Thus a more global, configural approach is proposed, one that chooses the parameters to account for essential sources of variance. *Principal components analysis* (PCA) provides a means of implementation.

PCA and eigenfaces

PCA is a mathematical technique, a vector space transform, used to reduce multidimensional data sets to a lower dimensional space with axes chosen to maximally describe the variance in the set. Depending on the field of application, it is also named the *discrete Karhunen-Loève transform*, the *Hotelling transform* or *proper orthogonal decomposition* (POD).

PCA was invented in 1901 by Karl Pearson (Pearson 1901). It involves the calculation of the *eigenvalue decomposition* of a data covariance matrix or *singular value decomposition* of a data matrix, usually after mean-centring the data for each attribute. The results of PCA are usually discussed in terms of component scores and coefficients, or weights. The data is thus transformed into a new coordinate system such that the greatest variance by any projection of the data comes to lie on the first coordinate (called the 1st principal component), the second greatest variance on the second coordinate, and so on. Lower order components can be discarded as noise, thus reducing the original dimensionality of the data. PCA is theoretically the optimum transform for a given dataset in least square terms and will be discussed in more detail in chapter 3.

Biological motivations for using PCA

After converting two-dimensional arrays of facial images into long vectors, Sirovich and Kirby showed that PCA could be used to extract the principal components of this face set, so-called *eigenfaces* (Sirovich & Kirby 1987). PCA effectively reduces the amount of information that needs to be stored in order to recognise individuals, since only one weight for each principal component is needed. The result is a representation parameterised in terms of the largest sources of variance (figs. 2.5 & 2.6).

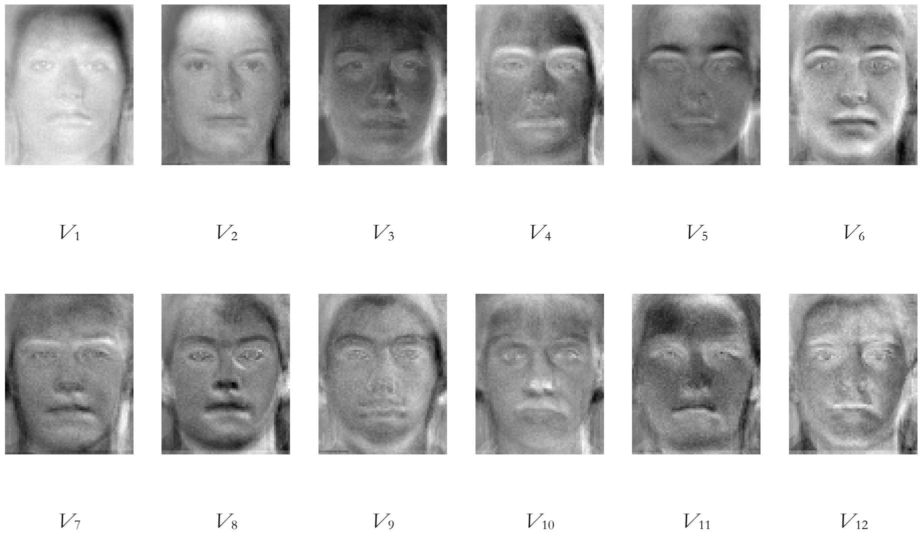


Figure 2.5 – The first 12 principal components (eigenfaces) extracted from a database set of 80 faces. Any face from that set can then be reconstructed (with some residual error) as a linear combination of the first N eigenfaces, with increasing precision for larger N.

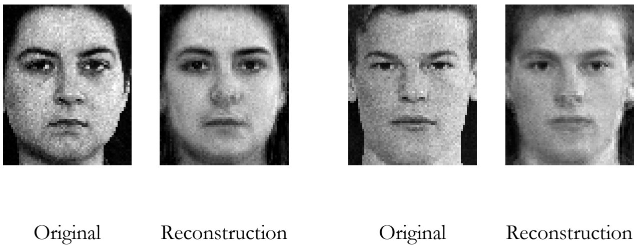


Figure 2.6 – Two novel faces constructed with 80 eigenfaces. For large face databases, reasonable reconstructions can be made of faces outside of the training set

Recognition experiments with eigenfaces show impressive results for faces captured under the same conditions as the training set, but slight variations in lighting, orientation, scale and position quickly degrade performance, although this is consistent with some psychophysical results discussed earlier.

The principal components of natural image patches have already been shown to closely resemble receptive fields of cells in the visual cortex (Hancock et al 1992) and it seems reasonable that cells should be tuned to the natural dimensions of variation inherent in the input concerned. The neural mechanisms behind the encoding of facial identity have previously been modelled with PCA on static images of faces and encouraging parallels have been found. O'Toole et al. (1994) found that how well a face could be reconstructed using eigenfaces could predict how memorable it was for human subjects. They also found that eigenfaces were much less efficient in the encoding of faces of race not contained within the generative database, mirroring the other-race effect (Valentine & Bruce 1986). Giese et al. (Giese & Leopold 2004; Giese et al 2004) presented further support for a prototype-based encoding like PCA from electrophysiological studies in primate visual cortex by exploiting a morphable 3D model of the face (Blanz & Vetter 1999). Neurones increased their firing rate to caricatured faces as a function of distance from the average face in the model space.

PCA has almost entirely been applied for encoding identity in faces, rather than facial movements. A rare exception is the work by Calder et al. (2001) in analysis of facial expressions. Principal components analysis was applied to Ekman and Friesen's face database (Ekman & Friesen 1976), containing a variety of people, demonstrating a variety of expressions. They found that their PCA-based system was capable of supporting facial expression recognition and noted a natural separation of identity and expression, with components tending to code for either just expression, or just identity.

Motion capture and facial animation

With an image-based model that uses PCA, the problem of hard-wiring complex muscle structures and actions is evaded. Furthermore, if a model is to be used interactively, it should include mechanisms that allow the user to manipulate or animate the faces described by it. So given a computer model of a face, a procedure is required for animating it. However, if realistic facial motion is required, where better to get it than from a true face. Such performance-driven animation requires tracking an actor's movements and relating those movements to the model.

Optic flow

Motion capture is done using a variety of techniques. Dot, contour and feature tracking can be used to capture movement at a small set of locations on the face but all require manual registration of these points within the model, using some type of marker or highlighting. *Optic flow* techniques, on the other hand, require no markers or highlighting.

In psychology, optic flow is referred to as the retinal velocity field induced by a moving observer (Marr 1982). A more precise definition describes it as apparent motion of local regions of the image brightness pattern from one frame to the next while preserving intensity patterns during frame-to-frame transitions (Simoncelli 1993). Optic flow algorithms (Barron et al 1992) provide estimates of speed and direction for locations in a frame of an image sequence. Mase (Mase 1991; Mase & Pentland 1991) was the first to introduce the method of tracking action units using optical flow. In this work no physical model is employed but the face motion is formulated statically rather than formulated within a dynamic optimal estimation framework. However, the results of this work confirmed the validity of optic flow computation for observing facial motion.

One such algorithm is the *Multi-channel Gradient Model* (McGM), modelled on the processing of the human visual system (Johnston et al 1999). This algorithm calculates a basis set of spatio-temporal derivatives by convolving the image sequence with derivative of Gaussian filters, and then combines them to form derivatives of the Taylor expansion in space and time. Ratios of the resulting terms then yield robust estimates of image motion for every pixel of every frame.

2D facial animation

Even when the movements have been successfully extracted using an optic flow algorithm, the process of animation is still non-trivial. While 3D models are still unable to produce animations that can deceive a human observer into believing that they are real faces, 2D animation techniques seem to fare better in this respect.

As an illustration of so-called *example-based modelling*, there's the work by Beymer, Shashua and Poggio where they demonstrated how novel views of objects varying rigidly and non-rigidly can be generated from an image-based model by interpolating between example images registered by application of an optic flow algorithm (Beymer et al 1993). Fidaleo and Neumann (2002) generated an example based virtual puppet from images. The face was first split into a small set of local regions (co-articulated regions or CRs) which represent small groups of facial muscles. Then a set of basic facial movements that activated movements independently in each

CR was chosen. Sequences were recorded in which an actor performed each of these movements individually, with markers attached to an empty frame in order to warp the sequences onto a standard position. Muscle actuations were separated using *Independent Component Analysis* (ICA), which maximises independence between components rather than variance, as in PCA (Bell & Sejnowski 1995). The face could then be parameterised and new footage of the same actor could then be analyzed in real-time to extract the high-level parameters, which could then be used to drive that model, or any other model handcrafted under that same parameterisation. This effectively enables photo-realistic performance driven facial animation, but it is limited to the one actor.

Such a limitation is removed if the example-based generation of puppets or avatars is done by way of PCA. All that is required is an example sequence of the target face in motion, where each frame of the sequence is considered to be an example configuration that can be provided in any vectorised format.

Chapter 2 summary

Building a good computational model of the face is not an easy business by any means. Due to the fact that the face is essentially a 3D structure, a number of 3D facial representations have been used in realistic head and face modelling, like polygonal mesh, implicit, and parametric representations of the surface of the face. Volume representations were also employed, together with the more complex dynamic representations of facial muscles. However, the greatest drawback of this class of models is that they require huge resources, be it in computational or human terms, to achieve any sufficiently compelling results to fool the human visual system. By considering the internal processes involved in the perception of faces, a more practical class of representations is introduced that takes into account the evidence which suggests that our visual system seems to process faces in a two-dimensional manner. A more global, configural approach is proposed, one that chooses the parameters to account for essential sources of variance, i.e. PCA. By reviewing this statistical modelling approach and the biological motivations for its use in this context, this section completes the formal literature review element of this report and leads onto the empirical study section in which PCA will be used to generate example-based avatars and use them in investigating problems such as the reconstruction of missing motion information due to facial occlusion.

Chapter 3

PCA-BASED PERFORMANCE-DRIVEN FACIAL MIMICRY

Example-based models of the face can be animated by projection of facial actions from another driving face (Cowe, 2003). All that is required is an example sequence of the target face in motion and an example sequence of the driving face in motion. Individual sequence frames represent example configurations and must be provided in a vectorised format; the better the quality of the vectorisation, the better the quality of the resulting model. Principal component analysis (PCA) is then applied to these example vectors in order to extract a smaller set of orthonormal vectors forming a basis that closely spans the set. A generative model of the target face can be produced, based on these principal components.

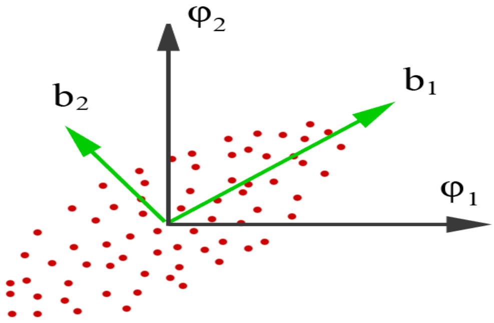


Figure 3.1 – In two dimensions, the set of principal components ($\mathbf{b}_1$, $\mathbf{b}_2$) accounting for more variance than the original set of variables (φ_1, φ_2)

Principal components for encoding facial actions

Principal component analysis is a mathematical technique that seeks to linearly transform a set of correlated N - dimensional variables, $\mathbf{\Phi} = \{\boldsymbol{\varphi}_1, \boldsymbol{\varphi}_2, ..., \boldsymbol{\varphi}_N\}$ (assumed without loss of generality to have zero mean), into an uncorrelated set that better describes the data, termed principal components or basis vectors, $\{\mathbf{b_1}, \mathbf{b_2}, \ldots, \mathbf{b}_N\}$ (fig.3.1). Any point **X** in our original dataset can then be described as a linear combination of these principal components $\mathbf{b}_i$, where *i = 1,…, N*. In general, most of the variation in **Φ** will be accounted for by *M* principal components, where $M = N$, thus giving:

$$\mathbf{X} \cong \sum_{i=1}^{M} c_i \mathbf{b}_i$$

It can be shown that these principal components, sequentially chosen to maximise the variance thus far accounted for, subject to the constraints of orthonormality, turn out simply to be the eigenvectors of the covariance matrix for the set $\{\boldsymbol{\varphi}_1, \boldsymbol{\varphi}_2, ..., \boldsymbol{\varphi}_N\}$ (Joliffe 1986).

Even though PCA has often been used in the encoding of identity, it has rarely been considered as a tool for encoding facial actions and motion. Some speech-related research, however, has involved the application of PCA to facial motion, with markers physically attached to the face and tracked while phonemes are uttered. The positional information of the dots over time was subjected to PCA as a means of dimensionality reduction for building codebooks relating acoustic data to mouth movements (Arslan & Talkin 1998; Kshirsagar et al 2001). This information is, however, sparse, and has not previously been used for the purpose of analysis.

PCA has also been applied to optic flow data around the mouth for extracting basis motion fields for motion recognition (Fleet et al 2000;

Yacoob & Black 1999). Their results were used to parameterise the movements of the mouth, but synthesis of facial motion was not a goal of the work and the resulting principal components were not discussed.

Calder et al. did apply PCA in the analysis of facial expression. They analyzed static images of faces posing a variety of expressions in order to acquire the statistical properties of the set (Calder et al 2001). The faces were taken from a database of photographs of several people performing several facial expressions. Landmarks were manually located on each picture, and all were warped onto the mean shape. PCA was applied to the shape and shape-free (texture) information. The methods involved were similar to that of the morph vectorisation discussed later in this chapter. However, Calder et al. calculated the principal components from a set of posed static images represented in the equivalent of the morph vectorisation, but these are not necessarily typical of natural experience of faces. The Cowe technique discussed here allows principal components to be obtained from natural sequences of facial motion.

Vectorisation of faces

The most basic vectorisation would probably be a list of the frame's grey-level pixel values (fig.3.2). An image of width w and height h can be considered to be an $h \times w$ matrix $\mathbf{X}$ of grey level intensity values - one value for each pixel of the image - where $\mathbf{X}_{ij}$ represents the value in the i^{th} row and j^{th} column. This can be converted into a vector, **x**, by simply concatenating the rows and transposing (Sirovich & Kirby 1987; Turk & Pentland 1991). This vector (of length $N = w \times h$) can be thought of as representing a point in an $N-$dimensional space. Now consider a set of P frames from a continuous recorded sequence of a face vectorised in this manner, $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_P$.

$$\rightarrow \begin{bmatrix} \mathbf{X}_{11} & \mathbf{X}_{12} & L & \mathbf{X}_{1w} \\ \mathbf{X}_{21} & \mathbf{X}_{22} & L & \mathbf{X}_{2w} \\ M & M & O & M \\ \mathbf{X}_{h1} & \mathbf{X}_{h2} & L & \mathbf{X}_{hw} \end{bmatrix} \rightarrow \left[\begin{array}{c} \mathbf{X}_{11} \\ \mathbf{X}_{12} \\ M \\ \mathbf{X}_{1w} \\ \mathbf{X}_{21} \\ \mathbf{X}_{22} \\ M \\ \mathbf{X}_{2w} \\ M \\ \mathbf{X}_{h1} \\ \mathbf{X}_{h2} \\ M \\ \mathbf{X}_{hw} \end{array}\right] \begin{array}{l} \left.\begin{array}{c} \\ \\ \\ \\ \end{array}\right] \text{1}^{\text{st}}\text{ row} \\ \left.\begin{array}{c} \\ \\ \\ \\ \end{array}\right] \text{2}^{\text{nd}}\text{ row} \\ \\ \left.\begin{array}{c} \\ \\ \\ \\ \end{array}\right] \text{row } h \end{array}$$

Image

$h \times w$ matrix, $\mathbf{X}$, of pixel intensity values

$N \times 1$ vector, $\mathbf{x}$, formed from matrix

Figure 3.2 – Vectorising an image by its pixel-wise intensity values (from Cowe 2003)

Since frames from a continuous recorded facial sequence tend to vary smoothly, these images will generally be clustered together in this space, centred approximately on their mean, $\boldsymbol{\mu} = \frac{1}{P}\sum_{i=1}^{P}\mathbf{x}_i$. Considering $\boldsymbol{\mu}$ as a reference, each face, $\mathbf{x}$, in the set can be considered as a linear translation, $\boldsymbol{\varphi}$, from this, $\boldsymbol{\varphi} = \mathbf{x} - \boldsymbol{\mu}$. RGB colour images are vectorised similarly, by simply concatenating the three colour planes (fig.3.3).

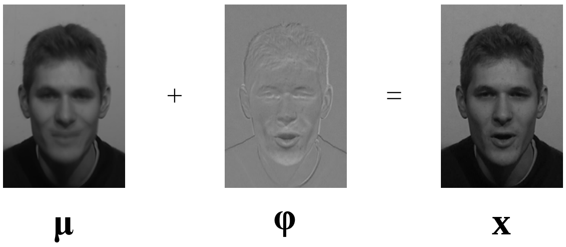


Figure 3.3 – Original frame $\mathbf{x}$ is broken down into its sequence mean $\boldsymbol{\mu}$ and a change vector $\boldsymbol{\varphi}$ (from Cowe 2003)

Simple linear combination of images results in an inherent blur, so a warping procedure is used in order to remove this blur. This procedure works by choosing an arbitrary frame to be a reference and defining all other frames in terms of warps from this single frame. A flow field relating each pixel in the target frame **T** to its source location in the reference **R** needs to be extracted in order to warp the reference to its target frame, and this is achieved using the McGM optic flow algorithm, modelled on the processing of the human visual system (Johnston et al 1999). Thus, all images in the sequence can be represented as warps from **R** and the entire sequence can be reconstructed by warping this one reference frame. Each vector field **[U,V]** - **U** and **V** are matrices containing the horizontal and vertical components of the field respectively, for each location (x, y) - can be vectorised, by concatenating each row of **U** and **V**, joining them and transposing to form one long vector. The whole sequence can thus be encoded by storing the one reference frame, **R**, and the vectorised flow field for each frame.

However, warping alone fails to capture iconic changes or lighting changes. These problems can be overcome by additionally encoding the image information for the target frame. This motivates a vectorisation based on *morphing*, a combination of warping and image blending. Specifically, by warping and simultaneously fading from the reference to the target, a better quality transition will result, with realistic synthesis of facial movement without blur, but without losing iconic or lighting changes (fig.3.4).

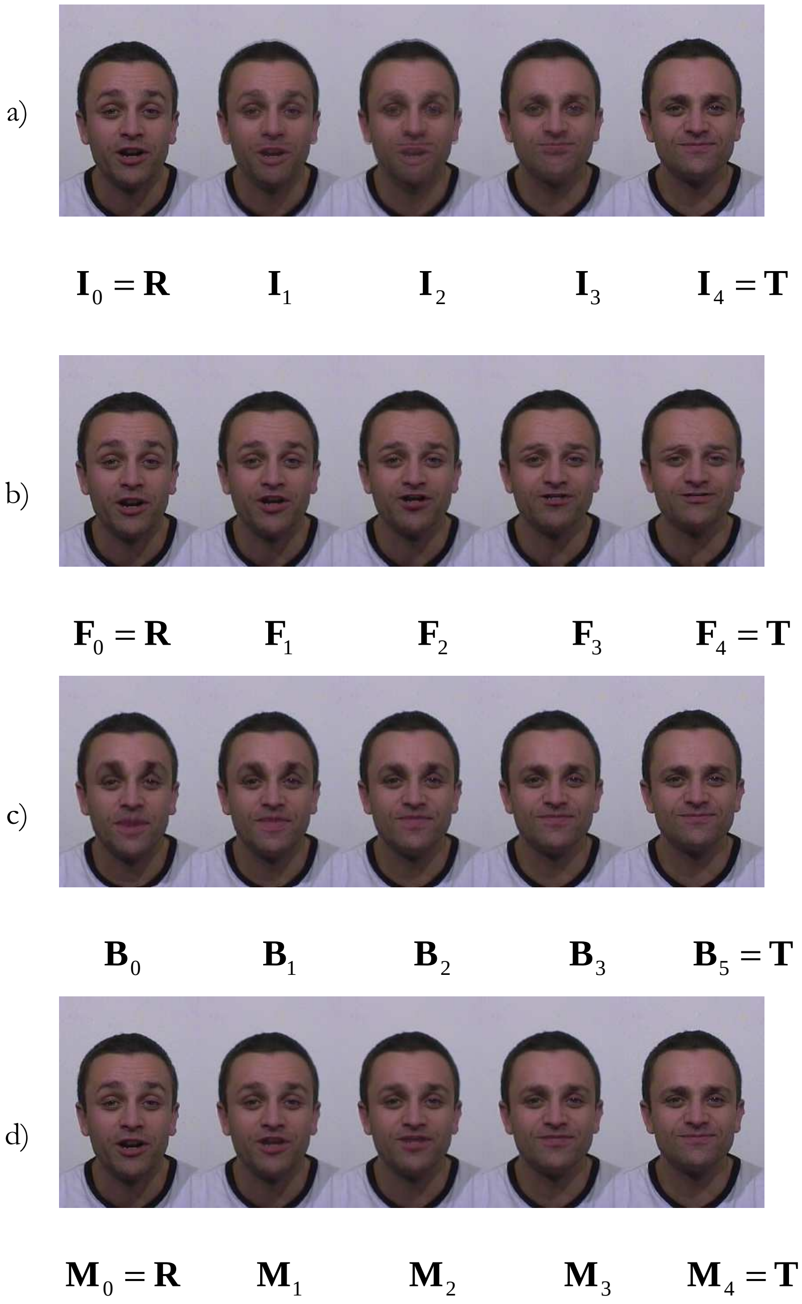


Figure 3.4 – Morphing: a) blending from image $\mathbf{R}$ to image $\mathbf{T}$ by weighted image addition; b) warping from $\mathbf{R}$ to $\mathbf{T}$ (left to right); c) warping from $\mathbf{T}$ to $\mathbf{R}$ (right to left); d) morphing – a combination of warping and blending

The frames are thus encoded by concatenating the shape information and texture information into one long vector (fig.3.5)

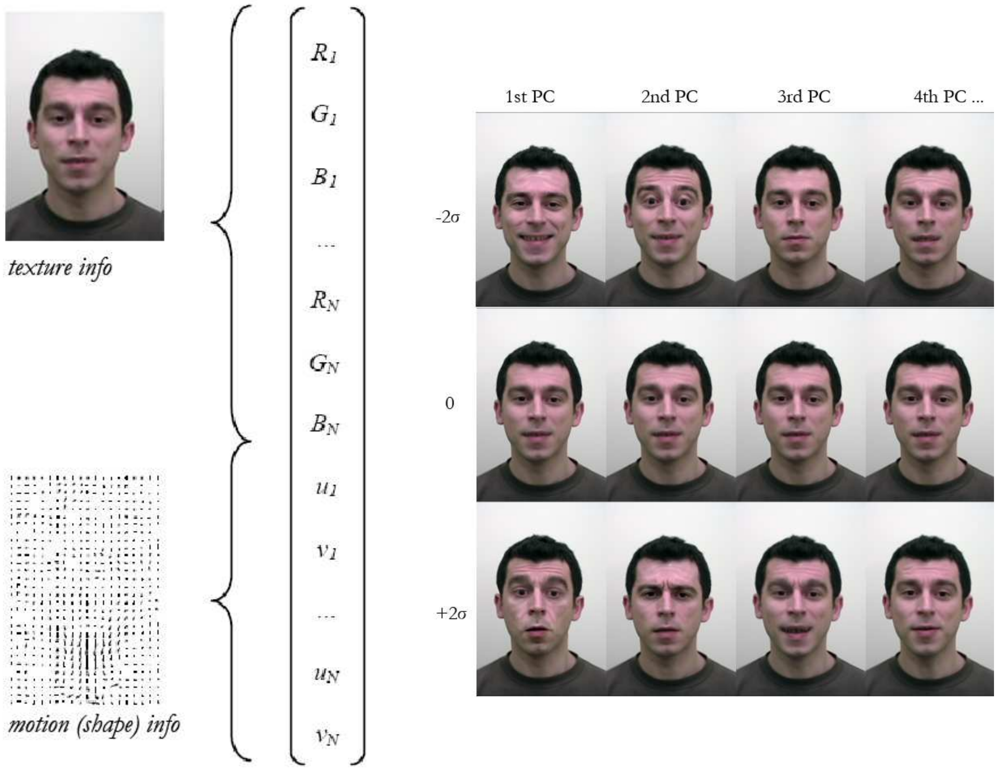


Figure 3.5 – Morph vector of face. On the right, 1st four PCs from a sequence of facial motion, vectorised as morphs. Top and bottom rows show the component morphs, -2 and +2 standard deviations from sequence mean (middle row)

By application of PCA on the covariance matrices of these face morph vectors, we can define a new improved orthonormal co-ordinate system centred on $\boldsymbol{\mu}$, which more efficiently spans this subspace, with axes chosen in order of descriptive importance, i.e. basis vectors are defined sequentially, each chosen to point in the direction of maximum variance, unaccounted for so far by their predecessors, due to the constraint of orthonormality. Since noise tends to be uncorrelated, vectors describing this will be of low importance in the hierarchy and can be later discarded by truncation to a lower dimensionality. With a generative model of the target face based on principal components constructed, facial movements performed by an actor can now be transferred onto the computer generated model.

Facial mimicry by projection

With this new co-ordinate system representing an individual's face space, any facial movement $\boldsymbol{\xi}$ from a sequence of any individual can be projected onto this basis, provided it's aligned & filmed from a similar view-point, vectorised in the same manner and centred on its own sequence mean.

Given a set of M_{train} training vectors from individual one (the face we wish to drive), $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{M_{train}}$, and a set of M_{drive} driving vectors from individual two (the face that will be doing the driving), $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_{M_{drive}}$, both sets are centred on their means and put into matrices $\boldsymbol{\Phi}$ and $\boldsymbol{\Psi}$, such that $\boldsymbol{\Phi} = \{\boldsymbol{\varphi}_1, \boldsymbol{\varphi}_2, \ldots, \boldsymbol{\varphi}_{M_{train}}\}$, where $\boldsymbol{\varphi}_i = \mathbf{x}_i - \boldsymbol{\mu}_{train}$, and $\boldsymbol{\Psi} = \{\boldsymbol{\psi}_1, \boldsymbol{\psi}_2, \ldots, \boldsymbol{\psi}_{M_{drive}}\}$, where $\boldsymbol{\Psi}_i = \mathbf{y}_i - \boldsymbol{\mu}_{drive}$. PCA extracts a set of basis vectors $\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_P$ from the training set, where $P \leq M_{train}$. To project into the new lower dimensional co-ordinate frame provided by the principal components, the basis transformation matrix $\mathbf{B} = \{\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_P\}$ is used, with the basis vector columns normalised to unit length. So, to project a N-dimensional vector, $\boldsymbol{\psi}_i$, into the P-dimensional subspace described by the principal components basis, apply:

$$\mathbf{c}_i = \mathbf{B}^T \boldsymbol{\psi}_i$$

Elements of $\mathbf{c}_i$ now represent weightings on the respective basis vector. In order to transform the projection, $\mathbf{c}_i$, back to N-dimensional space translated to the standard origin, the inverse transformation is applied and the training mean is added. In the case of principal component bases, the set

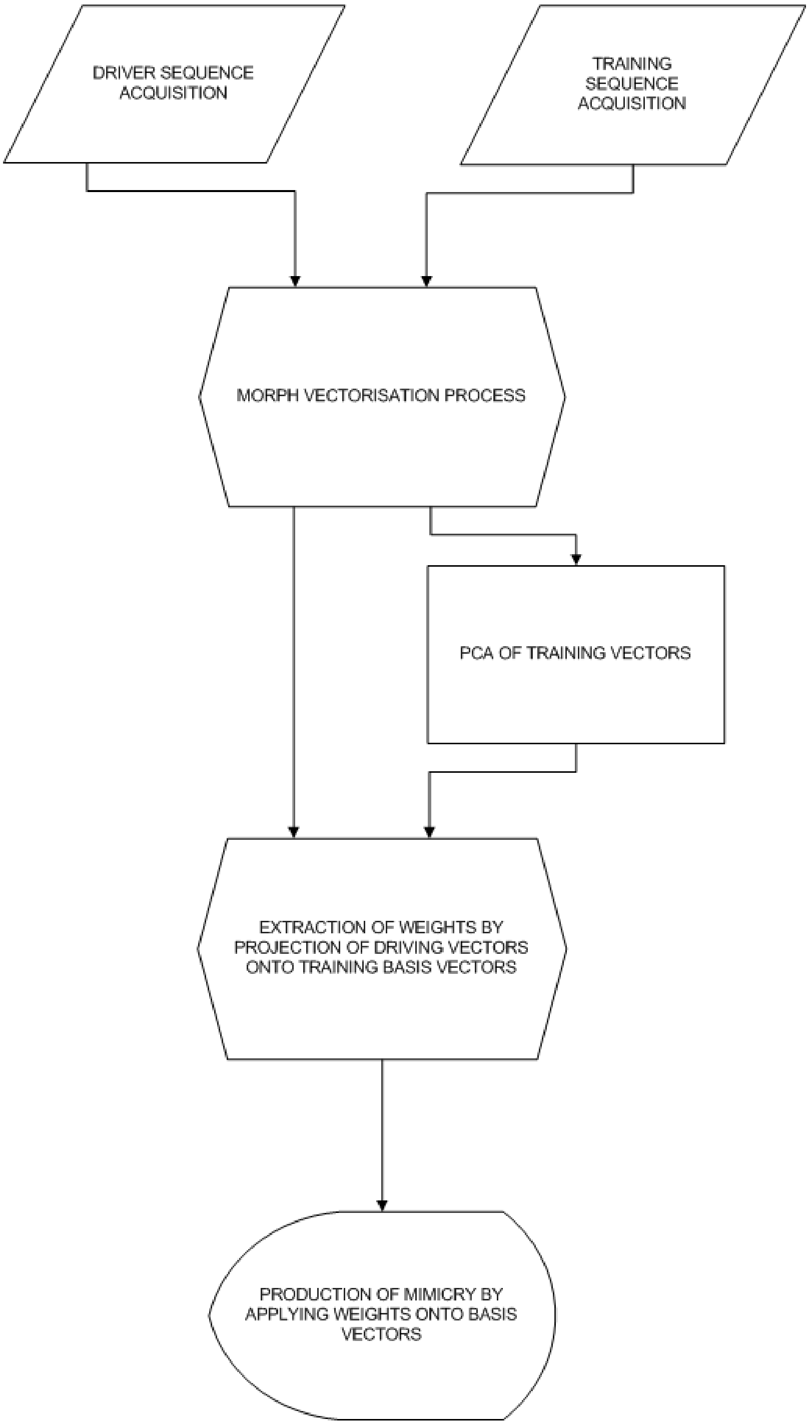

DRIVER SEQUENCE ACQUISITION
TRAINING SEQUENCE ACQUISITION
MORPH VECTORISATION PROCESS
PCA OF TRAINING VECTORS
EXTRACTION OF WEIGHTS BY PROJECTION OF DRIVING VECTORS ONTO TRAINING BASIS VECTORS
PRODUCTION OF MIMICRY BY APPLYING WEIGHTS ONTO BASIS VECTORS

Figure 3.6 – Block diagram of the mimicry generation process

is orthonormal, so $\mathbf{B}^T\mathbf{B} = \mathbf{I}$, which implies that $\mathbf{B}$ is the inverse transformation, thus

$$\mathbf{z}_i = \mathbf{B}\mathbf{c}_i + \boldsymbol{\mu}_{train}$$

In the case of the morph vectorisation, the new $N \times 5$ vector, $\mathbf{z}_i$, is then rearranged into h rows of w elements, to form a $w \times h$ image frame of the sequence of facial motion, and the sequence of those frames represents the result, an avatar driven by the facial actions of another.

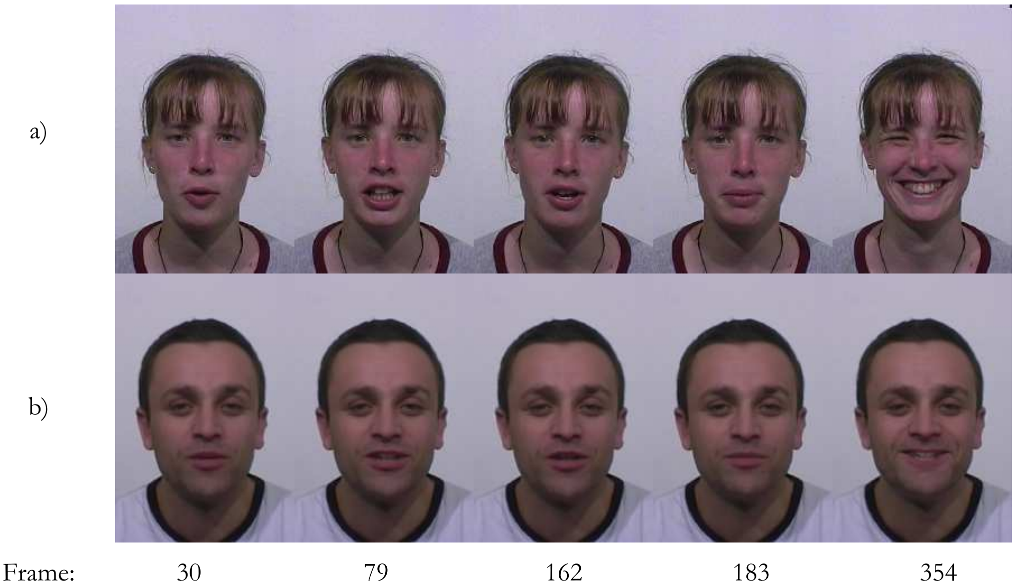


Figure 3.7 – Results from the mimicry with morph vectorisation: a) Selected frames from a sequence of facial motion; b) frames from a) are morph-vectorised, then projected into the male face space (from Cowe 2003)

Figure 3.7 demonstrates typical results from this process for the morph-vectorisation defined previously. A 15-dimensional face space was defined for face *a* since these first 15 components account for approximately 90% of the variance in the original sequence. Faces that are more expressive may require more principal components in order to capture the same amount of variance. The top frames in fig.3.6 are from a real image sequence of a person telling a joke which constituted the driving sequence. An affine transform was applied to all frames from the driving sequence. The two eyes and two mouth corners of the mean frames were used as references, in order to ensure that the centres of the eyes and edges of the mouth in face *a* were aligned with the respective points in face *b*. These frames were then vectorised in the same manner as face *b* and then projected onto face *b* space using the procedure defined at the beginning of this subsection. The resulting vectors were then transformed back to image space and are shown in the above figure, below their corresponding frames.

It should be mentioned that the generative phase of the process, in which the avatar is created, is computationally intensive due to the extraction of very dense optic flow information and principal components. For sequences of around 300 frames, 160 by 240 pixels, the morph vectorisation and PCA extraction can take up to half an hour on a 3.6GHz Pentium D processor. The EM-algorithm for PCA scales most favourably in complexity, scaling linearly both with number of components and dimensionality of the data (Roweis, 1998).

The driving stage, however, requires only the multiplication of a M ´ N matrix by an N - dimensional vector, followed by the multiplication of an N ´ M matrix by a M - dimensional vector (recalling that N is the dimensionality of the vectorisation and M is the number of basis vectors used). Matrix arithmetic can be calculated extremely fast on modern computers and even with the conversion of the resulting projection into a viewable image, driving can be comfortably achieved at frame rate (i.e. at under 40 ms), using the above vectorisation and configuration. Since features overlap well, the vectors of the driving face project strongly onto the target basis set. The procedure constrains movement, forcing it to be consistent with those movements which face *b* is capable of making. It also allows for further manipulation of the mimicries, such as exaggerating or rescaling of projected actions, by simply multiplying their respective mean-centred vectors by some factor, k, in order to magnify or reduce the departure from the mean. This last property of the model was proved to be fairly useful in enhancing the retrieval of occluded facial information, as described in ch. 5.

Chapter 3 summary

In chapter 3, a method for automatically creating computer-generated avatars was described. The model can be driven by real actors, simply by aligning features and projecting vectorised sequences of their motion into the target space. Novel footage can then be produced of the computer-generated avatar mimicking the actor's movements, all in reasonable amount of time despite quite high computational costs of the generating and driving procedures.

Resulting animations are confined to vary as a linear combination of movements from the example set, so the generated footage is realistic. This may seem to be a limitation, but is advantageous in preventing the avatar from doing anything that the original face was incapable of doing. Provided a sufficiently rich set of motion is captured for the generation of the model, these constraints do not pose a problem. The coefficients for a sequence can be transformed in the target face space, in order to exaggerate, or rescale movements to be consistent with the example footage. This is a useful processing step in conditions where the facial geometry is such that the novel vectors do not project strongly onto the target basis set, such as for example in the case of a driver with occluded facial regions.

Chapter 4

IDENTIFYING IMPORTANT REGIONS FOR HIGH FIDELITY PCA-BASED FACIAL MIMICRY

Most human observers perform very well in object classification tasks. Depending on the expertise of the observer, an object of interest can be classified in different levels of detail or generalization as a car, a vehicle or as some specific make of that car (Rosch et al 1976). Such a classification is achieved based on the same visual input, but intuitively it is clear that specific features within this visual input enable observers to succeed in performing such tasks. For instance, the badge and brand markings will enable an expert observer to place a car in a specific subordinate level category such as Ford or Renault. Other features could enable new categorizations and/or support existing ones. Similarly, visual information from different areas of the face does not appear to contribute equally to human observer's ability to process faces (Buchan et al 2007).

Diagnostic information for categorisation

In categorization tasks such as gender recognition and expression detection, subjects were shown to use different visual information from the same visual input (Gosselin & Schyns 2001). They went about revealing the diagnostic information for the above categorization tasks by introducing the so-called *bubbles* method. This method is now used for identification of diagnostic stimuli for a great variety of categorization tasks, such as infant perceptual categorization (Humphreys et al 2006), perception of ambiguous figures (Bonnar et al 2002), categorization of natural scenes (McCotter et al

2005), spatio-temporal dynamics of face recognition (Vinette et al 2004) and even pigeons' visual discrimination behaviour (Gibson et al 2005).

In the majority of these works, the stimuli consisted of static images of faces, the tested subjects were human or animal and the tasks were binary categorisation tasks, i.e. "is face male or female, expressive or non-expressive?"

Here, in order to produce dynamic facial stimuli, the model described in the previous chapter was used. Actor's facial movements were extracted and automatically projected onto another person's face model, or avatar, without any need for markers. The test subject is the computer system itself, with reproduction fidelity as a simple diagnostic criterion for comparison with a ground-truth mimicry. It essentially constitutes an ideal observer type set-up that can investigate whether any regions of a face in motion are more important for performance-driven, photo-realistic mimicry, generated using our computer model of the face.

During preliminary experiments, where performance-driven mimicries using driver faces with arbitrary rectangular occlusions placed over specific facial features were generated, it became clear by way of visual inspection that visual information from certain areas of the face was more important for hi-fidelity driving of our face model and overall recovery of actions from those occluded facial features. Here we try to locate these areas more precisely using a principled method such as bubbles.

The *bubbles* method

The original bubbles method introduced by Gosselin and Schyns (2001) involves partly occluding the facial stimuli using masks that are punctured by a number of randomly located Gaussian windows, or bubbles. Across trials, masks that revealed enough facial information for their human subjects to correctly categorise the occluded face were added up and divided by the sum of all masks, resulting in the so-called *ProportionPlane*. The averaged ProportionPlane is a measure of the relative importance of the image areas for the given task (fig. 4.1).

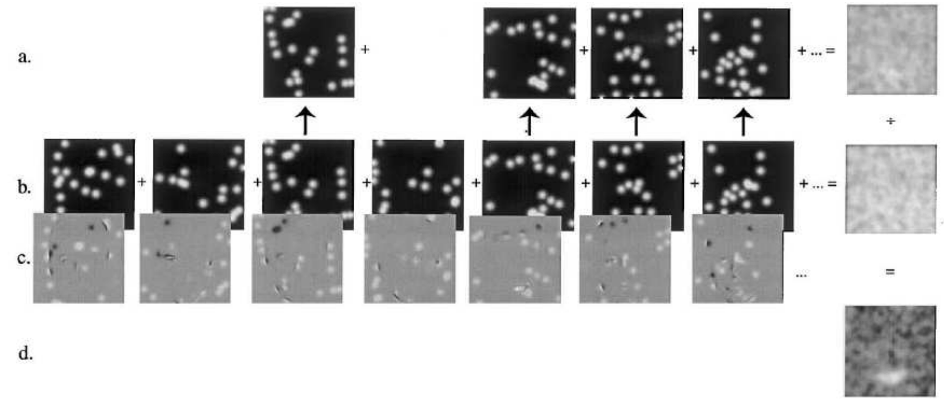


Figure 4.1 - This figure is taken from the original bubbles paper (Gosselin & Schyns 2001). Section (a) shows the bubbles leading to a correct categorisation added together to form the CorrectPlane (the rightmost greyscale picture). In (b), all bubbles (those leading to a correct and incorrect categorisations) are added to form TotalPlane (the rightmost greyscale picture). Section (c) shows examples of faces as revealed by the bubbles of (b). It is illustrative to judge whether each sparse stimulus is expressive or not. ProportionPlane (d) is the division of CorrectPlane with TotalPlane. Note the whiter mouth area found to be important for this categorisation (expressive or non-expressive) task.

Our face model, or avatar, was driven by instances of the same sequence (31 frames, 120x80px), processed in the same way, but occluded with 5000 random bubble-masks (fig. 4.2). These masks contained 23 bubbles each, with standard deviation of 5 pixels. A ground-truth mimicry was generated by driving the avatar with a non-occluded sequence.

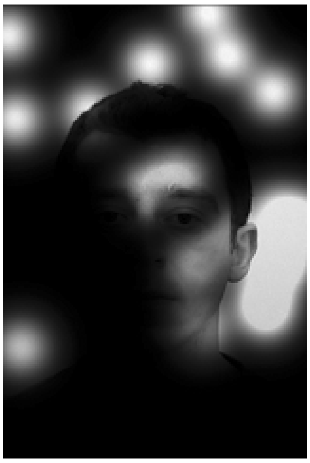


Figure 4.2 - Applied bubble-mask. This is one example of the 5000 random bubble-masks applied to the moving face driving sequence. The sequences were then processed and used to drive the avatar, resulting in 5000 mimicries.

Diagnostic criterion in bubble-occluded facial mimicry

The resulting occlusion-affected mimicries were compared to the ground-truth mimicry using a Pearson correlation metric. Initially, this metric was used to make image-based comparisons of actual mimicries, comparing RGB values between ground-truth and bubble-masked mimicry, frame by frame and pixel by pixel. This method failed to pick up on the comparatively subtle image changes in mimicry caused by the occlusions. It was presumably heavily biased by the inherent similarity of the images compared, i.e. all faces of the same person. Thus the resulting correlation values between the ground-truth and other mimicries were all in the interval between 0.988 and 0.99, and were not corresponding well with visual inspection results.

With this in mind we decided to measure the Pearson correlation between principal component weightings $\overset{\vee}{c}_i$ extracted from the 5000 occluded driver sequence vectors $\overset{\vee}{\psi}_i$, and those from the ground-truth. The correlation values obtained this way ranged between 0.16 and 0.91 in value (fig. 4.3). They corresponded very well with visual inspection results thus turning out to be a good metric for our categorization task.

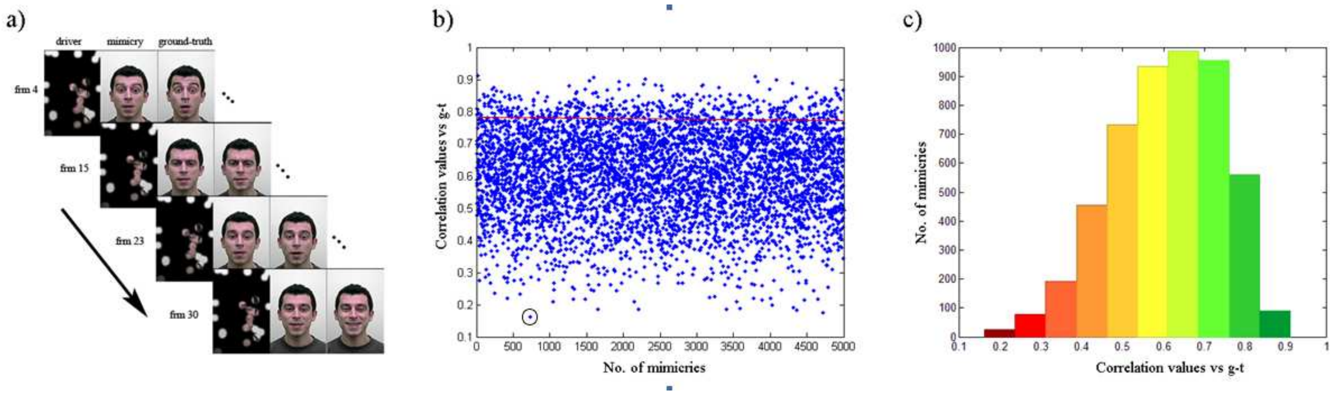


Figure 4.3 – Mimicry generation. a) The bubble-mask occluded driver sequence (left-most faces, occluded) was used to drive the avatar and produce the mimicries (middle faces). These were compared using our correlation metric to the ground-truth mimicry (right-most faces), which was produced by a non-occluded driver. It can be seen that the model recovers the facial expressions quite successfully despite the occlusions, albeit in a somewhat muted form. b) Points denote PC coefficient similarity between ground-truth and bubble-masked mimicries. Take for instance the mimicry generated by applying random mask number 728 (the circled point). Its generating PC coefficients are shown in the graph to have very low correlation with those of the ground-truth (corr = 0.16169) and upon visually inspecting the produced mimicry 728 there was hardly any facial movement reproduction, due to the applied occlusion mask. c) The histogram view of section b). Only the masks that produced the mimicries with PC coefficients highly correlated with ground truth coefficients (over the red line in section b), or represented by darker green colours in our histogram in c)) were classified as "good" and used to derive the proportion plane.

Face-map generation

We used the original bubbles method described in the previous section to derive a ProportionPlane for both our test sequences. We chose the masks resulting in the top 10% correlational values to be our "good" masks. It should be mentioned that, quite remarkably, visual inspection of the mimicries produced by drivers occluded with these "good" masks confirmed a high fidelity of the reproduction of facial actions, despite the massive occlusions. This demonstrated that the PCA face model successfully recovers aspects of expressions in those areas occluded in the driver sequence; however, the reproduced expressions in the avatar are slightly muted (fig. 4.3a).

A standard MATLAB *k*-means clustering routine was employed to partition the full set of masks into *k* clusters by using their corresponding coefficient correlation to the ground-truth as a criterion. So for example, when *k* = 2 the full set of masks was partitioned in two, with the partition containing masks corresponding to the top-tier of correlational values representing the "good" masks. Adding only these "good" masks together and dividing them with the TotalPlane yielded a rather noisy ProportionPlane image, i.e. the face-map (fig. 4.4). For a diagram of the whole procedure see fig.4.5.

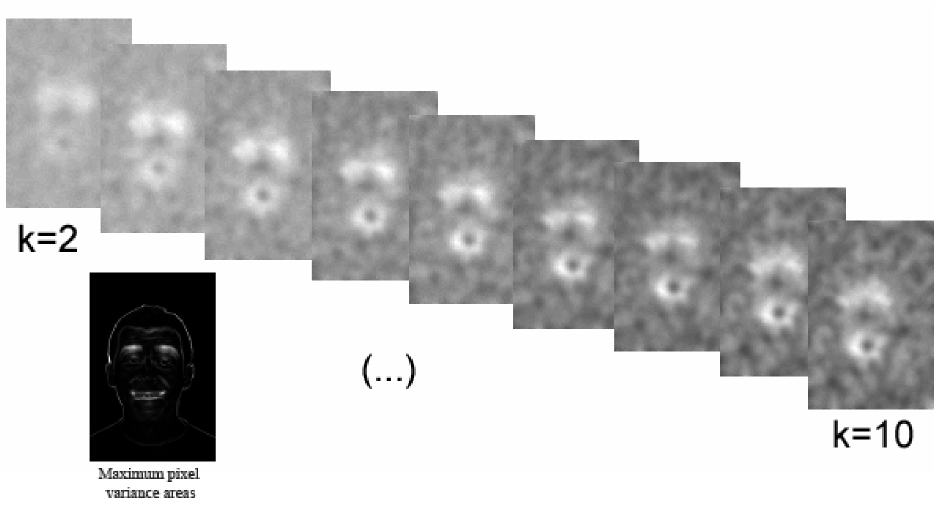


Figure 4.4 – Face-map. As we gradually selected out the bubble-masks corresponding to low correlation values brighter areas emerged from the noisy image. At this stage they are a rough representation of facial areas important for photo-realistic animation of our face model and not just the maximum pixel-value variance areas.

However as *k* approaches 10, a much clearer face-map depicting facial areas important for photo-realistic animation of our face model was gradually emerging. This face-map can be seen as a measure of the relative importance of the regions of the 2D image for the task at hand. These were the facial regions of the mouth and the eyes.

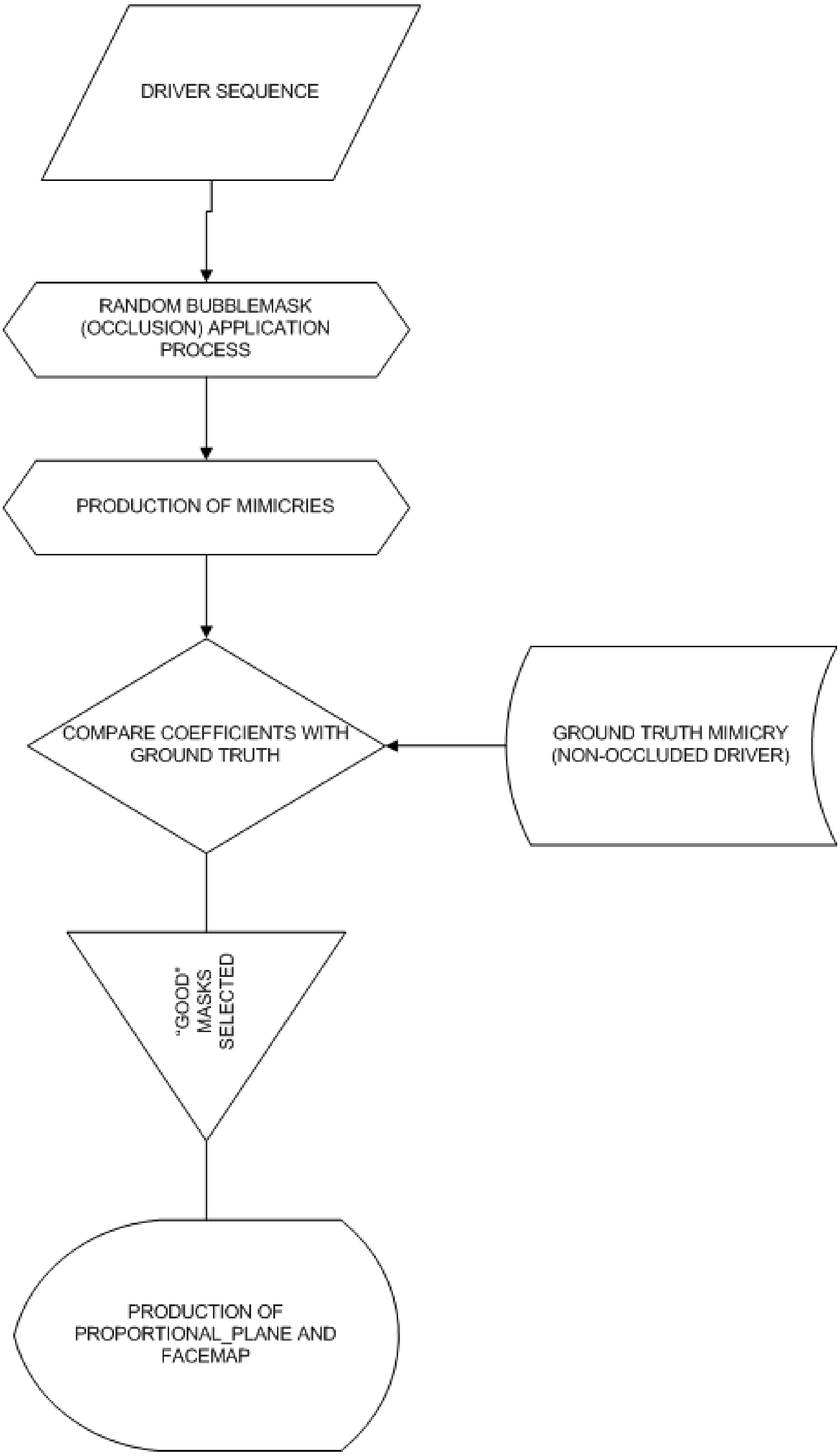

DRIVER SEQUENCE
RANDOM BUBBLEMASK (OCCLUSION) APPLICATION PROCESS
PRODUCTION OF MIMICRIES
COMPARE COEFFICIENTS WITH GROUND TRUTH
GROUND TRUTH MIMICRY (NON-OCCLUDED DRIVER)
"GOOD" MASKS SELECTED
PRODUCTION OF PROPORTIONAL_PLANE AND FACEMAP

Figure 4.5 – Block diagram of the face-map generation process using bubblemasks

Results

The face-map regions seem to overlap with the areas of maximum pixel-value variance, but importantly, they are not identical to them. This suggests that the method is not simply locating parts of the face that display most movement in our test sequences.

To derive the statistical significance of diagnostic regions, we used an accurate statistical test for smooth classification images (Chauvin et al., 2005). This test is based on the probability that, above a threshold t, a certain pixel-size cluster in our Z-scored classification image has occurred by chance. The derivation of the significant regions in our face map was done with a standard cluster test technique from the *Stat4Ci* MATLAB toolbox, with $p \leq .05$, $\sigma = 5$px, and threshold = 3.1. Figure 4.6 displays the thresholded classification images from both sequences. The areas that attained statistical significance are shown using the red pixels and the actual face used in the experiment was overlaid to facilitate interpretation. These areas are indeed the ones suggested by the face map.

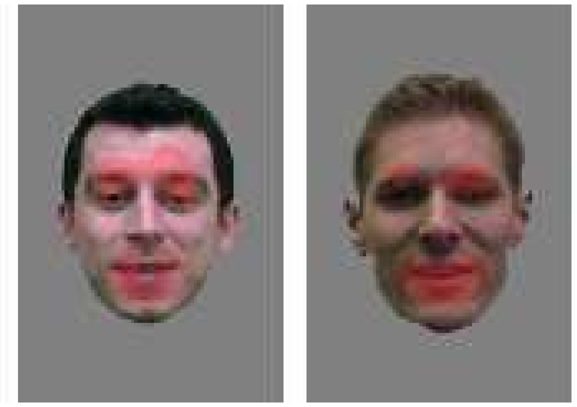


Figure 4.6 – Statistically significant diagnostic regions – Red areas denote the regions that attained statistical significance using our cluster test.

To demonstrate the actual importance of these areas we generated mimicries by using only the information from these statistically significant diagnostic regions in order to drive the avatar. As expected, the resulting mimicries were, visually, of very high fidelity. The correlation between these face-map driven mimicries and the ground-truth was very high (0.88438 and 0.89984, well within the top 1% of original bubble-mask coefficients).

Conclusions

The bubbles technique has been used as a way to identify regions of a stimulus that are important to perceptual discrimination. However here we show by studying the quality of image reconstruction, using a machine learning technique with no explicit knowledge of faces, that regions are significant because they reflect the important sources of variation in the facial image. Thus these areas are not necessarily important because of their functional roles (e.g. in visual speech or non-verbal communication) or because they are encoded by specialised neural modules but because these regions carry most information about the facial configuration.

There is growing evidence that faces are represented in terms of their deviation from a prototype. Leopold revealed high-level after-effects for static faces in tests involving stimuli like "anti-faces" (Leopold et al 2005). Curio also showed similar after-effects for dynamic facial expressions, by using "anti-expressions" as stimuli, created by a 3D morphable model for facial expressions based on laser scans (Curio et al 2007). These experiments indicate that a viable representation system for faces could be based on some mean prototype with axes of deviation radiating from this mean, which is the basis of the face model used in our experiment.

Peterson also studied the information distribution of face identification and its relation to human strategies in this task. This was done using a Bayesian ideal observer analysis. They found that both the ideal observer and the human subjects consistently use the visual information around the eye and mouth regions of the face when identifying individuals. This suggests that the human strategy of using the information from these regions for such tasks is commensurate with the concentration of visual information in real world faces (Peterson et al 2007).

We can ask what advantages does this form of representation offer. We identified significant regions by occluding them and then determining how effectively one could recover the original PCA based description of the face. Firstly this shows that PCA allows whole face information to be recovered from partial data. Secondly knowledge of important regions can improve the efficiency of encoding by only encoding critical information. Thirdly it shows that facial features (mouth and eyebrows) can be distinguished from the whole face in terms of their information content by spatially sampling a global PCA, thus linking features and configurations.

Chapter 4 summary

Our ability to process faces is known to depend on the spatial location of visual facial information we receive. A good method for revealing such diagnostic facial information for different categorisation tasks is the bubbles method. Here it succeeds in revealing diagnostic information for a performance-driven mimicry task carried out by a computer model of the face, built to a degree on biologically motivated principles. The face model was generated by vectorising a sequence of images of a talking face, extracting motion fields via an optic flow algorithm and calculating a set of basis actions using principal component analysis. The standard bubbles technique revealed the areas around and including the mouth and eyes as the most important ones for our task. These regions overlapped with but were not identical to areas of maximum pixel-value variance. Visual inspection also showed that the PCA face model recovers aspects of expressions in those areas occluded in the driver sequence. Until now bubbles were only used as a human search for diagnostic features in faces. Here, a system using reconstruction fidelity as diagnostic criterion and indifferent to the content of the stimulus, mimics the behaviour of human observers in face discrimination tasks. This information could be also very useful in further analysing and retrieving non-randomly occluded facial information, which is considered in more detail in the following chapter.

Chapter 5

FACIAL OCCLUSION PROCESSING

Randomised occlusions of the face such as those seen in the previous chapter are dealt with satisfactorily by the PCA face model, with global facial behaviour reproduced best when the areas around the mouth and eyes are not occluded. However, such randomised occlusions do not readily replicate real-life examples of facial occlusion. These are important since in many instances there is a need to identify subjects or facial behaviour from faces that are occluded with realistic occluders, placed in strategic positions around the face, such as the mouth and the eyes.

Psychologists have studied such occlusions - which fall in the category of non-systematic appearance variation - a lot less than systematic appearance variation. This latter type of variation comprises changes in viewpoint, expression or illumination direction and intensity. These types of changes can all have as a result occlusion of areas of the face which in turn causes a drop in identity recognition rates in human subjects.

Existing computer vision approaches

The few papers that have touched upon the face occlusion problem are motivated by the effect of occlusions in conventional face processing tasks such as face recognition and tracking. The majority of the face recognition literature presents data and results obtained under highly controlled scenarios with little variance in pose, illumination or set of expressions, but some have applied artificial and natural occlusions and tested the performance of their models with these impaired inputs. All of the papers described here use PCA and the information contained in the eigenface subspace to compensate for the details lost due to partial occlusions of the face.

The effect of occlusions on the performance of automatic face tracking models was investigated by Gross et al. (2006). They used Active Appearance Models (AAM) to track a face non-rigidly in a video. This, as mentioned in chapter two, requires a placement of markers on the face in order to extract the shape of the AAM by passing a triangular mesh through the marker vertices. Then the AAMs are constructed by applying PCA from a collection of such training images and they are fitted to input videos to track the face(s). Their main goal in this paper was to construct AAMs from occluded training images and then find the best-fit model parameters for face tracking in video. They managed to show empirically that AAMs computed from up to 45% occluded data were very similar to non-occluded data AAMs and to demonstrate successful video tracking of faces affected by various degrees of occlusion. They were also concerned with algorithmical issues within their approach and investigated speed vs performance trade-offs in different algorithms used in the construction and fitting of the AAMs. However, having to work with hand-placed markers to construct the AAMs is a major disadvantage of this method. It also has additional drawbacks with all the calibration and self-occlusion difficulties at the training stage, when all feature points have to be visible.

A method that doesn't use markers and that performs face recognition under partial occlusion is suggested by Tarres and Rama (Tarres & Rama 2005). This method consists of acquisition of the training images of full frontal faces, applying five different occlusions to them (fig. 5.1)

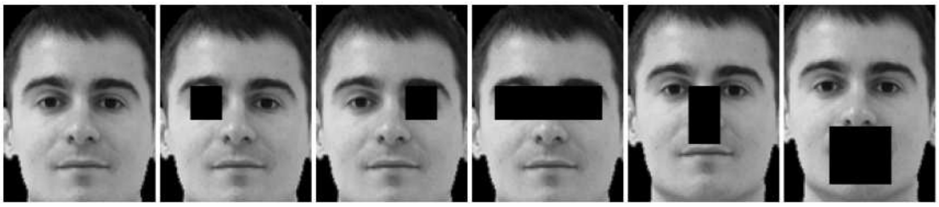


Figure 5.1 – Examples of training images showing various occlusion types from the Tarres and Rama database

Thus they created six different training subsets. PCA was then computed for each subset, together with the weights for each subject in the database which was done by projecting the six different images on the respective face subspace. The recognition stage consisted of the computation of the PC weights by projecting the image to be tested onto each face subspace, then classifying the subjects by verifying the weights of each face subspace and computing the reconstruction error. And finally they combine the different eigenface subspaces using a minimum reconstruction error strategy. This methodology yields good results in recognition of partially occluded faces, in fact better than the classical eigenfaces recognition method (O'Toole et al 1994). However, the main drawback is the fact that this combined method obviously carries a six times higher computational cost than the simple PCA recognition method.

Another automatic recognition system that uses PCA of local regions of the face for correct recognition of partially occluded faces was presented by Aleix Martinez (Martinez 2002). The faces are all divided in k local parts and an eigenface subspace is found for each of the training image parts. Then the occluded test face (divided in same number of parts) is projected respectively onto the eigenface subspaces and the closest match is found by means of Mahalanobis distance. Martinez demonstrates experimentally that occlusion of $1/6^{th}$ of a face does not decrease recognition accuracy. Even for cases where $1/3^{rd}$ of the face is occluded the identification results are shown to be very close to those obtained in the non-occluded case.

A final paper worth mentioning is that of Lanitis (Lanitis 2004). The model presented in this paper completes a number of tasks, such as:

- localisation of the face using AAMs
- creation of the face model using PCA of all training examples
- detection of occluded area by calculating the residual difference of the occluded test face image and PCA reconstructed face obtained by projecting image onto eigenface subspace.
- implementation of face recognition algorithm that makes use of information only from the non-occluded facial regions.

The experimental results published in this paper suggest that this model outperforms standard minimum distance classifier recognition methods such as Martinez's. However, this approach is computationally more expensive due to the additional localisation and occlusion detection steps.

Effect of occlusions on the PCA-based face mimicry model

The above mentioned approaches do not deal with dynamic facial behaviour. However, they all use PCA and the information contained in the eigenface subspace to compensate for the details lost due to partial occlusions of the face. Principal components tend to track global changes, and are therefore ideal for retrieving facial motion lost due to occlusion since facial muscles rarely function independently. Indeed, the photo-realism of the PCA performance-based mimicry model shows that they capture well the correlations between various muscles of the face. They seem to be able to encode with good precision a large number of high level facial movements and systematic variance in the face, such as constituent mouth shapes for speech, or the appearance and disappearance of a smile.

But what will happen if the driving information is incomplete, due to some non-random occlusion of a region of the face? Since the PCA-based mimicry system described in this thesis constrains movement by forcing it to be consistent with movements the training face can produce, and because this type of input yields a strong correlation between rigid and non-rigid motion, will the model be able to build a representation of the missing information based on the rest of the information it receives?

The following sub-sections explore these questions and attempt to measure the performance of the model in dealing with occlusions.

Test stimuli: the database of occluded sequences

In order to test the model, a set of facial motion sequences was created. One serves as the ground-truth and represents a capture of the face in natural motion in clear frontal view while the others include occlusions, but are captured from the same view-point and in same lighting conditions. This control over lighting conditions, pose and occlusion types was the main reason for creating this new database and not using an existent one. Types of occlusions to be used were selected on the basis of real-life occurrence frequency and recognition difficulty (Murphy and Bray 2003) and appearance in other databases, like the AR Face Database, Computer Vision Center, University of Barcelona. They were captured using a JVC GR-DVL 9600 digital video camera, at a rate of 25 frames per second, resolution 160 by 240 pixels. The sequences were of an expressive face telling a one-liner joke and they lasted around 8 seconds. The sequences included:

- mouth area occlusion (by a pointing finger and by a hand over mouth)
- lower face occlusion (by a scarf)

- eye occlusion (by dark wraparound sunglasses)
- top of head and hair occlusion (by scarf/hat)
- side of face occlusion (by newspaper sheet)
- and combinations thereof

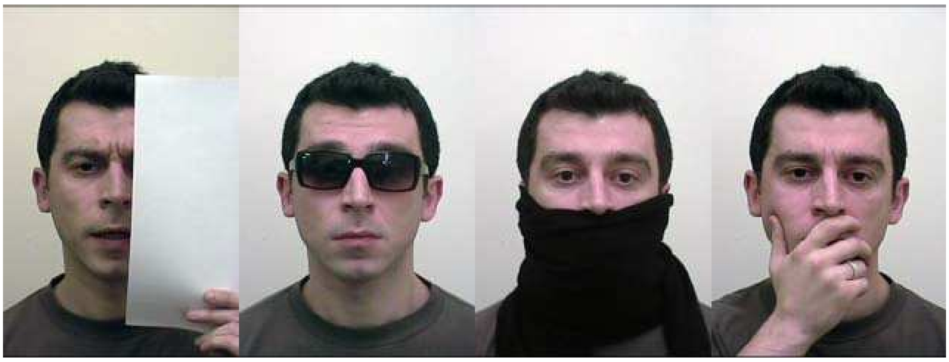


Figure 5.2 – Examples of natural occlusion types from the database

In addition to these natural occlusions, an identical set of occlusions was added onto the ground-truth sequence thus creating a database with perfect temporal alignment since it was produced from only one actual sequence (fig. 5.3).

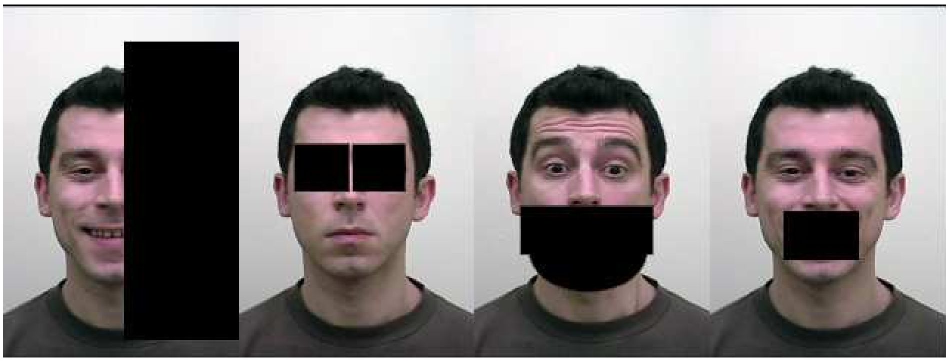


Figure 5.3 – Examples of artificial occlusion types from the database

Finally, two sequences with dynamic occlusions were also captured (fig. 5.4), one with a face moving from side to side and the other with an object entering and leaving the scene (in this case a waving hand).

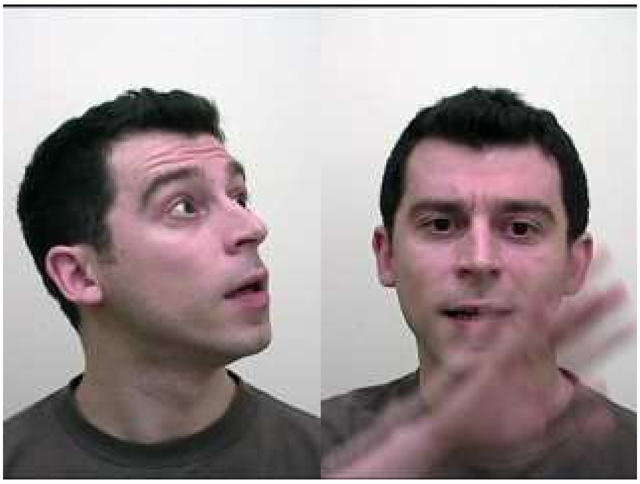


Figure 5.4 – Examples of dynamic occlusion types from the database

Occlusion-affected face mimicry results

The sequences were all morph vectorised as described in 3.1.1. The non-occluded sequence was processed using PCA and basis vectors were extracted, forming thus the generative model of the target face based on principal components. The facial movements from the vectorised occluded sequences can now be projected onto the computer generated model.

The results show that the motion information from the occluded mouth area can indeed be correctly reproduced by the system, albeit in a weaker and somewhat muted form. The non-occluded areas are reproduced almost perfectly, as can be seen from the examples in the figure 5.5.

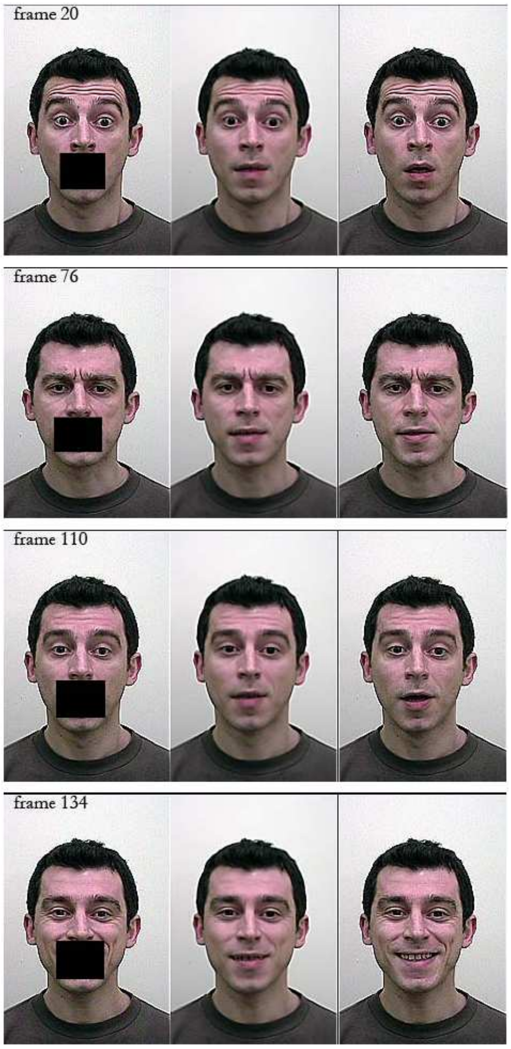


Figure 5.5 – Selected frames showing the occluded "driver"-face (left column), the ground-truth (right column), and the avatar mimicking occluded mouth motion weakly, but correctly

The PCA mimicry's frames shown in fig.5.5 seem to do relatively well and provide a realistic reproduction of occluded facial behaviour, as compared to the avatar sequence obtained by the non-occluded driver face, i.e the ground-truth. The relatively good success rate of this automated model in retrieving missing facial information was to an extent expected since the artificial mouth-occlusion driving example is the one with the smallest occlusion area and since there's a great deal of correlated information from the visible areas in the immediate vicinity of the mouth that is projected onto the basis vectors of the avatar. On the other hand, the motion from the sequence with a scarf around the lower face area, for example, results in almost no mouth movement reproduction in the avatar, and this is because there's hardly any mouth-area-correlated information coming from that input. The performance evaluation was done by visual inspection of the reproduced facial behaviours and its comparison with the ground-truth mimicry.

Exaggerating actions through basis coefficient scaling

Since all sequences of facial motion used here are vectorised and mean-centred, each vector can be considered to be a departure from the mean for that sequence. Frames from the original sequences can thus be exaggerated, or made subtler, by simply multiplying their respective mean-centred vectors by some factor k in order to magnify, or reduce, the departure from the mean. Adding back the mean, then converting back into images, results in the exaggerated frame. This effectively increases, or decreases, the distance of the point in face-space from the origin along the vector from the origin to the point's original position.

This can similarly be applied to performance-driven animations. Resulting sequences in which the driving vectors do not project strongly onto the basis, for example, can be corrected or exaggerated by applying a scaling factor to the coefficients in face-space in order to compensate. This can also be applied in order to make facial motion more coherent in situations where subtle or weak movements would otherwise be missed. Crucially, it is this capability of the model that allows for an enhancement of the muted facial behaviour retrieved from the occluded areas of the face.

Simple pixel-wise intensity vectorisations encode facial movements as changes in brightness, so an exaggeration will necessarily make regions that become darker even darker and regions that become lighter even lighter. This does not correspond particularly well to exaggerated gestures, but accentuates the image changes. The warp and morph vectorisations capture changes in a more realistic manner, so exaggeration will result in more extreme positional changes away from the reference.

Thus, the facial behaviours can be exaggerated by introducing a scaling factor k to the coefficients applied to our avatar basis set:

$$\mathbf{\Omega} = \mathbf{B}(k\mathbf{C})$$

where columns of $\mathbf{\Omega}$ represent mean-centred mimicry frames, $\mathbf{B}$ the basis vector (or principal components) set and $\mathbf{C}$ the set of coefficients of principal components, obtained by projecting mean-centred driving vectors onto $\mathbf{B}$, as explained in subsection 3.1.2 of this thesis.

This will compensate for the occlusion-affected, weakly projecting driving vectors. In the important mouth occlusion case seen earlier, this makes the mouth movement more realistic and recognisable. It almost allows lip-reading an avatar driven by a face with a “masked” mouth area! But since this operation multiplies the whole coefficient matrix by some factor k, there’s an inherent risk of an anomalous exaggeration of the other facial areas, which become rather like caricatured animation when the multiplier is too high (x4.5 in fig.5.6). Such exaggerations also accentuate colour changes in the three colour planes in the case of morph vectorisation used by the PCA face model described here.

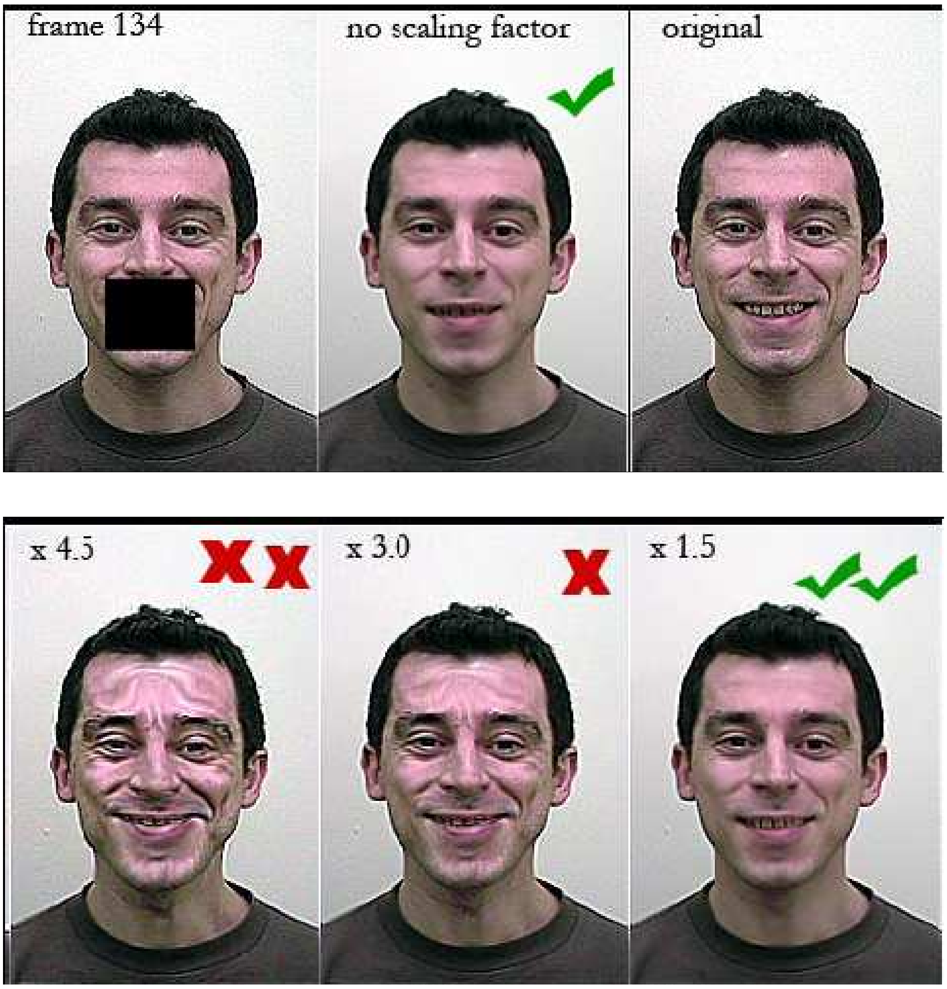


Figure 5.6 – The example frame can be seen here in its occluded and original form, together with its mouth-occlusion mimicry reproductions with scaling factors *k* = {1, 1.5, 3.0, 4.5}. Visual inspection confirms that out of these examples, multiplying the coefficients by 1.5 seems to give the best reproduction quality of the missing mouth motion information, but higher ones, like *k* = 4.5, create grotesque caricatures!

It's worth noting that in contrast with the results obtained using static occlusions of the face, the system failed to generate satisfactory mimicries when driven by dynamic occlusion sequences. This is because such a set of facial motion and deformation was never captured in the generation of the avatar, rendering it impossible to reproduce with only linear combinations of movements from the existing example set. Also, iconic changes caused by the hand entering and leaving the view and the head turning from side to side (fig.5.4) are probably too large or too fast for the optic flow algorithm to successfully extract the flow fields necessary for warping.

Performance metrics

In all occlusion-affected mimicries described in this chapter, the quality of the mimicries, i.e. their similarity with the ground-truth produced by a non-occluded driver sequence, was evaluated using visual inspection. Such a subjective evaluation method brings with it a multitude of inconveniences and doesn't quantify the performance of the model. This is why it is necessary to devise a performance metric that will provide objective evaluation and the basis for an optimization of the model.

Two basic metrics that can be used for this purpose are the mean and standard deviation of respective pixel intensity differences between two corresponding frames from the reconstructed sequence and the ground-truth. This mean difference represents a special form of the widely used Minkowski error metric, which for images **x** (say the frame to be tested) and **y** (the corresponding ground-truth sequence frame) is defined as:

$$E_p = \left(\sum_{i=1}^{n} \left| x_i - y_i \right|^p \right)^{1/p}$$

where x_i and y_i are the *i*-th pixels in images **x** and **y** respectively and n is the number of pixels, i.e. height times width in pixels of pre-processed (scaled, aligned, filtered, etc) images **x** and **y**. The constant exponent p, which can vary in the range $p \in [1, \infty)$, typically takes values between 1 and 4. For our mean difference metric we just use p as 1 and multiply the above Minkowski metric equation with $\frac{1}{n}$, giving:

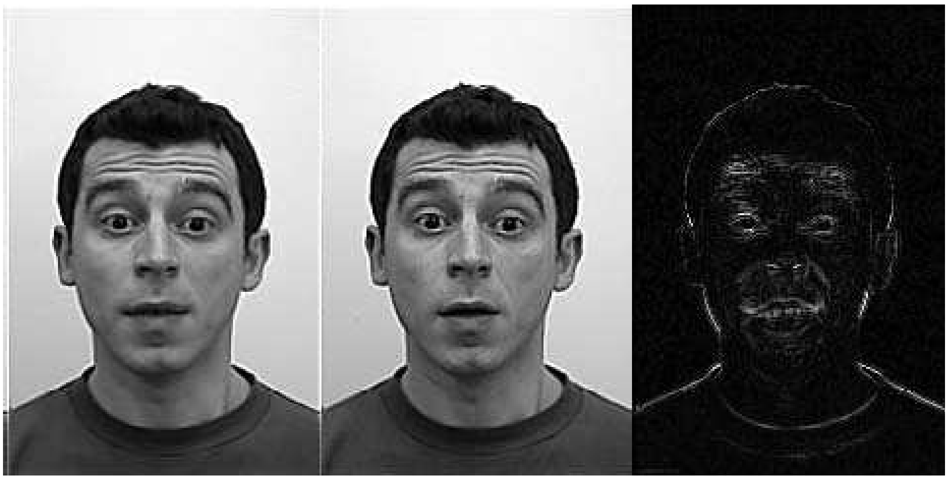
$$\mu_{x,y} = \frac{1}{n} \sum_{i=1}^{n} (x_i - y_i)$$

This metric represents the difference of the *luminance* of images **x** and **y**, since the mean pixel intensity estimates the luminance of an image**.** For the standard deviation difference we have:

$$\sigma_{x,y} = \left(\frac{1}{n} \sum_{i=1}^{n} ((x_i - y_i) - \mu_{x,y})^2 \right)^{1/2}$$

This metric represents the difference of the *contrast* of images **x** and **y**, since the standard deviation of the pixel intensity estimates the contrast of an image.

A C++ routine was developed to visualise (see fig.5.7) and calculate the mean and standard deviation of these pixel value differences.



reconstruction *ground-truth* *Pixel value difference*

Figure 5.7 – This example visualizes a map of the pixel value difference between the intensity values of a mimicry frame reconstructed using an occlusion-affected driver and the ground-truth.

The routine is then used to test the model's performance on mimicking the mouth occlusion sequences, i.e. make a comparison of an animated avatar from the sequence inputs with no occlusion (ground-truth), with artificially placed mouth occlusion and no scaling factor (mouth-occlusion-to-face, 'mo2f'), scaling factor 1.5 ('mo2f x1.5'), 3.0 ('mo2f x3.0') and 4.5 ('mo2f x4.5').

The luminance and contrast metrics seem to perform well in comparing the performance of the model, insofar as they show quite clearly which mimicries are further away from the ground-truth (fig.5.8). The results seem to agree with the visual inspection of the performance of the model in mimicking the mouth-occluded sequence. It confirms that the non-scaled and the ×1.5 scaled mimicries are doing best (a perfect reconstruction would be equal to the ground-truth and would lay flat on the x-axis). It even detects that on frame 134, the animation produced by applying a 1.5 scaling factor should be closer to the ground-truth than the one with no scaling factor, which is what was detected perceptually in fig.5.6.

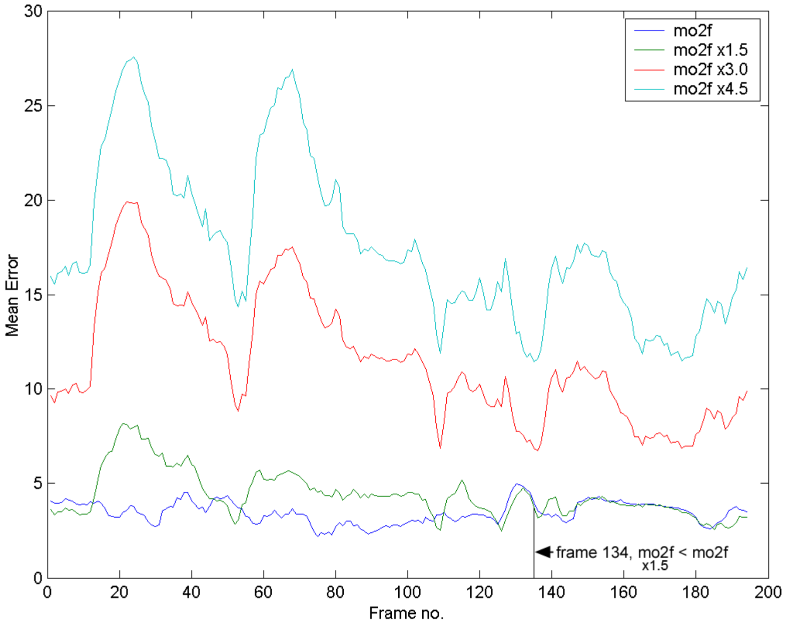


Figure 5.8 – This shows that the mimicry with no scaling factor and the one with a small 1.5 scaling factor are closest to the ground-truth (x-axis), seemingly agreeing with the visual inspection results.

The standard deviation error graph looks very similar to the mean error one but with accentuated differences between the respective frames (fig.5.9).

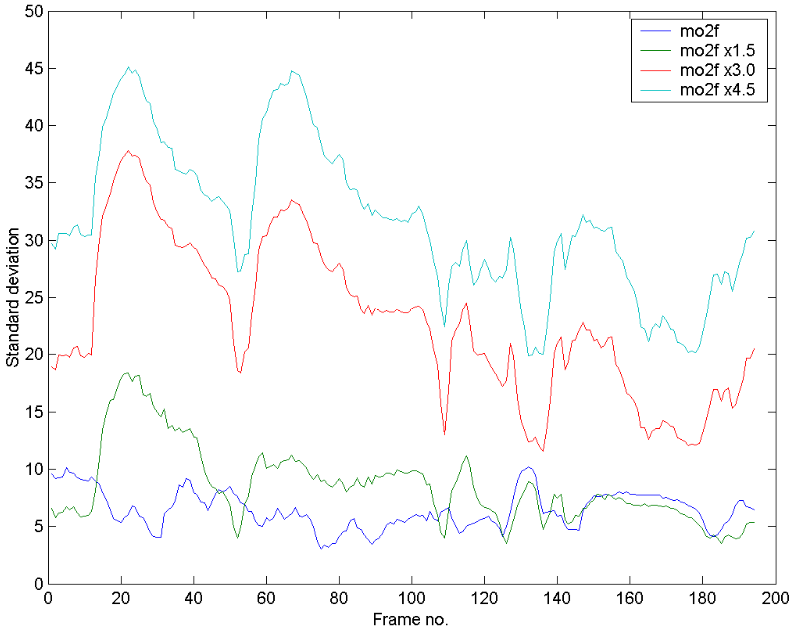


Figure 5.9 – Standard deviation differences

A single numerical value for these errors can be obtained by taking the mean values of both luminance and contrast differences across frames of the produced occlusion-affected mimicries:

	mo2f	mo2fx1.5	mo2fx3.0	mo2fx4.5
Mean of all $ME_{x,y}$	3.4639	4.3966	11.5584	17.5835
Mean of all $SD_{x,y}$	6.4410	8.5793	22.8282	31.5885

Table 1 The table indicates that the mimicry produced without rescaling the PC coefficients (mo2f) has the smallest mean value of mean errors and standard deviations per frame. This doesn't correspond with visual inspection results, which seem to indicate that the mo2fx1.5 mimicry is closer to the ground-truth!

However, the results of table 1 show the mimicry produced without rescaling the PC coefficients as the closest one to the ground truth, which is not what was perceived by visual inspection by human subjects. So even though the graphs 5.8 and 5.9 seem to suggest that the mean and standard deviation of the frame differences could work as good bases for image quality metrics, averaging (or summing for that matter) these values in order to obtain single-numerical quantitative measures of reproduction quality doesn't produce an accurate prediction of perceived quality of the mimicry, as compared with the ground-truth.

A good explanation for the failure of this class of metrics is given by Wang et al. (2004). They point out that while these types of quality metrics are simple to calculate and have clear physical meanings they do not take into account the fact that the human visual system is highly adapted to extract structural information from the visual scene. The mean and standard deviation just don't do that. In fact, these metrics implicitly assume that all pixel intensities are independent, which means that the ordering of the pixel intensities should have no effect on the overall distortion measurement. This is in sharp contrast to the fact that natural image signals are highly structured and that the ordering and pattern of the signal samples carry most of the visual information in the image. Therefore, an accurate metric of image quality should be able to capture the structural information and/or sense the structural changes in the image signals.

A simple demonstration can show why these metrics are unsuitable to predict image quality as perceived by the human visual system. It is quite possible to take an image and distort it by operations of mean shifting, contrast stretching, Gaussian blurring and compressing, while keeping the *ME* value constant. However, visual inspection of the resulting images clearly shows great variation in perceived quality (fig.5.10)

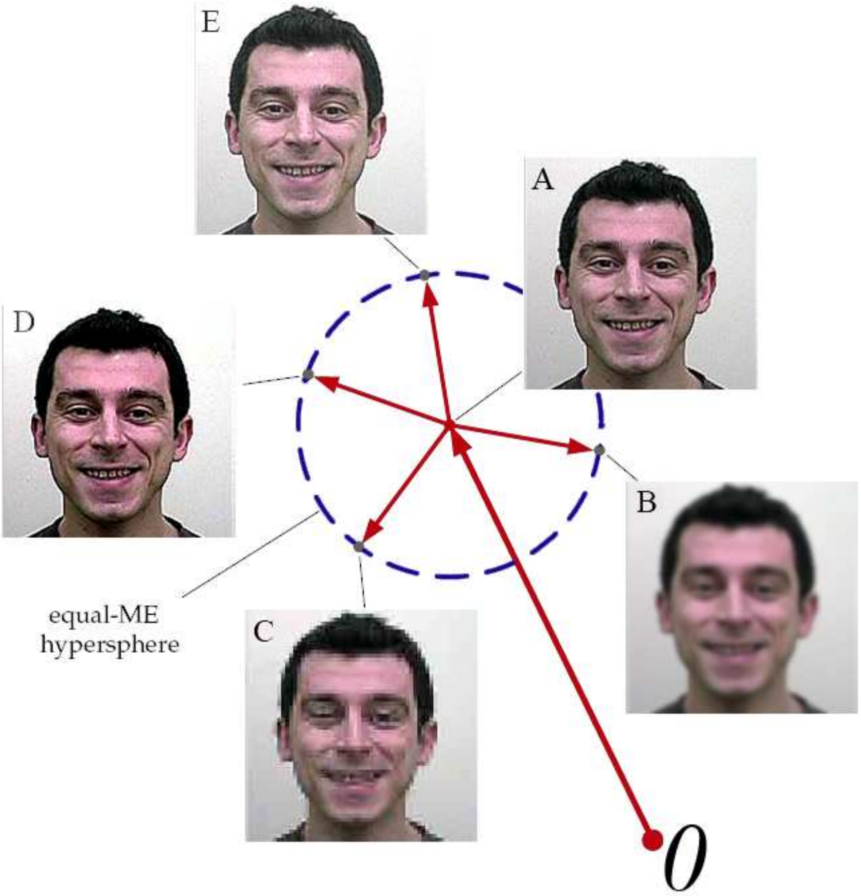


Figure 5.10 – Images are represented here as points (or vectors) in this *n*-dimensional image space. A is the original image, B the blurred image with ME = 15.4, C the JPEG compressed image with ME = 15.6, D the contrast stretched image with ME = 15.3, and E the mean shifted image with ME = 15.3. Despite the fact that these images are points on virtually the same hypersphere they display noticeably different visual quality.

What Wang et al. suggest is a metric that combines the luminance and contrast estimates, together with an estimate of the normalised signals. So if the luminance of an image **x** was defined as:

$$\mu_x = \frac{1}{n}\sum_{i=1}^{n} x_i$$

then the luminance comparison function *l*(**x**, **y**) is defined as $l(\mu_x, \mu_y)$. The mean intensity values are then removed from the image signals (zerocentring) and the contrast is defined as an unbiased estimate by the standard deviation:

$$\sigma_x = \left(\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \mu_x)^2 \right)^{1/2}$$

The contrast comparison $c\,(\mathbf{x},\ \mathbf{y})$ is then defined as $c\,(\sigma_x, \sigma_y)$. Then, the signals are normalised by their own standard deviation, thus forming the third component of the metric:

$$s(\mathbf{x},\mathbf{y}) = s\left(\frac{\mathbf{x}-\mu_x}{\sigma_x}, \frac{\mathbf{y}-\mu_y}{\sigma_y} \right)$$

Finally the three elements of the metric are combined to form an overall similarity measure:

$$S(\mathbf{x},\mathbf{y}) = f\,(l(\mathbf{x},\mathbf{y}), c(\mathbf{x},\mathbf{y}), s(\mathbf{x},\mathbf{y}))$$

They complete the definition of their similarity measure by defining the three functions $l(\mathbf{x},\mathbf{y}), c(\mathbf{x},\mathbf{y})$ and $s(\mathbf{x},\mathbf{y})$, as well as the combination function *f*. Details can be found in Wang et al. (2004). The result is a class of image similarity measures which they call collectively as *Structural SIMilarity* (SSIM) *Indices* between signals **x** and **y**:

$$SSIM\,(\mathbf{x},\mathbf{y}) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{x,y} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}$$

where C_1 and C_2 are constants introduced to avoid instability when $\mu_x + \mu_y$ or $\sigma_x + \sigma_y$ are very close to zero. The design of this metric ensures symmetry, $S(\mathbf{x},\mathbf{y}) = S(\mathbf{y},\mathbf{x})$; boundedness, $S(\mathbf{x},\mathbf{y}) \leq 1$; and a unique maximum, where $S(\mathbf{x},\mathbf{y}) = 1$ iff $\mathbf{x} = \mathbf{y}$.

Applying the SSIM metric to faces from figure 5.10 we get results that correspond very well with human visual inspection results, as can be seen in fig.5.11.

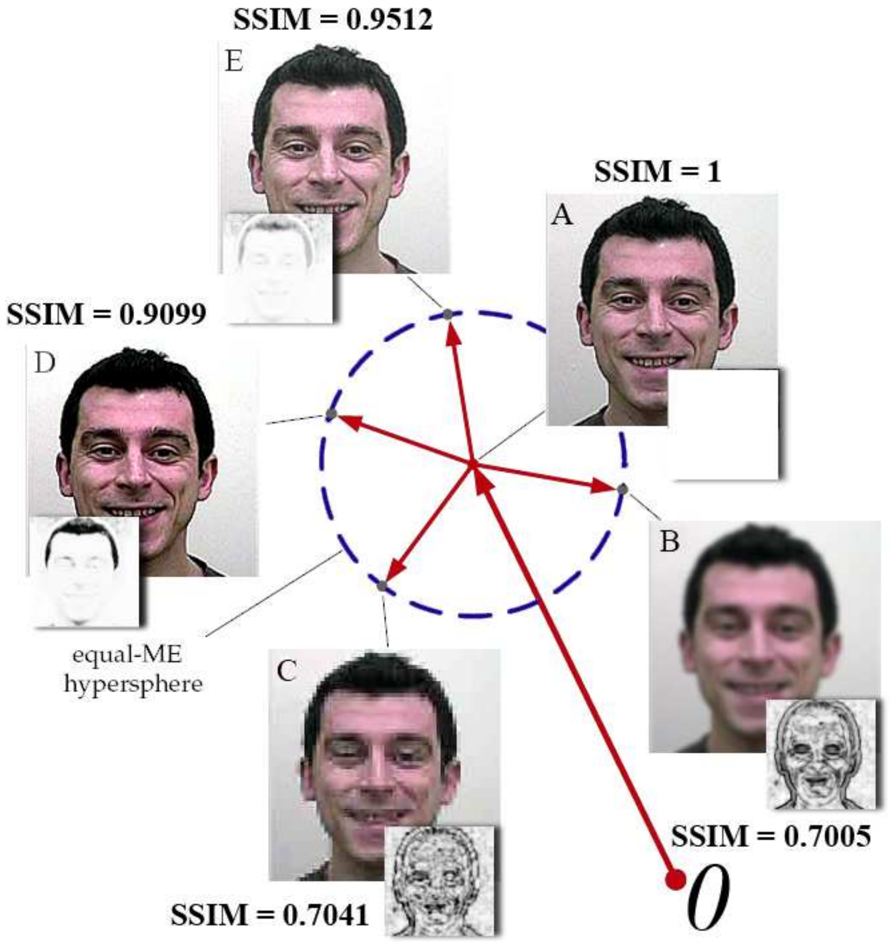


Figure 5.11 – SSIM values for E (mean-shifted) and D (contrast-stretched) images are a lot higher than those for B (blurred) and C (JPEG-compressed), which is consistent with the perceptual quality of the images. The smaller boxes show SSIM index maps of the respective images, where brighter indicates better quality.

The same correspondence is observed when applying SSIM to the mimicry sequences (single frame SSIM indices are added up and divided by the number of frames, giving mean SSIM, or MSSIM). This is described in table 2.

	mo2f	mo2fx1.5	mo2fx3.0	mo2fx4.5
ME	3.4639	4.3966	11.5584	17.5835
MS	6.4410	8.5793	22.8282	31.5885
MSSIM	0.7936	0.8399	0.5621	0.2375

Table 2 – MSSIM indicates that the mo2fx1.5 mimicry is closest to the ground-truth, followed by mo2f, then mo2fx3.0 and finally by mo2fx4.5. This corresponds with visual inspection results.

Wang et al. expanded the testing of the SSIM metric by comparing the results of the metric with the visual evaluations of a number of JPEG compressed images by 25 human subjects. They repeated this experiment with other metrics. SSIM came top vs every other metric tested (Wang et al 2004). This, together with our results, validates the use of this metric in quantitative evaluation of mimicry quality and in optimising the parameter settings of the PCA face model for even more photorealistic retrieval of occluded information. The metric was applied to all other types of artificial occlusion mimicries from the database and correctly predicted the best perceived mimicries.

So this metric performs well in comparisons between artificially-occluded sequence mimicries and the ground-truth from which the artificially-occluded sequences were generated. But what happens if the drivers are naturally-occluded sequences? In such cases there will be some temporal misalignment, since no actor can repeat saying a sentence in exactly the same time as they did during the capture of the non-occluded ground-truth sequence.

In tests using the luminance and contrast differences alone, this misalignment caused the lower end of mean errors and standard deviation values to more than double, suggesting that the animation from natural-occlusion drivers is much worse (fig.5.12). However, on visual inspection of the mimicries there is very little difference, in fact if we compare like for like, an animated avatar driven by a sequence with a computer generated mouth occlusion with one driven by a sequence with a natural mouth occlusion, like a hand over the mouth, is virtually impossible to tell apart.

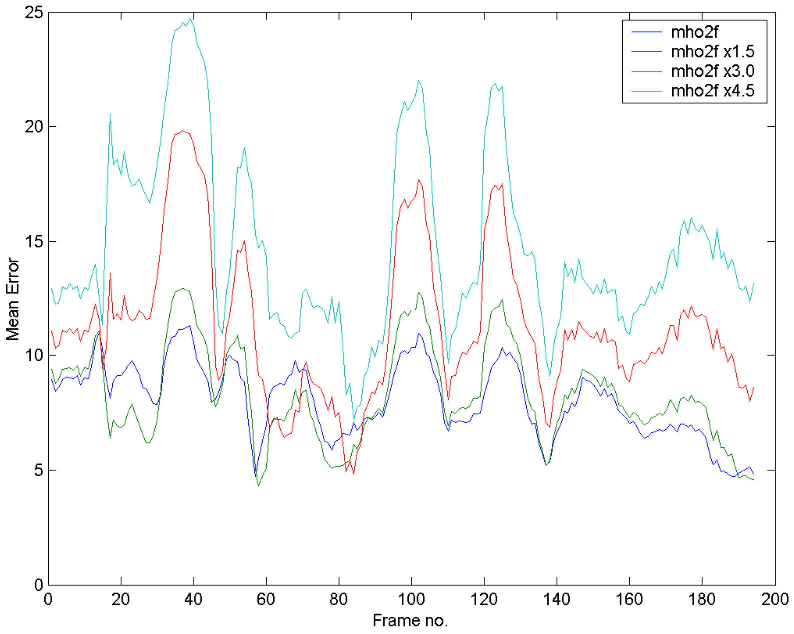


Figure 5.12 – This shows no clear separation between the mimicries with a considerable number of frame mean error comparison results not agreeing with the visual inspection results.

MSSIM performed better yet again, but the similarity measures are a bit lower than the artificial occlusion ones, which can be attributed to the dissimilarity contributed by the temporal misalignment of the natural occlusion-affected mimicries and the ground-truth (table 3).

	mo2f	mo2fx1.5	mo2fx3.0	mo2fx4.5
ME	7.9584	10.2336	15.3504	21.0026
MSSIM	0.6654	0.7596	0.4005	0.2068

Table 3 – MSSIM indicates that the mo2fx1.5 mimicry is closest to the ground-truth, followed by mo2f, then mo2fx3.0 and finally by mo2fx4.5. This corresponds with visual inspection results.

Chapter 5 summary

Effects of non-systematic facial appearance variation such as partial occlusions on face perception and analysis have not been studied to the same extent as the effects caused by changes in viewpoint, expression or illumination direction and intensity. However most studies of facial occlusion that have been conducted have indeed tended to make use of PCA and the information contained in the eigenface subspace to compensate for the details lost due to partial occlusions of the face. This property was demonstrated using the PCA-based mimicry system in chapter 4, where occlusions were randomised. In this chapter, a database of video sequences affected by non-random occlusions (artificial, natural and dynamic) was created. These sequences were used to drive a PCA-built avatar and the mimicries obtained in this way were compared to the ground-truth mimicry which in turn was produced using a non-occluded driver.

Visual inspection of the resulting mimicries results show that the motion information from the occluded areas can indeed be correctly reproduced by the system, albeit in a weaker and somewhat muted form as compared to the ground-truth (the occluded mouth-area results were used throughout as a representative case in this chapter, mainly due to the superior complexity of behaviour stemming from this area and its importance in human communication and interaction). To amplify the muted reproductions of facial behaviour, the facial actions were exaggerated through principal component coefficient scaling by arbitrary scalars. Visual inspection again showed that certain factors did indeed appear to increase the similarity of the occlusion-affected mimicries as compared to the ground-truth. In order to provide an objective evaluation and a basis for an optimization of the model, a number of performance metrics were suggested. Simple error quantisation metrics such as mean error (frame luminance difference) and standard deviation error (frame contrast difference) didn't seem to correspond well to visual inspection results due to their inability to take into account any structural information with the images compared. However, a metric - suggested by Wang et al. and named Structural Similarity Index (SSIM) - that combined these two measures together with a normalisation component did indeed succeed in predicting perceived image, and consequently, mimicry quality in all occlusion cases.

All improvements by coefficient scaling described in this chapter were achieved by arbitrary and heuristic scaling of the whole matrix of coefficients. The side-effect of such an approach was an anomalous exaggeration of the other facial areas, which became rather like caricatured animation when the multiplier is too high. To combat these shortcomings ways of manipulating specific PC coefficients, together with other methods of component analyzes, are developed in the next chapter.

Chapter 6

INDEPENDENT COMPONENT ANALYSIS FOR IMPROVED FACIAL MIMICRY

PCA carries out a decorrelation of the input data, but does not deal with dependencies of the higher order (Joliffe 1986). What decorrelation means is that the variables cannot be predicted from each other using a simple linear predictor. This means that there can still be nonlinear dependencies between them that PCA would not be able to analyse. A clear example of this is, for instance, a trigonometric function, say cosine. In $y = \cos(x)$, x and y are clearly related and highly dependent of each other, but their correlation value would still be zero! In image processing, edges are another example of a high-order dependency, as are elements of curvature and shape.

Independent component analysis (ICA) is a generalization of PCA that separates the high-order dependencies in the input, in addition to the second-order dependencies. PCA encodes second-order dependencies in the data by rotating the axes to correspond to directions of maximum covariance. It models the data as a multivariate Gaussian and places an orthogonal set of axes such that the two distributions are completely overlapping. ICA does not constrain the axes to be orthogonal, and attempts to place them in the directions of statistical dependencies in the data (fig.6.1). Each weight vector in ICA attempts to encode a portion of the dependencies in the input, so that the dependencies are removed from between the elements of the output (Comon 1994). This characteristic of ICA is advantageous for the purposes of this study because in the context of face behaviour it promises to produce a set of statistically independent basis actions which encode local rather than global changes.

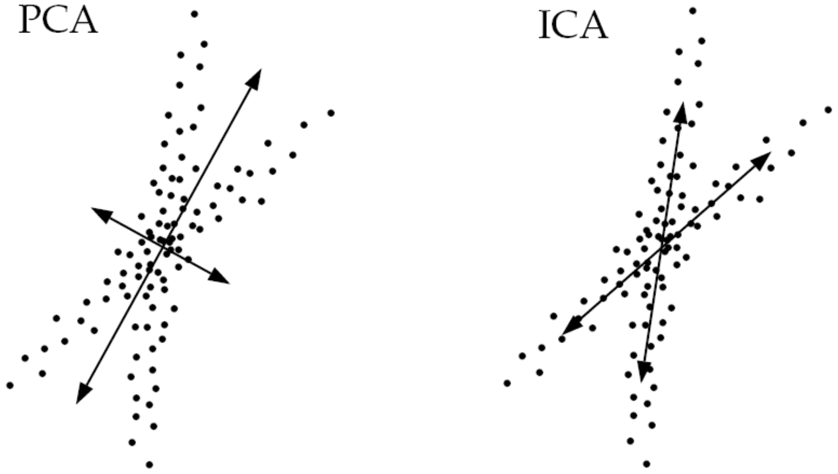


Figure 6.1 – A 2-D data distribution and corresponding PC and IC axes (Lewicki & Sejnowski 2000).

IC representations of face information vectors

As mentioned in the previous chapter, any exaggeration of specific principal components has unwanted side-effects in the mimicry results because PCs don't describe single localised facial actions but combinations thereof, i.e. more global changes. Since this behaviour generally and inherently consists of combinations of localised unit actions, it is these combinations that will be encoded by single principal components. So as a result, when an amplification of the action of the mouth and lips was attempted by exaggerating the relevant PC, it also amplified the actions and behaviour of eyes and eyebrows in the resulting mimicry. This was an unwanted side-effect that invariably deteriorated the quality of the mimicry.

ICA, in the other hand, offers a way of encoding more local basis actions that are independent from each other and thus can be manipulated individually without the above mentioned side-effects. Just as with PCA, a set of M_{train} training vectors containing both shape and texture information from the morph-vectorised sequence of the individual whose the face we wish to drive, $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_{M_{train}}\}$, is zero-centred. This is then put into matrix $\mathbf{\Phi}$ such that $\mathbf{\Phi} = \{\boldsymbol{\varphi}_1, \boldsymbol{\varphi}_2, ..., \boldsymbol{\varphi}_{M_{train}}\}$, where $\boldsymbol{\varphi}_i = \mathbf{x}_i - \boldsymbol{\mu}_{train}$. ICA then extracts a set of independent basis vectors $\mathbf{U} = \{\mathbf{u}_1, \mathbf{u}_2, ..., \mathbf{u}_P\}$ from the training set, with $P \leq M_{train}$.

The task of finding these independent components is formulated as an estimation of the source signals $\mathbf{S}$ and identification of a mixing matrix $\mathbf{H}$ assuming only the statistical independence of the primary sources, i.e. independent components, and the linear independence of columns of $\mathbf{H}$.

The generative model can be represented in batch or matrix form as:

$$\mathbf{\Phi} = \mathbf{HS}$$

The inverse form used to find the source signals is $\mathbf{S} = \mathbf{W\Phi}$, where one estimates the pseudo-inverse un-mixing (or separating) matrix $\mathbf{W}$ which then yields the source independent components on matrix multiplication with the input matrix. The unknown matrices are estimated so that the rows of $\mathbf{S}$ and columns of $\mathbf{W}$ are as independent as possible. This independence is measured by an information-theoretical cost function such as the Kullback-Leibler distance or other criteria like sparseness or linear predictability.

So the face vectors in $\mathbf{\Phi}$ were all assumed to be a linear combination of an unknown set of statistically independent source images $\mathbf{S}$ mixed by an unknown mixing matrix $\mathbf{H}$. The sources were then recovered by a matrix of learned filters $\mathbf{W}$ which produced the statistically independent outputs $\mathbf{U}$ (fig.6.2)

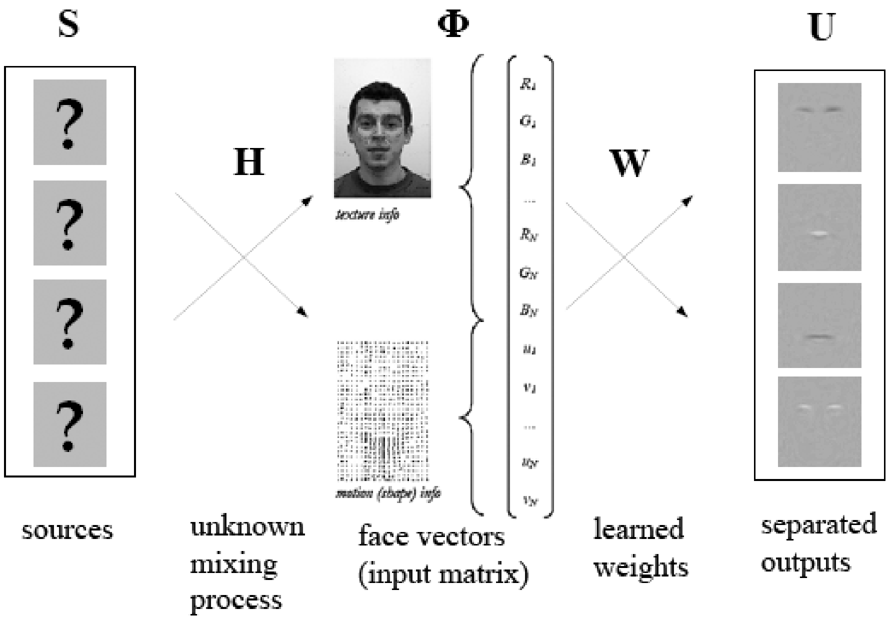


Figure 6.2 – Vectors in $\mathbf{\Phi}$ were all assumed to be a linear combination of an unknown set of statistically independent source images **S** mixed by an unknown mixing matrix **H**. The sources were then recovered by a matrix of learned filters **W** which produced the statistically independent outputs **U**.

Application of ICA to occlusion-affected PCA-based mimicries

Can these independent component outputs be used firstly to produce mimicries and secondly to improve the quality of the occlusion-affected mimicries? The full mimicry method used in this thesis was described on chapter three. After completing the morph-vectorisation of both the training and the driving set and the extraction of the PCs from the training set, there was the key step of projecting the *N*-dimensional vector containing the vectorised driver information, $\boldsymbol{\psi}_i$, into the *P*-dimensional subspace described by the principal components basis $\mathbf{B} = \{\mathbf{b}_1, \mathbf{b}_2, ..., \mathbf{b}_P\}$:

$$\mathbf{c}_i = \mathbf{B}^T \boldsymbol{\psi}_i$$

which produced the coefficients $\mathbf{c}_i$ that were used with the PCs to construct the mimicry.

However, this method only works in the case of principal component bases, because this basis set is comprised of eigenvectors that are orthonormal by definition, so $\mathbf{B}^T\mathbf{B} = \mathbf{I}$, which implies that $\mathbf{B}$ is the inverse transformation and:

$$\mathbf{z}_i = \mathbf{B}\mathbf{c}_i + \boldsymbol{\mu}_{train}$$

thus giving the image vectors $\mathbf{z}_i$, representing the result, an avatar driven by the facial actions of another. In the case of ICA, the resulting basis vectors or independent components are not mutually orthonormal and form a subspace totally different from that formed by the driving vectors, thus rendering the orthogonal projection step, and consequently the performance-based mimicry, impossible. Simply put, projecting the driving vectors onto **U** wouldn't make any sense because these spaces are neither mutually orthonormal nor equivalent as they are in the case of PCA.

There is also the additional problem of component ordering. In PCA the components are ordered automatically based on their corresponding eigenvalues, starting with the one with the highest eigenvalue. This way the top few PCs encode a very high percentage of the variation in the data, thus a projection onto those first few PCs provides a good reproduction of the data. ICA doesn't automatically provide a way of ordering the resulting ICs so arbitrary measures have to be used to create some type of ordering. This means that even if projection was feasible, selecting the correct necessary ICs is a much more difficult task.

Manipulating the ICs of facial mimicries

This doesn't mean however that ICA cannot be used to improve the quality of occlusion-affected mimicries. The PCA produced mimicries could be morph-vectorised post-production and at that point ICA could be performed on them. This should generate the ICs of the mimicry which could then be manipulated by exaggerating their coefficients without the drawbacks displayed by the equivalent manipulations of PCs.

So firstly, a PCA-based performance-driven mimicry was produced using a driver with an artificially occluded mouth area. To strengthen the case, a driver of a different identity from the avatar was used (Glyn as driver, Fatos as avatar). This, as described in chapter 5, results in a mimicry displaying rather muted mouth area actions. The morph-vectorised frames of this mimicry are then processed using the ICALAB toolbox for Matlab (Cichocki et al 2003). The Fast or Fixed Point algorithm (Hyvärinen & Oja 1997) was used as standard throughout the experiments, due to its superior speed. No pre-processing was done and the mixing matrix **H** was selected to be an identity matrix, which is standard procedure for real-world data.

Results

As expected, ICA produced components that generally describe more localised facial behaviour and actions. Results looked very promising, since behaviour seemed to be encoded across the ICs, as opposed to 95% of variance usually encoded in first 4 or 5 PCs (figs.6.3, 6.4 and 6.5). The fact that individual components were now encoding more straightforward, local actions and less complex, global behaviour meant that manipulation of these components should allow for a more precise recovery of occlusion-affected facial behaviour and actions.

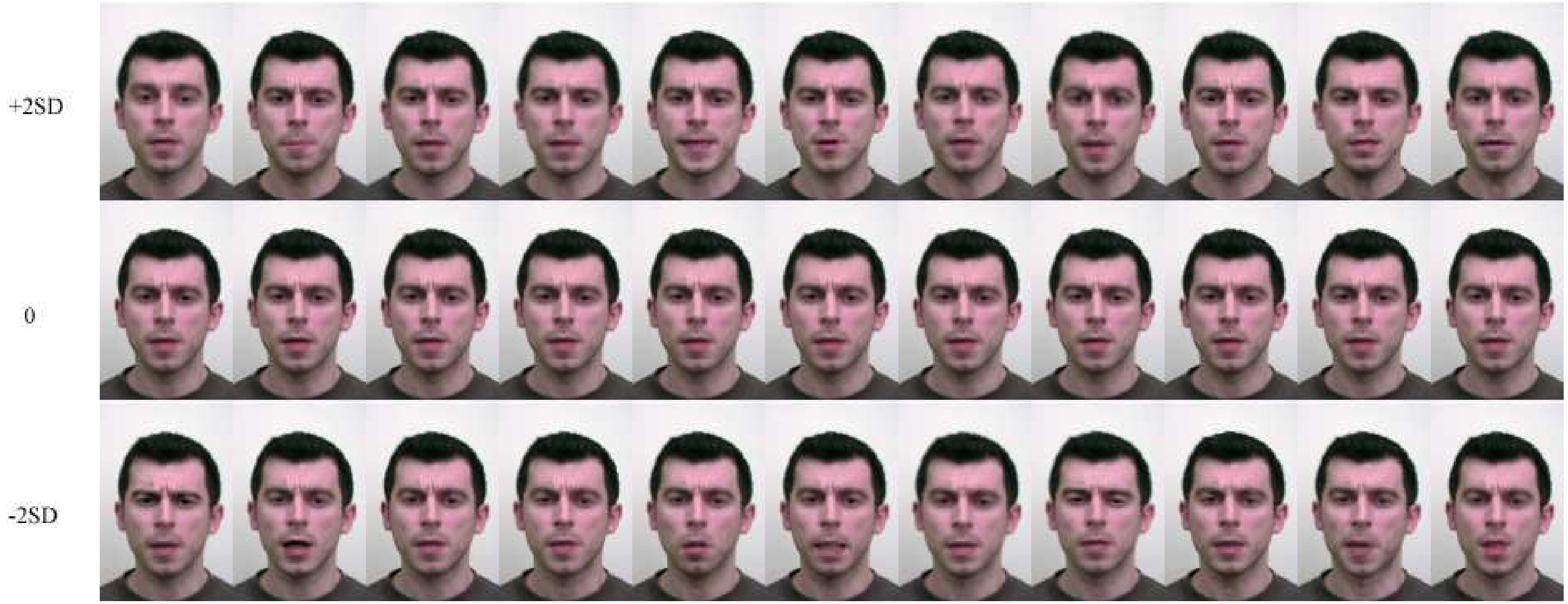


Fig 6.3 - The first 11 ICs extracted from a mouth-occlusion-affected mimicry. Middle row represents the mean while top and bottom row represent +2 and -2 standard deviations (SD) from this mean, respectively

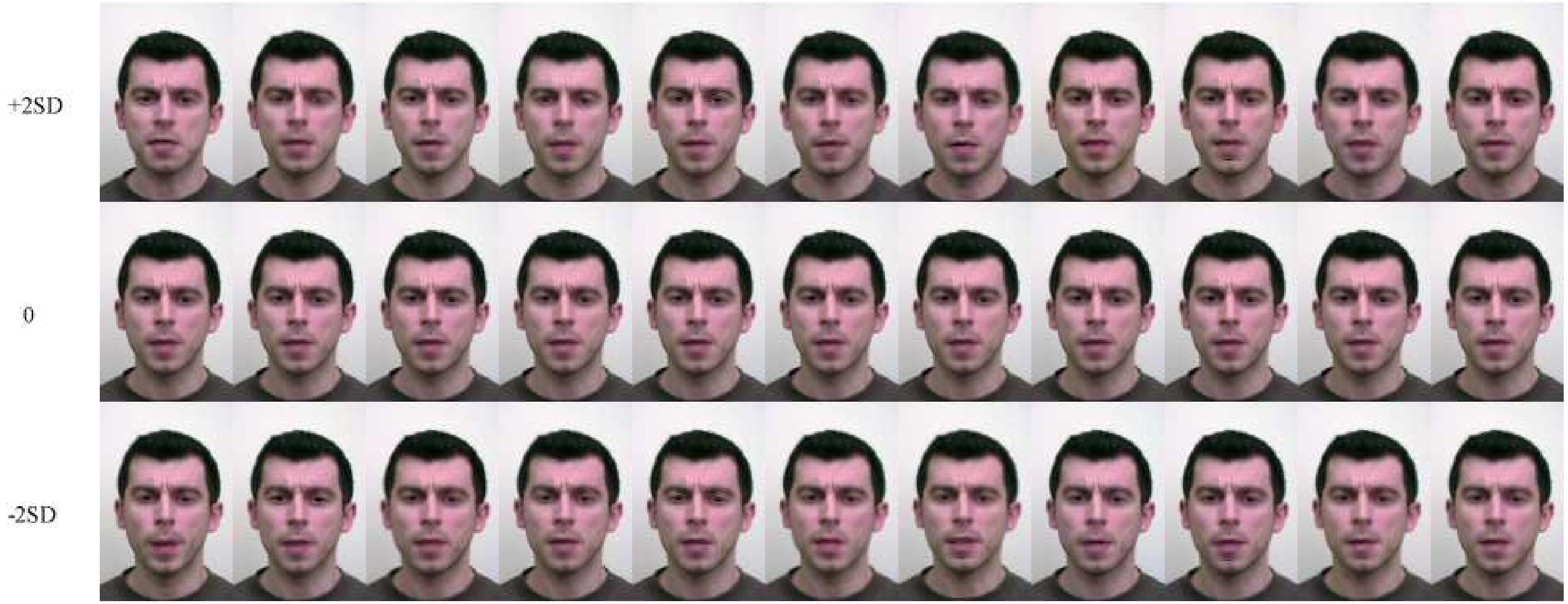


Fig 6.4 - The next 11 ICs (12 to 22) extracted from a mouth-occlusion-affected mimicry.

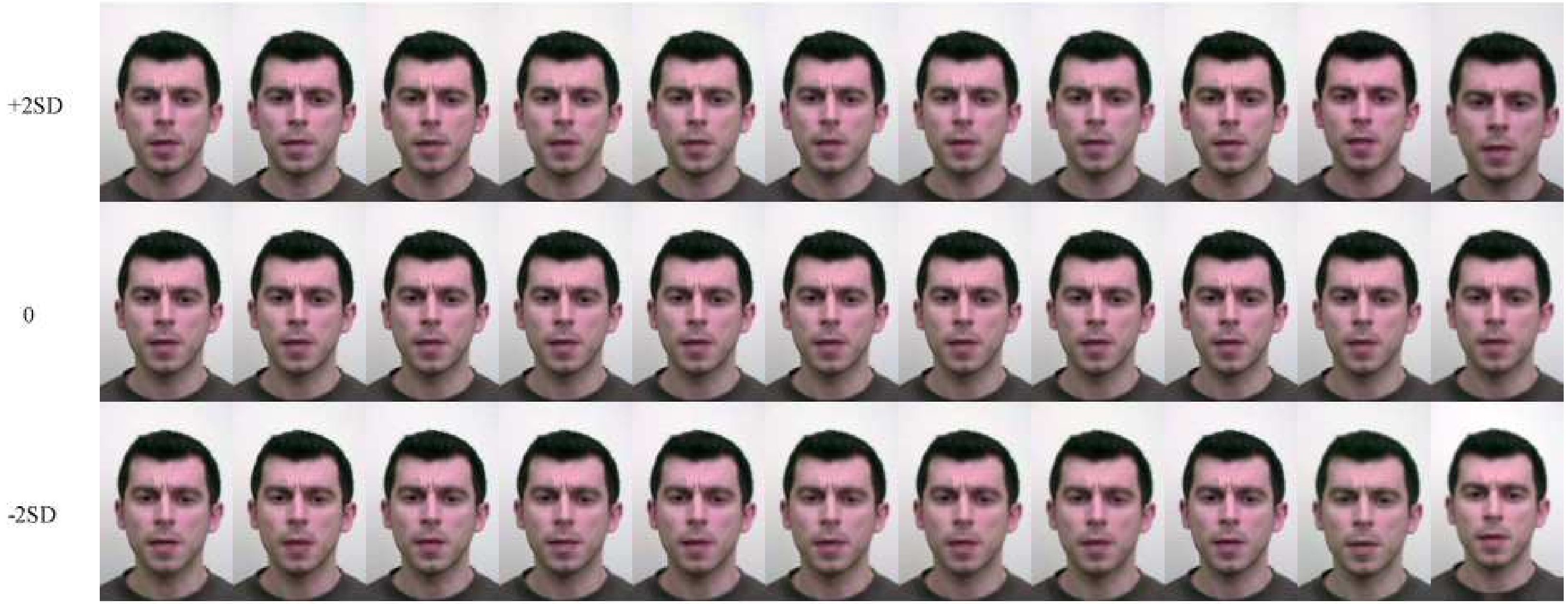


Fig 6.5 - The third set of 11 ICs (23 to 33) extracted from a mouth-occlusion-affected mimicry.

The relevant ICs (in this case, the ones that encode mouth movements) were identified by applying a complementary occlusion to the original occlusion onto the ICs and then calculating a simple SSIM value between the visible facial fragment from each of the ICs and their +2SD version. ICs with SSIM values below an arbitrary threshold were selected and scaled to produce a movie containing ICA-enhanced frames (see figure 6.6 for a diagram of the full ICA-enhancement procedure).

This scaling was done heuristically by settling on a scaling combination that produced the frames displaying smallest error as compared with the ground-truth mimicry frames. Component or basis coefficients cannot be compared this time, so a metric calculating the warp field (WF) between the respective image frames of the resulting ICA-enhanced mimicry and the ground-truth was used to evaluate the quality of the reproduction. Note that MSSIM works just as well; WF was used as a possible alternative in this experiment and because the procedure was already coded in the morph vectorisation routines used throughout this thesis.

Both visual inspection and the WF metric show good improvement in ICA-enhanced mimicries (figs.6.7 and 6.8)

The experiment was then expanded to include 10 more subjects plus Glyn as drivers, with Glyn as the avatar. This is to strengthen the case and prove that even facial behaviour that is not at all identical to that of the avatar can be quite successfully reproduced and enhanced using this ICA method.

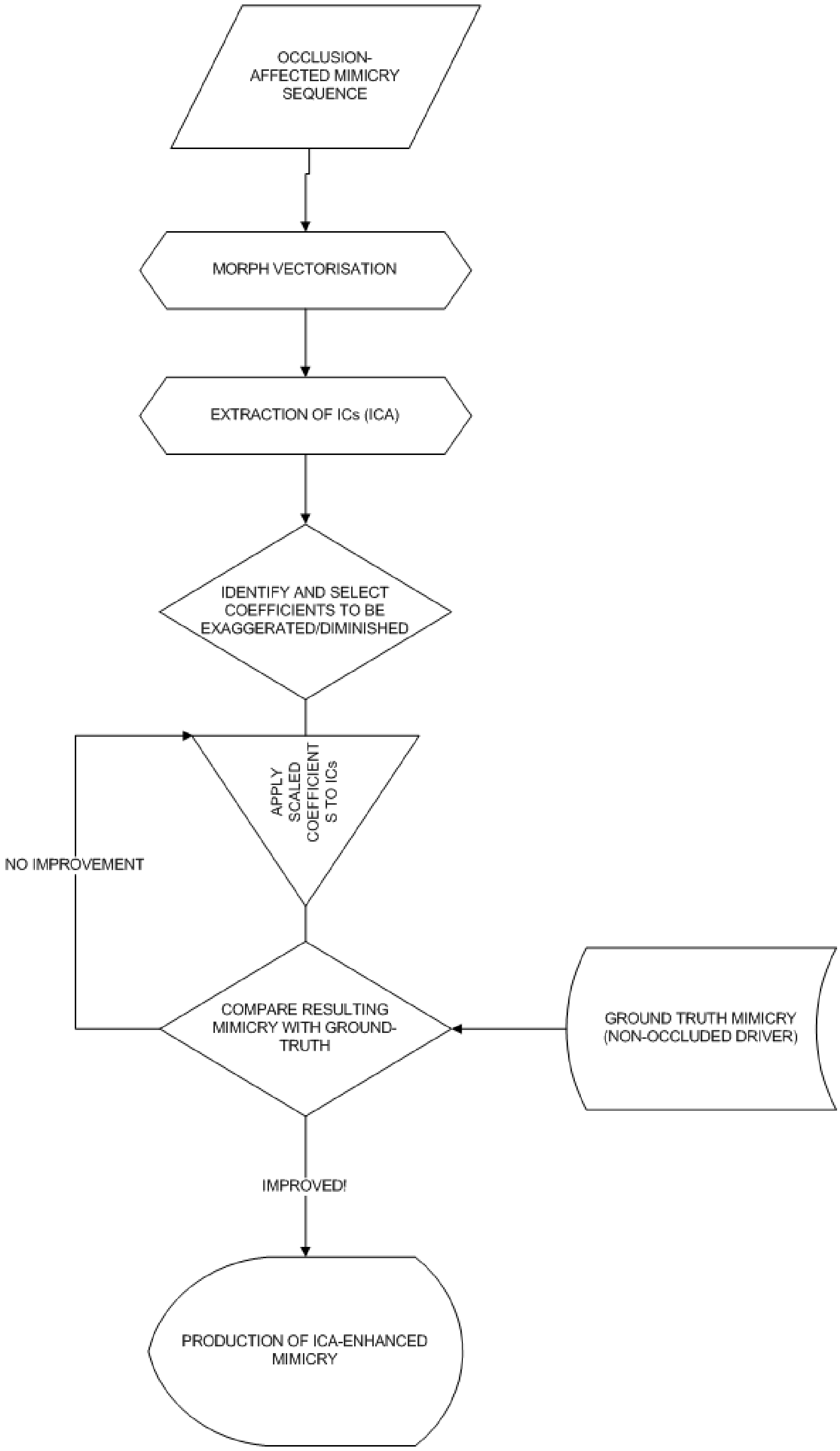

OCCLUSION-AFFECTED MIMICRY SEQUENCE
MORPH VECTORISATION
EXTRACTION OF ICs (ICA)
IDENTIFY AND SELECT COEFFICIENTS TO BE EXAGGERATED/DIMINISHED
APPLY SCALED COEFFICIENT S TO ICs
NO IMPROVEMENT
COMPARE RESULTING MIMICRY WITH GROUND-TRUTH
GROUND TRUTH MIMICRY (NON-OCCLUDED DRIVER)
IMPROVED!
PRODUCTION OF ICA-ENHANCED MIMICRY

Figure 6.6 – Block diagram of the ICA-enhancement of occlusion-affected mimicries

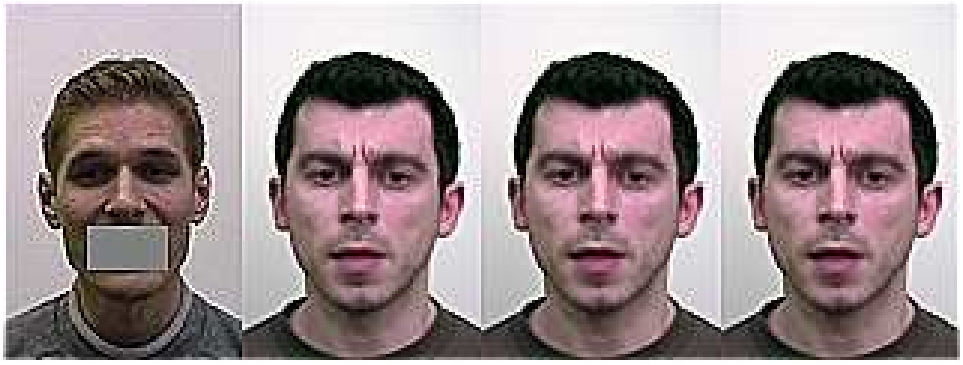


Figure 6.7 – The left-most frame displaying a mouth-occlusion is the driver used to produce the mimicry. The second frame from the left represents the PCA-based mimicry frame reproduction, without any enhancements. The third frame from the left is the ICA-enhanced reproduction while the right-most frame is the ground-truth, obtained by using a non-occluded driver. It can be just seen that the mouth area in the ICA-enhanced reproduction is more similar to the ground-truth than to the PCA-based mimicry.

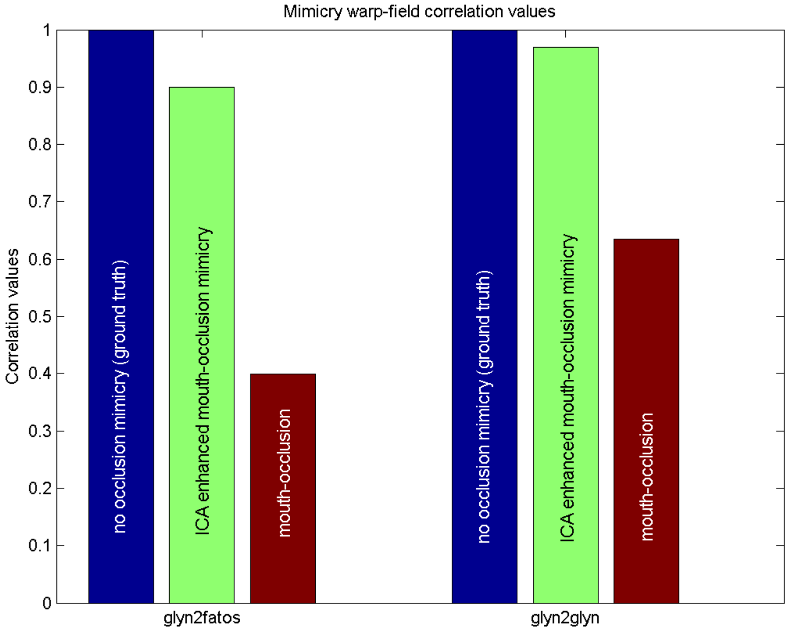


Figure 6.8 – It is clear that in both cases (Glyn driver -> Fatos avatar and Glyn driver -> Glyn avatar) ICA-enhanced mimicry does much better than a simple PCA-based mouth-occlusion driven mimicry, as compared with the ground-truth.

The results show a clear improvement when compared with the PCA-based mimicry (figs. 6.8, 6.9, 6.10). Dependent sample *t*-test also shows that the probability of obtaining such an increase in quality of mimicries by pure chance is well below the *p* = 0.05 limit (table 4). Thus we can safely conclude that the results support the experimental prediction that ICA post-processing significantly improves warp-field correlation of mimicries with the ground truth, corresponding to a perceptual increase in quality of mimicry.

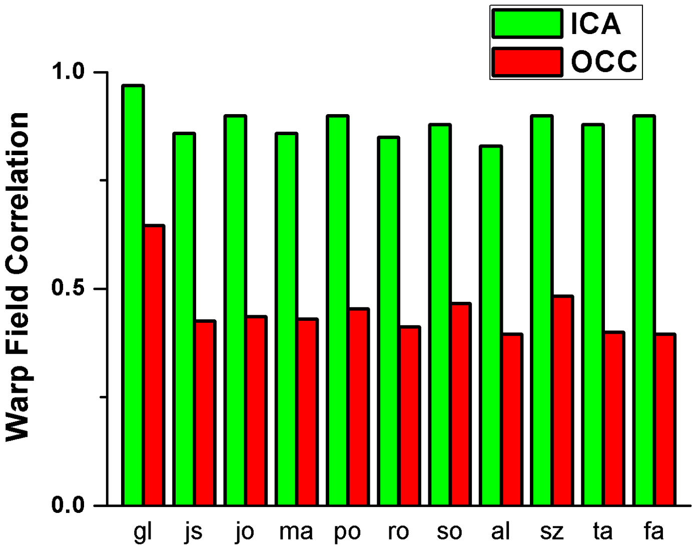


Figure 6.9 –.The green bars show the correlation as found by means of WF calculation between respective frames of the tested ICA-enhanced occlusion-affected mimicry and the ground-truth. Red bars display the correlation between the non-ICA occlusion-affected mimicry and the ground-truth.

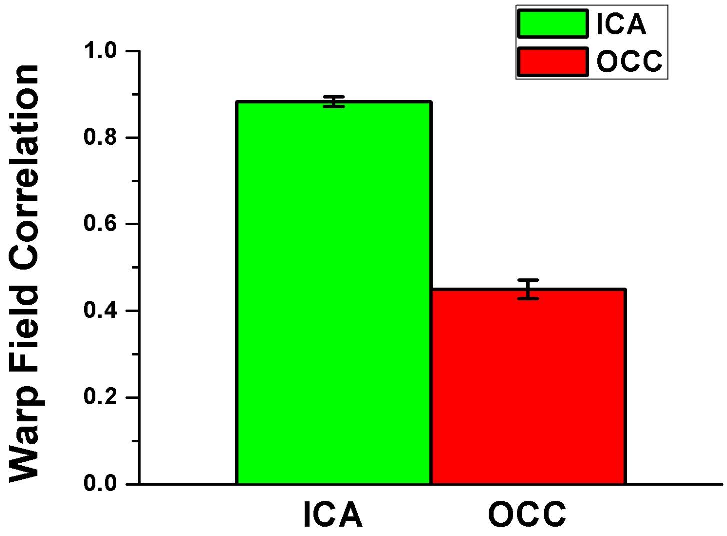


Figure 6.10– Mean values of the correlations displayed in fig.6.9.

Table 4

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	ICA	.8836227	11	3.764829E-02	.0113514
	OCC	.4494873	11	7.122080E-02	.0214739

Paired Samples Correlations

		N	Correlation	Sig.
Pair 1	ICA & OCC	11	.841	.001

Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
					95% Confidence Interval of the Difference				
		Mean	Std. Deviation	Std. Error Mean	Lower	Upper			
Pair 1	ICA - OCC	.4341355	4.448095E-02	.0134115	.4042527	.4640182	32.370	10	.000

Summary of chapter 6

Coefficient scaling of PCs was shown to cause some unwanted side-effects in the attempts to enhance muted occlusion-affected mimicries. This is because PCA tends to describe more global changes, thus grouping together basic facial actions that occur together during normal facial behaviour. PCA seems to pick this global correlated activity, with the added benefits of natural ordering and orthogonality. ICA on the other hand encodes into its components more local actions that are mutually independent, even though in reality a lot of these movements are correlated and go together.

So ICA seems very suitable for “adding manual strings to the puppet” that can clearly enhance affected mimicries, but it was shown in this chapter that it is not suitable at all for purely automatic, performance-driven mimicry. This is due to two factors that are intrinsic to the way ICA works. The first one is the fact that ICs are not mutually orthonormal, which given the length of the vectors makes it virtually impossible to project onto this subspace and extract the weights needed to drive the mimicry. The second factor is the lack of a natural ordering of the ICs. This makes it difficult to keep and discard important ICs in a principled way.

However, because behaviour seems to be encoded across the ICs in a rather uniform way, as opposed to having 95% of variance encoded in first few PCs, ICA is as a much more powerful tool in post-processing enhancements of PCA-based mimicries. The ability to localise actions meant this method could be used successfully to enhance occlusion-affected mimicries. The ICA-enhanced reproductions are of high quality and far superior than the simple occlusion affected PCA-based mimicry. This claim is backed up by both visual inspection and warp field calculation against the non-occluded ground-truth mimicry. The results prove that this method is a good candidate for post-processing enhancement of occlusion-affected mimicries.

Chapter 7

CONCLUSIONS

Summary of thesis

The human face is a complex organ of communication that is used to transmit a wide spectrum of social signals and messages. Its structure is specialised and the signals it conveys can contribute to perceptions of the bearer's facial identity, expression, sex and more. The perception of these signals is achieved by a system with multiplicity of components processing separate signals, and there is ample evidence to suggest that both the component processes and the location of early processes of face recognition (FFA) are face-specific. Psychophysical and neuropsychological experiments also seem to suggest that while object recognition is analytic and part-based, face perception and recognition is holistic and configural. Looking at other psychophysical evidence, it appears that a viewpoint or lighting invariant representation of faces is not present in our visual system (although possibly size-invariant representations are), and that most importantly, we seem to store faces in a two-dimensional manner. Thus a two-dimensional, image-based approach could be very effective in encoding facial identity, expressions and sex.

Due to the fact that the human face is essentially a 3D structure, a number of 3D facial representations and computational models have been used in realistic head and face modelling, like polygonal mesh, implicit, and parametric representations of the surface of the face. Volume representations were also employed, together with the more complex dynamic representations of facial muscles. Advanced morphable 3D models of Blanz and Vetter also produce good results and show a lot of promise in the context of face modelling. However, the greatest drawback of this class of models is that they require huge resources, be it in computational or human terms, to achieve any sufficiently compelling results to fool the human visual system. By considering the internal processes involved in the perception of faces, a more practical class of representations is introduced that takes into account the evidence which suggests that our visual system seems to process faces in a two-dimensional manner. A more global, configural approach is proposed, one that chooses the parameters to account for essential sources of variance, i.e. PCA. This statistical modelling approach and the biological motivations for its use in this context are reviewed.

In chapter 3, a method for automatically creating computer-generated avatars was described. The model can be driven by real actors, simply by aligning features and projecting vectorised sequences of their motion into the target space. The target space itself is nothing but a basis set extracted by performing PCA on a video footage of the target face in action. Novel footage of the computer generated avatar can thus be produced, mimicking the driving actor's movements, all in reasonable amount of time despite quite high computational costs of the generating and driving procedures. Resulting animations are confined to vary as a linear combination of movements from the example set, so the generated footage is realistic, seemingly without ever drifting into the "uncanny valley" (Mori 1970). This confinement of the set of possible animation elements may look like a limitation, but in fact it is largely advantageous in preventing the avatar from doing anything that the original face was incapable of doing, this possibly being the very factor that prevents the "uncanny valley" effect from occurring here. And provided a sufficiently rich set of motion is captured for the generation of the model, these constraints do not pose a problem. The coefficients for a sequence can be transformed in the target face space, in order to exaggerate, or rescale movements to be consistent with the example footage. This is a useful processing step in conditions where the facial geometry is such that the novel vectors do not project strongly onto the target basis set, such as for example in the case of a driver with occluded facial regions.

Our ability to process faces is known to depend on the spatial location of visual facial information we receive. A good method for revealing such diagnostic facial information for different categorisation tasks is the "bubbles" method. Here it succeeds in revealing diagnostic information for a performance-driven mimicry task carried out by a computer model of the face, built to a degree on biologically motivated principles. The face model was generated by first vectorising a sequence of images of a talking face affected by a total occlusion interspersed with a number of randomly spaced Gaussian windows and then by extracting motion fields via an optic flow algorithm and calculating a set of basis actions using principal component analysis. This standard bubbles technique revealed the areas around and including the mouth and eyes as the most important ones for our task. These regions overlapped with but were not identical to areas of maximum pixel-value variance. Visual inspection also showed that the PCA face model recovers aspects of expressions in those areas occluded in the driver sequence. Until now bubbles were only used as a human search for diagnostic features in faces. Here, a system using reconstruction fidelity as diagnostic criterion and indifferent to the content of the stimulus, mimics the behaviour of human observers in face discrimination tasks. This information could be also very useful in further analyzing and retrieving non-randomly occluded facial information.

Effects of non-systematic facial appearance variation such as partial occlusions on face perception and analysis have not been studied to the same extent as the effects caused by changes in viewpoint, expression or illumination direction and intensity. But most studies of facial occlusion have tended to make use of PCA and the information contained in the eigenface subspace to compensate for the details lost due to partial occlusions of the face. This property was demonstrated using the PCA-based mimicry system in chapter 4, where occlusions were randomised. In this chapter, a database of video sequences affected by non-random occlusions (artificial, natural and dynamic) was created. These sequences were used to drive a PCA-build avatar and the mimicries obtained in this way were compared to the ground-truth mimicry which in turn was produced using a non-occluded driver. Visual inspection of the resulting mimicries results show that the motion information from the occluded areas can indeed be correctly reproduced by the system, albeit in a weaker and somewhat muted form as compared to the ground-truth (the occluded mouth-area results were used throughout as a representative case in this chapter, mainly due to the superior complexity of behaviour stemming from this area and its importance in human communication and interaction). To amplify the muted reproductions of facial behaviour, the facial actions were exaggerated through principal component coefficient scaling by arbitrary scalars. Visual inspection again showed that certain factors did indeed appear to increase the similarity of the occlusion-affected mimicries as compared to the ground-truth. In order to provide an objective evaluation and a basis for an optimization of the model, a number of performance metrics were suggested. Simple error quantisation metrics such as mean error (frame luminance difference) and standard deviation error (frame contrast difference) didn't seem to correspond well to visual inspection results due to their inability to take into account any structural information with the images compared. However, a metric - suggested by Wang et al. and named Structural Similarity Index (SSIM) - that combined these two measures together with a normalisation component did indeed succeed in predicting perceived image, and consequently mimicry, quality in all occlusion cases.

All improvements by coefficient scaling described in chapter 5 were achieved by arbitrary and heuristic scaling of the whole matrix of coefficients. The side-effect of such an approach was an anomalous exaggeration of the other facial areas, which became rather like caricatured animation when the multiplier is too high.

To combat these shortcomings other ways of manipulating specific PC coefficients, together with other methods of component analyses, are developed in chapter 6. ICA represents a generalised form of PCA that encodes into its components more local actions that are mutually independent, even though in reality a lot of these movements are correlated and go together. It is a very suitable method for "adding manual strings to the puppet" that can clearly enhance affected mimicries, but it is not suitable at all for purely automatic, performance-driven mimicry due to two factors that are intrinsic to the way ICA works. The first one is the fact that ICs are not mutually orthonormal, which given the length of the vectors makes it virtually impossible to project onto this subspace and extract the weights needed to drive the mimicry. The second factor is the lack of a natural ordering of the ICs. This makes it difficult to keep and discard important ICs in a principled way. Nevertheless, because behaviour seems to be encoded across the ICs in a rather uniform way, ICA is as a much more powerful tool in post-processing enhancements of PCA-based mimicries. The ability to localise actions meant this method could be used successfully to enhance occlusion-affected mimicries and ICA-enhanced reproductions are of high quality and far superior than the simple occlusion affected PCA-based mimicry. This claim is backed up by both visual inspection and warp field calculation against the non-occluded ground-truth mimicry. The results prove that this method is a good candidate for post-processing enhancement of occlusion-affected mimicries.

Discussion of contributions

Most current methods of animating a face from an actor's movements have focussed mostly on modelling the face as a 3D polygonal surface. Realism is then attained through many hours of work by a talented artist, or through capturing the geometry of a real face with equipment such as a laser scanner. A complex underlying muscle model is then usually added, in order to make the model moveable and a lot of skill is still required on the part of the operators of such models to achieve photorealism. Still, with addition of movement and behaviour such synthetic models begin to usually look very unrealistic and less than natural, rolling quickly into Mori's "uncanny valley". The Cowe PCA-based face mimicry model used in this work discards the complexities of 3D and adopts a simpler 2D representation, which could also be a closer match with our own brain's representation of faces. With this tool photorealism is achieved automatically because the resulting animations are confined to vary as a linear combination of original movements from the example set, so the generated footage is always realistic.

PCA was shown to be a form of auto-associative memory (Valentin et al 1994). A few years previous to that, Cottrell demonstrated that in auto-associative networks a whole face can be recovered from a partial, static facial input (1990). Motivated by these findings, the PCA-based mimicry model was used for the first time to attempt photorealistic reconstruction of missing dynamic facial information from occlusion-affected faces. This task was successfully achieved for the same reasons Cowe's non-occlusion PCA-based mimicry was successful - because again the reconstruction is essentially a linear combination of the shape and texture vectors on which PCA was performed. The PCA eigenvectors are linear combinations of the original data. So given that the occlusion is generally distant from the portion of image space spanned by the PCs, the projection of the face vectors onto the component axes will be dominated by the face portions of the vector, and will reconstruct images and actions that are very similar to the images and actions of the original face.

Similarity, or reconstruction quality, depends on a number of factors. Visual inspection showed that some of these factors, such as the size and especially the position of the face occlusion, are very important for PCA-based mimicry. In this thesis, a novel experiment was designed that employs a principled method previously only used with human subjects as evaluators. The task of this experiment was to locate the pertinent areas of the face for high fidelity mimicry. This *bubbles* experiment in an ideal-observer, in-silico set-up was successful in mapping out areas of the face that convey the best information used in automatic PCA mimicry, with reproduction fidelity as a simple diagnostic criterion for comparison with the ground-truth mimicry. These areas convey the best information about global facial configuration. This fact was confirmed when mimicries created with dynamic driving information emanating only from these areas of the face produced very high quality mimicries. The high quality of these reproductions was observed by way of visual inspection, but also by using a Pearson correlation metric to compare the set of respective PC coefficients used in the generation of those same reproductions vs the ground-truth coefficients.

Scaling the same PC coefficients by a constant can increase or decrease the amplitude of facial actions in the mimicry. This thesis manages to use this scaling method to improve occlusion affected mimicries. It also carries out a detailed analysis to identify an image-based metric that can be used with dynamic facial image and action data and reliably evaluate the quality of the mimicries.

Scaling PCs though can cause a lot of unwanted effects and artefacts in the produced mimicries. This is because PCs encode global facial behaviour, thus their scaling too will affect the reproduced facial behaviour globally. In order to overcome this drawback, this thesis provides a novel way of using

ICA to enhance or modify facial actions recorded in video footage. The idea involves vectorising the shape and texture information of all the sequence frames and then extracting a set of statistically independent components from that dataset. These components encode local facial behaviour and thus are more suitable for applying specific modifications of the image/sequence. They were then used as a novel method for enhancing generally all types of footage but specifically and most importantly for this thesis, enhancing muted occlusion-affected mimicries without causing unwanted side-effect in the reconstruction of facial behaviour. A new metric based on calculations of the warp field between two corresponding images was proposed and successfully used. As a final result, it was shown that this method, post-processing ICA-enhancement of footage, can be successful in improving occlusion-affected mimicries by heuristic manipulation of the IC coefficients.

Future work and applications

In the future, a number of of opportunities is available for testing more powerful and versatile vectorisations, such as encoding multiple camera views into one long vector. This type of vectorisation could help the PCA-model dealing better with the types of dynamic occlusions like a moving hand or a head turning from side to side. As mentioned in chapter 5, such dynamic iconic changes cause the system to fail and produce mimicries that are rather flawed. A vector containing 3D info would offer a solution to such problems.

The procedure of coefficient scaling is currently heuristic, i.e. a user selects arbitrary scaling values for the relevant IC (or PC) coefficients according to the visually assessed quality of the mimicries. An automated procedure could be implemented that deploys possibly a gradient-descent-type algorithm with a SSIM cost function in order to arrive at an optimal combination of scaling values, producing thus even better mimicry reproductions in a much more principled and efficient way.

Overall, the combination of performance-driven PCA-based mimicry and ICA-enhancements produces good results in the retrieval of occluded facial information. The technology can be seen, in its present form, to have the capabilities to improve animation quality, and speed up an expensive process in the film and entertainment or computer games industry. Two-dimensional models may not be appropriate for the current needs of these industries, but the performance-driven techniques could be used to animate handcrafted models.

Facial animation in these industries, however, is considered to be an artistic activity and animators are wary of full automation. An automated first-pass animation could provide time savings, but it would not be easy introducing such a product into an industry with such a strong culture of manual production. Where the ICA-enhancement manual methods could come in handy is in post-production roles. It could be possible to extract ICs from any type of footage, sound or a combination thereof, thus offering endless possibilities of further manipulation.

Together with the general findings about the structural importance of the face for photorealistic mimicry this thesis could provide a set of tools and information that could find applications also security, telecommunications and image processing software industries. Finally, there could be applications in psychological research, for instance to create and manipulate facial stimuli in order to generate specific facial configurations and expressions.

References

Anaki D, Zion-Golumbic E, Bentin S. 2007. Electrophysiological neural mechanisms for detection, configural analysis and recognition of faces. *Neuroimage* 37:1407 - 16

Arslan LM, Talkin D. 1998. 3-D Face Point Trajectory Synthesis Using An Automatically Derived Visual Phoneme Similarity Matrix. *Auditory-Visual Speech Processing Conference*, pp. 175-80. Terrigal, NSW, Australia

Barron J, Fleet D, Beauchemin S, Burkitt T. 1992. Performance of optical flow techniques. *Proceedings of the IEEE on Computer Vision and Pattern Recognition*, pp. 236-42

Bassili JN. 1978. Facial motion in the perception of faces and of emotional expression. *Journal of experimental Psychology: Human Perception and Performance* 4:373-9

Bassili JN. 1979. Emotion recognition: The role of facial movement and the relative importance of upper and lower areas of the face. *Journal of Personality and Social Psychology* 37:2049-58

Bell A, Sejnowski TJ. 1995. An information-maximization approach to blind separation and blind deconvolution. *Neural Computation* 7:1129-59

Bentin S, Taylor MJ, Rousselet GA, Itier RJ, Caldara R, et al. 2007. Controlling interstimulus perceptual variance does not abolish N170 face sensitivity. *Nature Neuroscience* 10:802 - 3

Benton C, Jennings S, Chatting D. 2006. Viewpoint dependence in adaptation to facial identity. *Vision Research* 46:3313-25

Beymer D, Shashua A, Poggio T. 1993. Example-based image analysis and synthesis. ed. M Artificial Intelligence Laboratory. Cambridge, Massachusetts

Blanz V, Basso C, Poggio T, Vetter T. 2003. Reanimating Faces in Images and Video. *Computer Graphics Forum* 22:641 - 50

Blanz V, Grother P, Phillips PJ, Vetter T. 2005. Face Recognition Based on Frontal Views Generated from Non-Frontal Images. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05).* San Diego, CA, USA

Blanz V, Romdhani S, Vetter T. 2002. Face identification across different poses and illuminations with a 3D morphable model. *Proceedings of the Fifth IEEE International Conference onAutomatic Face and Gesture Recognition, 2002*, pp. 192 - 7. Washington, DC, USA

Blanz V, Vetter T. 1999. A morphable model for the synthesis of 3D faces. *Proceedings of the 26th annual conference on Computer Graphics and interactive techniques*, pp. 187-94

Blanz V, Vetter T. 2003. Face recognition based on fitting a 3D morphable model. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 25:1063 - 74

Bodamer J. 1947. Die Prosopagnosie. *Archiv für Psychiatrie und Nervenkrankheiten* 179:6-53

Bonnar L, Gosselin F, Schyns PG. 2002. Understanding Dali's Slave Market with the Disappearing Bust of Voltaire: A case study in the scale information driving perception. *Perception* 31:683-91

Bradshaw JL, Taylor MJ, Patterson K, Nettleton NC. 1980. Upright and inverted faces, and housefronts in two visual fields. *Journal of Clinical Neuropsychology* 2:245-57

Bruce V, Valentine T, eds. 1988. *When a nod's as good as a wink: The role of dynamic information in face recognition*, Vols. 1. Chichester: Wiley

Bruce V, Young A. 1986. Understanding Face Recognition. *British Journal of Psychology* 77:305-27

Buchan JN, Par, M, Munhall KG. 2007. Spatial statistics of gaze fixations during dynamic face processing. *Social Neuroscience* 2:1-13

Burke D, Taubert J, Higman T. 2007. Are face representations viewpoint dependent? A stereo advantage for generalising across different views of faces. *Vision Research* 47:2164-9

Calder AJ, Burton MA, Miller P, Young A, Akamatsu S. 2001. A principal component analysis of facial expression. *Vision Research* 41:1179-208

Cichocki A, Amari S, Siwek K, Tanaka T, Phan AH. 2003. ICALAB Toolboxes.

Comon P. 1994. Independent component analysis - a new concept? *Signal Processing* 36:287-314

Cootes TF, Edwards GJ, Taylor CJ. 1998. Active Appearance Models. In *Computer Vision - ECCV'98*, ed. H Burkhardt, B Neumann, pp. 484 - 98. Freiburg, Germany: Springer-Verlag Berlin / Heidelberg

Cottrell GW. 1990. Extracting features from faces using compression networks: Face, identity, emotion, and gender recognition using holons. *Connectionist Models Summer School*. San Diego, California, USA: Morgan Kaufmann Publishers

Cowe G. 2003. *Example-based computer-generated facial mimicry*. PhD Thesis. UCL, London

Curio C, Giese MA, Breidt M, Kleiner M, Bulthoff H. 2007. High-level after-effects in the recognition of dynamic facial expressions. *Visual Sciences Society 7th Annual Meeting 2007 Abstracts*, p. 283

Darwin C. 1872. *The expression of the emotions in man and animal*. London: John Murray

Diamond R, Carey S. 1986. Why faces are and are not special: An effect of expertise. *Journal of Experimental Psychology: General* 115:107-17

Ekman P, Friesen WV. 1975. *Unmasking the face. A guide to recognising emotions from facial clues*. Englewood Cliffs, New Jersey: Prentice-Hall

Ekman P, Friesen WV. 1976. *Pictures of facial effect*. Palo Alto, California: Consulting Psychologists Press

Ekman P, Friesen WV. 1978. *Facial action coding system: A technique for the measurement of facial movement*. Palo Alto, Calif.: Consulting Psychologists Press

Essa I, Pentland AP. 1997. Coding, analysis, interpretation and recognition of facial expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 19:757-63

Fang F, He S. 2005. Viewer-centred object representation in the human visual system revealed by viewpoint aftereffects. *Neuron* 45:793-800

Farah MJ, Rabinowitz C, Quinn GE, Liu GT. 2000. Early commitment of neural substrates for face recognition. *Cognitive Neuropsychology* 17:117-23

Farah MJ, Wilson KD, Drain M, Tanaka JN. 1998. What is "special" about face perception? *Psychological Review* 105:482-98

Fidaleo D, Neumann U. 2002. CoArt: Co-articulation region analysis for control of 2D characters. In *IEEE Computer Animation Conference*. Geneva, Switzerland

Fleet DJ, Black MJ, Yacoob Y, Jepson AD. 2000. Design and Use of Linear Models for Image Motion Analysis. *International Journal of Computer Vision* 36:171-93

Flevaris AV, Robertson LC, Bentin S. 2008. Using spatial frequency scales for processing face features and face configuration: an ERP analysis. *Brain Research* 1194:100 - 9

Forsey DR. 1990. *Motion control and surface modelling of articulated figures in computer animation*. PhD Thesis. University of Waterloo, Waterloo

Fraser I, Parker D. 1986. Reaction time measures of feature saliency in a perceptual integration task. In *Aspects of face processing*, ed. HD Ellis, MA Jeeves, F Newcombe, A Young. Dordrecht: Martinus Nijhoff

Gauthier L, Logothetis NK. 2000. Is face recognition not so unique after all? *Cognitive Neuropsychology* 17:125-42

Gauthier L, Tarr MJ. 2002. Unraveling mechanisms for expert recognition: Bridging brain activity and behavior. *Journal of experimental Psychology: Human Perception and Performance* 28:431-46

Gauthier L, Williams P, Tarr MJ, Tanaka JN. 1998. Training "greeble" experts: a framework for studying expert object recognition processes. *Vision Research* 38:2401-28

Gibson BM, Wasserman EA, Gosselin F, Schyns PG. 2005. Applying bubbles to localize features that control pigeons' visual discrimination behavior. *Journal of Experimental Psychology-Animal Behavior Processes* 31:376-82

Giese MA, Leopold DA. 2004. Physiologically inspired neural model for the encoding of face spaces. *Neurocomputing* 65-66:93-101

Giese MA, Sigala R, Wallraven C, Leopold DA. 2004. Psychologically inspired neural model for the prototype-referenced encoding of faces. *Journal of Vision* 4:213a

Gosselin F, Schyns PG. 2001. Bubbles: a technique to reveal the use of information in recognition tasks. *Vision Research* 41:2261-71

Grill-Spector K, Knouf N, Kanwisher N. 2004. The fusiform face area subserves face perception, not generic within-category identification. *Nature Neuroscience* 7:555-62

Gross R, Matthews I, Baker S. 2006. Active appearance models with occlusion. *Image and Vision Computing* 24:593-604

Hancock PJB, Baddely RJ, Smith LS. 1992. The principal components of natural images. *Network: Computation in Neural Systems* 3:61-70

Hasselmo ME, Rolls ET, Baylis GC, Nalwa V. 1989. Object-centred encoding by face selective neurones in the cortex in the superior temporal sulcus of the monkey. *Experimental Brain Research* 75:417-29

Haxby JV, Ungerleider LG, Clark VP, Schouten JL, Hoffman EA, Martin A. 1999. The effect of face inversion on activity in human neural systems for face and object perception. *Neuron* 22:189-99

Hietanen JK, Perret DI, Oram MW, Benson PJ, Dittrich WH. 1992. The effects of lighting conditions on responses of cells selective for face views in the macaque temporal cortex. *Experimental Brain Research* 89:157-71

Hill HCH, Bruce V. 1993. Independent effects of lighting, orientation, and stereopsis on the hollow face illusion. *Perception* 22:887-97

Hill HCH, Bruce V. 1994. A comparison between the hollow-face and "hollow-potato" illusions. *Perception* 23:1335-7

Hill HCH, Bruce V. 1996. Effect of lighting on the perception of facial surfaces. *Journal of Experimental Psychology* 22:986-1004

Hill HCH, Johnston A. 2001. Categorising sex and identity from the biological motion of faces. *Current Biology* 11:880-5

Hill HCH, Schyns PG, Akamatsu S. 1997. Information and viewpoint dependence in face recognition. *Cognition* 62:201-22

Humphreys K, Gosselin F, Schyns PG, Johnson MH. 2006. Using "Bubbles" with babies: A new technique for investigating the informational basis of infant perception. *Infant Behavior & Development* 29:471-5

Hyvärinen A, Oja E. 1997. A Fast Fixed-Point Algorithm for Independent Component Analysis. *Neural Computation* 9:1483-92

Jing X, Baker S, Matthews I, Kanade T. 2004. Real-time combined 2D+3D active appearance models. *IEEE Computer Science Conference on Computer Vision and Pattern Recognition (CVPR'04)*, pp. 535 - 42. Washington, DC, USA

Johansson G. 1973. Visual perceptionof biological motion and a model for its analysis. *Perception and Psychophysics* 14:201-11

Johnston A, Hill HCH, Carman N. 1990. Recognising faces: effects of lighting, direction, inversion and brightness reversal. *Perception* 21:365-75

Johnston A, McOwan PW, Benton CP. 1999. Robust velocity computation from a biologically motivated model of motor perception. *Proceedings of the Royal Society of London, B* 266:509-18

Joliffe IT. 1986. *Principal Component Analysis*. New York: Springer-Verlag

Kanwisher N, McDermott J, Chun MM. 1997. The fusiform face area: A module in human extrastriate cortex specialised for face perception. *Journal of Neuroscience* 17:4302-11

Knappmeyer B, Thornton IM, Bulthoff H. 2003. The use of facial motion and facial form during the processing of identity. *Vision Research* 43:1921-36

Knight B, Johnston A. 1997. The role of movement in face recognition. *Visual Cognition* 4:265-73

Kshirsagar S, Molet T, Magnenat-Thalmann N. 2001. Principal Components of Expressive Speech Animation. *Proceedings Computer Graphics International 2001, IEEE Computer Society*, pp. 38-44

Lander K, Christie F, Bruce V. 1999. The role of movement in the recognition of famous faces. *Memory and Cognition* 27:974-85

Lanitis A. 2004. Person identification from heavily occluded face images. In *Proceedings of the 2004 ACM symposium on Applied computing*. Nicosia, Cyprus: ACM

Lee Y, Matsumiya K, Wilson HR. 2006. Size-invariant but viewpoint-dependent representation of faces. *Vision Research* 46:1901-10

Leopold DA, Bondar IV, Giese MA. 2006. Norm-based face encoding by single neurons in the monkey inferotemporal cortex. *Nature* 442:572-5

Leopold DA, O'Toole AJ, Vetter T, Blanz V. 2001. Prototype-referenced shape encoding revealed by high level after effects. *Nature Neuroscience* 4:89-94

Leopold DA, Rhodes G, Muller KM, Jeffery L. 2005. The dynamics of visual adaptation to faces. *Proceedings of the Royal Society B-Biological Sciences* 272:897-904

Lewicki MS, Sejnowski TJ. 2000. Learning overcomplete representations. *Neural Computation* 12:337-65

Liu CH, Collin CA, Burton AM, Chaudhuri A. 1999. Lighting direction affects recognition of untextured faces in photographic positive and negative. *Vision Research* 39:4003-9

Marr D. 1982. *Vision*. San Francisco: W. H. Freeman & Company

Martinez AM. 2002. Recognizing Imprecisely Localized, Partially Occluded, and Expression Variant Faces from a Single Sample per Class. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 24:748-63

Mase K. 1991. Recognition of facial expressions for optical flow. *IEICE Transactions, Special Issue on Computer Vision and ifs Applications* E 74

Mase K, Pentland AP. 1991. Lipreading by optical flow. *Systems and Computers* 22:67-76

Maurer D, LeGrand R, Mondloch CJ. 2002. The many faces of configural processing. *Trends in Cognitive Neurosciences* 6:255-60

McCotter M, Gosselin F, Sowden P, Schyns P. 2005. The use of visual information in natural scenes. *Visual Cognition* 12:938-53

McKone E, Kanwisher N, Duchaine BC. 2007. Can generic expertise explain special processing for faces? *Trends in Cognitive Sciences* 11:8-15

Mori M. 1970. The Uncanny Valley. *Energy* 7:33-5

Moscovitch M, Moscovitch DA. 2000. Super face-inversion effects for isolated internal or external features, and fractured faces. *Cognitive Neuropsychology* 17:201-19

O'Toole AJ, Deffenbacher KA, Valentin D, Abdi H. 1994. Structural aspects of face recognition and the other race effect. *Memory and Cognition* 22:208-24

O'Toole AJ, Roark DA, Abdi H. 2002. Recognizing moving faces: a psychological and neural synthesis. *Trends in Cognitive Sciences* 6:261-6

Parke FI. 1972. *Computer generated animation of faces.* Masters Thesis. University of Utah, Salt Lake City

Pearson K. 1901. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine* 2:559-72

Perret DI, Oram MW, Harries MH, Bevan JK, Hietanen JK, et al. 1991. Viewer-centred and object-centred coding of heads in the macaque temporal cortex. *Experimental Brain Research* 86:159-73

Perret DI, Smith PA, Potter DD, Mistlin AJ, Head AS, et al. 1985. Visual cells in the temporal cortex sensitive to face view and gaze direction. *Proceedings of the Royal Society of London, B* 223:293-317

Peterson M, Abbey C, Eckstein M. 2007. Information distribution for face identification and its relation to human strategies. *Visual Sciences Society 7th Annual Meeting 2007 Abstracts*, p. 53

Phelps MT, Roberts WA. 1994. Memory for pictures of upright and inverted primate faces in humans (Homo sapiens), squirrel monkeys (Saimiri sciureus), and pigeons (Columba livia). *Journal of Comparative Psychology* 108:114-25

Pighin F, Hecker J, Lischinski E, Szeliski R, Salesin DH. 1998. Synthesizing realistic facial expresions from photographs. In *Computer Graphics, Annual Conference Series*, pp. 75-84: Siggraph

Platt SM, Badler NI. 1981. Animating facial expression. *ACM SIGGRAPH Conference on Computer Graphics*. Dallas, Texas

Rinn WE. 1984. The neuropsychology of facial expression: A review of the neurological and psychological mechanisms for producing facial expressions. *Psychological Bulletin* 95:52-77

Rosch E, Mervis CB, Gray WD, Johnson DM, Boyesbraem P. 1976. Basic Objects in Natural Categories. *Cognitive Psychology* 8:382-439

Roy-Chowdhury A, Chellappa R. 2003. Stochastic approximation and rate-distortion analysis for robust structure and motion estimation. *The International Journal of Computer Vision* 55:27-53

Schyns P, Petro LS, Smith ML. 2007. Dynamics of visual information integration in the brain for categorizing facial expressions. *Current Biology* 17:1580 - 5

Shepherd JW, Davies GM, Ellis HD, eds. 1981. *Studies of cue saliency*. London: Academic Press

Simoncelli EP. 1993. *Distributed representation and analysis of visual motion*. PhD Thesis. MIT, Cambridge, Massachusetts

Sirovich L, Kirby M. 1987. Low-dimensional procedure for the characterization of human faces. *Journal of the Optical Society of America A* 4:519-24

Tarr MJ, Pinker S. 1989. Mental rotation and orientation dependence in shape recognition. *Cognitive Psychology* 21:233-82

Tarres F, Rama A. 2005. A novel method for face recognition under partial occlusion or facial expression variations. *ELMAR, 2005. 47th International Symposium*, pp. 163-6

Terzopoulos D, Waters K. 1993. Analysis and synthesis of facial image sequences using physical and anatomical models. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 15:569-79

Thierry G, Martin CD, Downing P, Pegna AJ. 2007a. Controlling for interstimulus perceptual variance abolishes N170 face selectivity. *Nature Neuroscience* 10:505 - 11

Thierry G, Martin CD, Downing P, Pegna AJ. 2007b. Is the N170 sensitive to the human face or to several intertwined perceptual and conceptual factors? *Nature Neuroscience* 10:802 - 3

Thompson P. 1980. Margaret Thatcher - a new illusion. *Perception* 9:483-4

Tong F, Nakayama K, Moscovitch M, Weinrib O, Kanwisher N. 2000. Response properties of the human fusiform face area. *Neuropsychology* 17:257-79

Turk MA, Pentland AP. 1991. Eigenfaces for recognition. *Journal of Cognitive Neuroscience* 3:71-86

Valentin D, Abdi H, O'Toole AJ, Cottrell GW. 1994. Connectionist models of face processing: a survey. *Pattern Recognition* 27:1209-30

Valentine T. 1988. Upside-down faces: A review of the effect of inversion on face recognition. *British Journal of Psychology* 79:471-91

Valentine T, Bruce V. 1986. The effect of race, inversion, and encoding activity upon face recognition. *Acta Psychologica* 61:259-73

Vinette C, Gosselin F, Schyns PG. 2004. Spatio-temporal dynamics of face recognition in a flash: it's in the eyes. *Cognitive Science* 28:289-301

Wallis G, Bulthoff HH. 1999. Learning to recognise objects. *Trends in Cognitive Neurosciences* 3:22-31

Wallis G, Bulthoff HH. 2001. Effects of temporal association on recognition memory. *Proceedings of the National Academy of Science USA* 98:4800-4

Wang Z, Bovik AC, Sheikh HR, Simoncelli E. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing* 13:600-12

Waters K. 1987. A muscle model for animating three-dimensional facial expression. *Computer Graphics* 21:17-24

Williams L. 1990. Performance driven facial animation. *Computer Graphics* 24:235-42

Yacoob Y, Black MJ. 1999. Parameterized modeling and recognition of activities. *Computer Vision and Image Understanding* 73:232-47

Yin RK. 1969. Looking at upside-down faces. *Journal of Experimental Psychology* 81:141-5

Yin RK. 1970a. Face recognition by brain-injured patients: A dissociable ability? *Neuropsychologia* 8:395-402

Yin RK. 1970b. Face Recognition: A dissociable ability? *Neuropsychologia* 23:395-402

Young A, Hay D. 1986. Configural information in face perception, Experimental Psychology Society, London