Example-based Computer-Generated Facial Mimicry

by

Glyn Cowe

BSc Mathematics, University of Bath (1997)

Submitted to the Department of Psychology

University College London

In partial fulfilment of the requirements for the Degree of

Doctorate of Philosophy

January 2003
Abstract

Computer-generated faces, or *avatars*, are becoming increasingly sophisticated, but are visually unrealistic, particularly in motion, and their control remains problematic. Previous work has implemented complex three-dimensional polygonal models, often generated from laser-scans, with intricate hard-coded muscle models for actuation of speech and expression. Driving the avatar through mimicry, or *performance-driven animation*, involves tracking a real actor’s facial movements and associating them with analogues on the model. Motion is tracked usually through markers physically attached to the actor’s face or by locating natural feature boundaries. Here, complex three-dimensional models are avoided by taking an image-based approach. Novel techniques are presented for automatically creating and driving photo-realistic moveable face models, generated from example footage of a face in motion. Image changes for each frame, coupled with dense motion fields extracted using an optic flow algorithm, are analyzed to extract a set of basis actions by application of principal components analysis. These techniques yield a virtual avatar onto which the movements of an actor can automatically be projected for convincing performance-driven animation, with no need for markers.
Contents

Example-based Computer-Generated Facial Mimicry ... 1
Abstract ..............................................................................2
Contents .............................................................................3
Chapter 1- Introduction.....................................................8
  1.1 - Applications for Photo-realistic Computer-Generated Facial Mimicry .... 9
  1.2 - History of Facial Animation ......................................................... 10
  1.3 - Computer-Generated Facial Animation ....................................... 12
  1.4 - Example-based Facial Mimicry .................................................. 12
Chapter 2- Modelling the face ........................................14
  2.1 - Computer-Generated Models of Faces........................................ 14
    2.1.1 Surface representations ............................................................. 14
    2.1.2 Volume representations ............................................................. 17
    2.1.3 Dynamic Representations ......................................................... 18
    2.1.4 Summary .................................................................................. 19
  2.2 - Biological Representations of Faces ........................................... 20
    2.2.1 Specialist Mechanisms ............................................................... 20
    2.2.2 Separate Modules .................................................................. 21
    2.2.3 3D Representation ................................................................. 22
    2.2.4 Invariance ............................................................................... 23
    2.2.5 Features or configuration? ....................................................... 25
    2.2.6 Principal sources of variance in the face .................................. 28
    2.2.7 Encoding of Motion ................................................................. 29
  2.3 - Biologically Inspired Facial Representation ............................. 31
    2.3.1 Eigenfaces ............................................................................... 31
    2.3.2 Conclusion ............................................................................... 33
Chapter 3- Related Work.................................................35
  3.1 - Motion Capture ....................................................................... 35
    3.1.1 Key-framing ........................................................................... 36
    3.1.2 Dot tracking ............................................................................ 36
    3.1.3 Deformable contours ............................................................... 37
    3.1.4 Feature tracking ................................................................. 37
    3.1.5 Optic flow ............................................................................... 39
    3.1.6 3D tracking ............................................................................ 40
    3.1.7 Muscle Tracking ................................................................. 41
  3.2 - Two-dimensional Facial Animation ........................................ 41
Chapter 4- Example-based Generation of Facial Avatars

4.1 - An Example-based Approach to Modelling Faces
4.1.1 Facial motion as pixel-wise intensity variations
4.1.2 Centring the Examples
4.1.3 Principal Components Analysis
4.1.4 First principal component
4.1.5 Second principal component
4.1.6 The remaining principal components
4.1.7 Reducing computation
4.2 - Vectorising Faces by Warping
4.2.1 Warping a reference
4.2.2 The Multi-channel Gradient Model (McGM)
4.2.3 Results
4.3 - Vectorising Faces by Morphing
4.3.1 Image Morphing
4.3.2 Morph Vectorisation Procedure
4.3.3 Results
4.4 - Summary

Chapter 5- Facial Movement Analysis

5.1 - Existing Facial Action Coding Schemes
5.1.1 The Facial Action Coding Scheme (FACS)
5.1.2 FACS+
5.1.3 MPEG-4
5.2 - PCA for Coding Facial Action
5.2.1 Previous Applications of PCA in Facial Motion Analysis
5.2.2 Methods
5.2.3 Comparison of results
5.3 - PCA and the human visual system
5.3.1 PCA on a corpus of sequences
5.3.2 Results
5.4 - Conclusions
E.3.3- Mimicry by Projecting Morph Variations onto a Second Person's Basis Set
E.4- Relating Vectorisations
   E.4.1- Related Basis Set
   E.4.2- Mimicry by Projecting onto a Related Basis Set
   E.4.3- Handcrafted Basis Movements
   E.4.4- Mimicry by Projecting onto Handcrafted Basis Set
E.5- PCA on a Corpus of Sequences Vectorised as Morphs
Chapter 1- Introduction

For most of us, faces play an integral role in identification and communication each day of our lives. We can easily and robustly recognise our friends, family and colleagues from the visual information contained within their faces, yet these faces are dynamic in nature and their frequent configural perturbations yield important cues for social interaction, without disturbing our perception of the subject's identity (Bruce & Green, 1990). The face is an intriguing and complex stimulus and its synthesis is a challenging task not to be underestimated.

This thesis concerns the computer-generated synthesis of human faces and their movements, focussing on realism. Novel techniques are introduced for the automatic generation of photo-realistic, moveable facial models, or avatars, from standard video footage. The models are automatically parameterised by a small set of basis movements, which can be linearly combined to produce a broad range of gestures. Given a second video sequence of another individual (or the same, if required) their movements can then be projected onto the generated model, empowering the model to convincingly mimic the original footage. The work has resulted in a patent application (Cowe & Johnston, 2001).

This chapter introduces the subject, initially by motivating the development of the techniques to be described, with a discussion of the applications. A brief overview of the rich history of synthetically mimicking faces follows, leading up to the computer age with a summary of the current state of the art and its weaknesses. Finally, in addressing these weaknesses, the example-based focus underlying the present approach is developed.
1.1 - Applications for Photo-realistic Computer-Generated Facial Mimicry

An immediately obvious theatre of application for the realistic synthesis of human motion is the entertainment industry. Some faces in films, television and computer-games cannot physically be portrayed by human actors, such as those of animals or non-human characters, and necessitate artificial manipulation, but it could be asked: Why there is a need for artificially manipulating real human faces? Surely it is easier, less time consuming and consequently cheaper just to hire genuine actors?

A number of scenarios can be envisaged where a particular personality is required, but is not available for filming. The protagonist of a film could be injured or die before it is finished, for example. This would necessitate starting again from the beginning, or finding another way of finishing off the film without them. An appearance by a well known, but long-dead political figure or celebrity may be required, where a sufficiently convincing double is not available. Alternatively, again, it may not be appropriate to use the actual individual portrayed, for example in satirical productions. In these situations, it can be seen how computer-generated mimicry could be applied.

Industry techniques for computer generated facial animation are indeed currently time-consuming and expensive, but with advances in automation with acceptable output, it is not unrealistic to imagine scenarios where it could be cheaper to artificially manipulate the face of a celebrity in high demand than to pay for their time. A good application for this could be in film dubbing, where a new audio track, usually in another language, is played over the original footage. It is impractical to get the actors to re-enact all of their scenes so lips often do not match the dialogue, but computerised techniques could be used to correct this.

There are also applications for synthetic facial manipulation outside the entertainment industry. In telecommunications, for example, bandwidth is expensive and limited, so video is often heavily compressed before transmission, frequently also at a low frame rate. The recipient's best reconstruction is consequently of extremely poor quality. The techniques
developed here reduce faces into high-quality models that can be manipulated with a small number of parameters. Given that the recipient is provided with such a model in advance, only the parameters need be sent for every frame for a high quality reconstruction, rather than the full image, reducing the quantity of information for transmission drastically, whilst increasing the quality of the reconstructed sequence.

The facility to realistically transfer the motion from one face to another is also useful for investigation into human face perception. Separating facial motion from other cues has long posed a problem for its analysis and only recently has it been possible to do this with facial animation techniques (Hill & Johnston, 2001; Knappmeyer, Thornton, & Bülthoff, 2001).

1.2 - History of Facial Animation

Replicating faces is far from new. For examples of synthetic faces in static pose, we need only consider the great works of art in sculpture and portraiture, with some evidence dating back to the very beginnings of civilisation.

The synthesis of moving people, however, is the art of the puppeteer, arising across many cultures, throughout human history, as an age-old tool for storytelling and entertainment. Egypt and India have the oldest string operated puppets ever found. In 1934, ancient Egyptian string operated figurines from the twelfth dynasty (circa. 1991 BC – 1786 BC) were discovered amongst the treasures in the excavation of a tomb at Lisht by the Metropolitan Museum of Art, New York (Mattil, 1971). Such early puppets had static faces and just their bodies alone could be manipulated, serving as further testament to the complexity of facial motion.

The ancient Greeks certainly had puppets, with numerous references to them in their literature (Speaight, 1955). From writings by Horace from 30BC, we have:

"you are moved like a wooden puppet by wires that others pull"

From Philo at around the birth of Christ:
“all these as in marionette shows, are drawn with strings... each in the attitudes and with the movements appropriated to it”

And from Apuleius around 200AD:

“those who impart gestures to the wooden figures of men, when they draw a string to the limb that they wish to move, the neck turns, the head nods, the eyes roll, the hands are ready for every purpose, and the whole is seen, not ungracefully, to live”

Again, little reference is made to moving faces, apart from rolling of eyes, but an artist and puppeteer, R. Bruce Inverarity, following his visit to the Queen Charlotte Islands in 1932, described traditional string-controlled masks that had been used before the arrival of the Europeans by the aborigines in fire-lit performances for showing the bonds between the human and spirit worlds (McPharlin, 1949).

Medieval puppets from England are described with moveable faces (Speaight, 1955). The Rood of Grace at Boxley, Kent, a crucifix preserved from medieval England, was said to have been made by an English carpenter whilst imprisoned in France. It is reported that the eyes did

“move and stare in the head like unto a living thing, and also the nether lip likewise to move, as though it would speak”.

A similar French example is held in the Cluny Museum.

Puppetry continues in modern times with today’s technologies providing powerful new tools in the pursuit of realism and versatility. The application of robotics in puppetry, or *animatronics*, is often utilised in filmmaking for animating complex artificial creatures, but the advent of computer graphics has
allowed a progression into the virtual world, removing all previous physical barriers, granting almost limitless power to the animator.

In computer graphics, much effort has been recently applied in the pursuit of creating convincing synthetic face models capable of fooling the viewer. However, the human visual system is expert in the processing of faces and has proven exceptionally difficult to deceive.

1.3 - Computer-Generated Facial Animation

In order to produce compelling facial animation, actions must be faithful to the audience’s daily experiences of faces in motion. Mimicry is thus always the basis of facial animation.

In filmmaking, where realism is vital, faces are currently animated almost entirely by hand. The voice-track is normally recorded first, and the animator manually manipulates the model with his own face in a mirror as a reference, or occasionally with the footage of the original actor accompanying the vocal track. All facial animation in the pioneering all-computer-generated 2001 movie, Final Fantasy, was executed entirely by hand (Final Fantasy, 2001).

Techniques for automating this process by tracking the movements of an actor’s face have been developed, generally either by tracking markers physically attached to the actor’s face, or by tracking prominent facial features. However, animations generated using this information alone have not been sufficiently compelling for its adoption into the industry, with its application seen generally only in low budget television productions, although Famous Faces, a commercial dot-tracking facial animation system boasts its application in the recent Lord of the Rings movie (Lord of the Rings, 2001).

1.4 - Example-based Facial Mimicry

The creation of computer-generated faces for animation is a complex and time consuming process and even the highest quality animations, such as Final Fantasy are not sufficiently compelling to fool the highly critical human visual system, particularly in motion.
As is detailed later, psychophysical evidence indicates that it is unlikely that the very system we are trying to fool stores a fully three-dimensional representation for the human face. An alternative 2D approach is thus presented in the following chapters, avoiding the complexity of three-dimensional models and exploiting the information available from real facial image sequences to define how faces can move and hence to control them accordingly. Given a set of examples of a face in motion, the novel approach outlined in this thesis enables the automatic generation of an image-based model from this information alone. It is shown how the natural layout of human faces is sufficiently congruous to allow the driving of this model, simply by projecting a second actor's movements onto it.
Chapter 2- Modelling the face

The first step in computer animation is building the model. It is important to choose a representation that is sufficiently powerful to successfully produce the full range of movements required, without being too complex for efficient animation. A variety of representations can be used for computer modelling of the face and here the more common are discussed. Inspiration is then gathered from the human brain. Not only is it the most powerful face processing system at our disposal, but it is also important to understand the very system photo-realistic facial animation strives to fool. This chapter concludes, showing how this information can be used, motivating a representation inspired by knowledge of our own visual systems.

2.1 - Computer-Generated Models of Faces

Although artificial facial animation is rarely viewed in conditions other than through a projection onto a two dimensional surface, the face is inherently three-dimensional in structure, so previous approaches have generally 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.

2.1.1 Surface representations

Surface representations have most commonly been used in realistic head modelling. Perhaps the simplest approach is the polygonal mesh, a surface described by a set of polygonal planar surfaces (usually triangular) connected at each vertex. The use of flat polygons to represent smoothly varying surfaces inevitably leads to errors, but these can be made arbitrarily small by increasing the number of polygons, although this predictably increases the rendering time and the storage requirements. Most graphics cards, however, now have fast, efficient, inbuilt polygon rendering routines for this purpose.
Parke initiated work on computer-generated facial animation with a crude polygonal model of the head (Parke, 1972). The position of each vertex in three-dimensional space was derived from real faces using a photogrammetric technique. A mesh was painted onto one side of a face that was then photographed from two viewpoints, frontal and profile, as shown in Figure 2.1. Co-ordinates of vertices were measured in 2D, and the 3D co-ordinates were geometrically recovered. The mesh was then constructed and the polygon faces were coloured.

![Figure 2.1- Example image pair for Parke's photogrammetric method (reproduced with permission from (Parke, 1972))](image)

Williams demonstrated how a more accurate polygonal mesh could be acquired by laser scanning a plaster cast of a human model's head (Williams, 1990). The scanned data was in cylindrical co-ordinates, with missing data (lost due to occlusions) mapped to the origin, so discontinuities needed to be smoothed by interpolating over the surrounding data. Photographs were taken of the model's head and painstakingly aligned and registered with the scanned data to map onto the computer head. It is now possible to scan real heads with custom-made laser scanners, such as Cyberware's, and simultaneously capture the localised texture map as the scanner rotates around the head, eliminating the need for the time-consuming alignment stage (see Figure 2.2).
Other surface representations include implicit surfaces and parametric surfaces. Implicit surfaces are 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 \((a, b, c)\).

\[
f(x, y, z) = (x - a)^2 + (y - b)^2 + (z - c)^2 - r^2 = 0
\] (2.1)

Although these implicit surfaces can be used to model shapes that can otherwise only be crudely approximated in alternative representations, and produce smooth shapes without polygonal edges, they are difficult to interact with since equations become unwieldy for complex surfaces, such as the face, and require much more processing. They are hence not a good choice for modelling 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 \( z \). 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 expensive to process than the simpler polygon. It is also difficult to seal up a surface using these surface patches; it is not possible, for example, to represent a sphere in this form without encountering the "north-pole problem", at least one point always exists where there is an irresolvable discontinuity. This can be
countered by basing the head model on a torus (which does not suffer this problem), since

“A human or animal figure is essentially a torus where the hole of the torus extends from the mouth to the anus as the alimentary canal” (Forsey, 1990).

An extension of parametric surfaces is hierarchical splines (Forsey & Bartels, 1988). Extra surface detail is added where necessary, by overlaying a more detailed surface with control points defined relative to the underlying surface, so that when this primary mesh deforms, the additional surfaces deform consistently. The facial animation system, Langwidere, was developed using hierarchical splines (Wang, 1993).

2.1.2 Volume representations

Three-dimensional objects can be represented 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).

While CSG is perfectly acceptable for simple face models, difficulties arise when trying to model the detail required for realistic faces. Spatial occupancy enumeration is often used for three-dimensional representations in biomedical applications for storing data from computerised axial tomography (CAT) or magnetic resonance imaging (MRI) scans (Foley, van Dam, Feiner, Hughes, & Phillips, 1997). This is a representation analogous to pixels in 2D images. Data is stored in identical cells, generally cubic in shape, known as voxels, which are arranged in a fixed rectangular grid. As with pixels, volume must be quantized; voxels can be either completely occupied or completely unoccupied, with no notion of partial occupancy. This representation requires a vast amount of storage and a level of internal detail that is necessary for MRI and CAT scan data, but not for face modelling, since, for faces, only surfaces are visible. Voxels are also difficult to animate.
2.1.3 Dynamic Representations

An important consideration in choosing a representation for facial animation is the matter of control. It is necessary to be able to produce any natural movements with simple manipulations of the model's parameters. This is not an easy task, since three-dimensional computer graphics is ideally suited for modelling rigid movements and biological movement is often non-rigid, particularly in the expressive movements of the face.

The complexity of Parke's polygonal model meant there were too many parameters to animate the mesh manually, so he was prompted to revise his model by parameterisation (Parke, 1974). In this new parametric model, numerical weights controlled subsets of vertices, representing specific facial features, such as the mouth opening height, width and protrusion, allowing for fast, easy animation.

With aspirations to more accurately model the control of the face to deal with the difficulties encountered with differing face topologies, 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-coding 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.

Platt and Badler introduced the first muscle model for facial animation (Platt & Badler, 1981). They modelled muscles as springs obeying Hooke's law:

\[ dl = \frac{df}{k} \tag{2.2} \]

Where the displacement of a point, \( dl \), is given in terms of the force, \( df \), and the sum of the spring constants at that point, \( k \). They demonstrated their system on a polygonal mesh with a few simple muscle actuations in the forehead region.

Keith Waters further developed the muscle model (Waters, 1987). He used a simplified model from research on facial muscles, based on Ekman and Friesen's *Facial Action Coding System* (FACS) (Ekman & Friesen, 1978). FACS is a scheme for coding facial movement that describes movement in terms of
action units (AU’s). Fifty AU’s are specified, each representing a muscle or a small group of muscles. Waters modelled ten of these to control a polygonal model. By locating approximately the key nodes of muscle attachment and extremes of displacement from the examination of faces, Waters modelled movement due to a particular muscle by moving the points of attachment maximally and the neighbouring points with diminishing strength as distance from the node increased. He modelled these effective zones of influence with a Hanning window (a 2D window based on the cosine function between $-\pi / 2$ and $\pi / 2$, with its maximum at the centre) (Figure 2.3).

Many improvements have been made to these preliminary muscle models. 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, 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). Underlying this surface layer was a muscle actuator model that reduced the set of voluntary facial muscles (in excess of 200) to a smaller set of about 20 muscle actuators. Again, a radial cosine blend function (as in Figure 2.3) described the zone of influence for each muscle fibre.

2.1.4 Summary

In realistic computer-generated animation, faces are usually modelled in three-dimensions with high-resolution polygonal mesh surfaces, generated either with the talent of highly skilled artists, or with expensive equipment such as laser
scanners. The sheer number of vertices makes them effectively uncontrollable, so intricate underlying muscle models are commonly implemented to constrain and parameterise movements. These methods are, in general, complex, time-consuming and often require the skills of experienced and talented artists, and, even then, results are still not sufficiently compelling to fool the human visual system. It thus seems appropriate to consider the internal processes involved in the perception of faces.

2.2 - Biological Representations of Faces

Perception can be thought of as computer rendering’s complementary inverse. Rather than creating a scene from a set of known parameters, perception involves the extraction of unknown causal parameters from the retinal image.

The human visual system seems to be specifically tuned for the processing of faces. Babies only a few minutes old have been found to show a preference for face-like patterns, indicating some degree of hard wiring for facial processing before birth (Johnson & Morton, 1991; Mondloch et al., 1999).

Facial animations are almost exclusively created for the human visual system, so it is important to understand it. Also, given its success in everyday processing of faces, it also seems only appropriate to study it for inspiration. Some of the known properties of the mechanisms involved are now discussed in the following sections.

2.2.1 Specialist Mechanisms

It is commonly thought that the brain has special mechanisms devoted to the sole purpose of processing faces. 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, aeroplanes 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 (Yin, 1969). Similar results have also been found when comparing faces to houses and words (Farah, Wilson, Drain, & Tanaka, 1998).
This could occur because our visual systems are hard-wired to process faces differently, but it has been postulated that this effect occurs simply because of our disproportionate exposure to upright faces. Indeed, for example, inversion has been shown to have more of an impairing effect on the recognition of pedigree dogs for experts over non-experts (Diamond & Carey, 1986). Similar results have also been found in the context of handwriting (Bruyer & Crispeels, 1992) and a specifically designed set of novel objects called greebles, which, like faces, share a common spatial configuration, (Gauthier & Tarr, 2002; Gauthier, Williams, Tarr, & Tanaka, 1998).

Experiments on the recognition of faces of other races, however, have presented a surprising result, seeming to contradict the hypothesis that disproportionate exposure of an object in a particular orientation leads to this excessive impairment of recognition when inverted. Valentine and Bruce tested Caucasian subjects on a recognition task with Caucasian and Black faces (Valentine & Bruce, 1986). Caucasian subjects were found to be not as good at recognising Black faces, arguably because they have less exposure to them, yet inversion was shown to have a greater detrimental effect on recognition of Black faces over Caucasian faces. It would appear, from this result, that other-race faces are encoded in a less efficient manner, and inversion further hinders their decoding. We are generally much worse in the processing of unfamiliar faces (Hancock, Bruce, & Burton, 2000), so it seems likely that encoding improves with familiarity in a constant learning process.

2.2.2 Separate Modules

It has also been proposed that there are parts of the brain dedicated to the processing of faces alone. Prosopagnosic patients, as first reported by Bodamer, suffer a deficit where all faces appear unfamiliar to them and, for some, recognition even of their own face is an impossible task (Bodamer, 1947). However, some have been known to retain the ability to differentiate between other very similar objects with little difficulty (Farah, 1995). Conversely, some patients have also been reported with deficits in general object recognition, but no impairment with faces (Moscovitch, Winocur, & Behrman, 1997). This double dissociation suggests the existence of an area of the brain dedicated to face processing, although many dispute this since most
prosopagnosics suffer also from deficits in general object recognition. Kanwisher counter-argues that

“the chances that a stroke or head trauma to visual cortex will obliterate all of the hypothesized face-processing region of cortex without affecting nearby cortical areas is similar to the chance that an asteroid hitting New England would obliterate all of the state of Rhode Island without affecting Massachusetts or Connecticut” (Kanwisher, 2000).

Studies on these face areas in primates, however, demonstrate that it is unlikely that they are uniquely dedicated to face processing; of cells in these locations responsive to stationary stimuli, in fact, at most 20% have been found to be responsive to faces (Rolls, 1992; Wallis & Rolls, 1997). This suggests instead that these regions are dedicated to a more general expert processing of highly familiar stimuli.

Domain specificity for faces is under debate, but evidence also exists for independent processing routes within face processing. Patients have been documented who lose the ability to recognise faces whilst retaining the facility to interpret facial expressions (Bruyer et al., 1983; Shuttleworth, Syring, & Allen, 1982) and, conversely, patients have been reported without the ability to interpret emotion in faces, but no deficit in recognition (Kurucz & Feldmar, 1979; Kurucz, Feldmar, & Werner, 1979). This double-dissociation suggests that recognition of identity and expression are processed independently of each other. Psychophysics also provides evidence for the independence of sex judgements and expression judgements. Although facial expressions were found to affect subjects' ratings of masculinity and femininity of faces, it was found, in a speeded classification task, that they could attend selectively to either dimension without interference from the other (Le Gal & Bruce, 2002).

2.2.3 3D Representation

The human visual system is remarkably effective at identifying familiar objects regardless of their positions or orientations. Marr and Nishihara suggested how this could be achieved with an object-centred co-ordinate system, independent of the viewer (Marr & Nishihara, 1978). In order to facilitate view invariance, several three-dimensional representations have been proposed for the visual
system. For example, Marr and Nishihara argued for volumetric representations with generalised cones as primitives (Marr & Nishihara, 1978). These are solids swept out by a curve in 3D space (see Figure 2.4). Pentland suggested a more flexible scheme consisting of deformable geometric solid primitives, such as cones, cubes, spheres, etc, called superquadrics (Pentland, 1986). Biedermann suggested an alternative scheme again, but with reference to the image plane. He proposed a set of 36 primitive 2D object shapes, or geons, with properties that are invariant over different views (Biederman, 1987). The advantage of these schemes is that the complex volume of the face can be reduced to a set of parameters describing each primitive, i.e.- the volume of a cone need only be described by the radius of its base, its height and its relative position. All these suggestions are analogous to schemes discussed in the previous section. Building shapes out of primitives, however, does not allow for high levels of surface detail. Bruce and Young also argue that faces have generally very similar underlying volumes, so a representation of this form allows little detail for distinguishing between them (Bruce & Young, 1986).

![Figure 2.4- Generalised cone](image)

### 2.2.4 Invariance

Invariant representation of objects is an attractive proposal since the visual system can robustly deal with changes in view and lighting, but surprisingly, evidence seems to show, certainly for faces, that this is not entirely the case for biological vision.

The principal evidence for view dependency in face processing is the previously mentioned inversion effect. Faces are much harder to recognise when inverted, rather than upright (Valentine, 1988; Yin, 1969, 1970).

Faces are rarely seen inverted, but it should be noted that even regularly experienced changes in viewpoint have a substantial effect. Hill et al.
demonstrated how subjects performed poorly in a recognition task where viewing conditions were altered (Hill, Schyns, & Akamatsu, 1997). Although all views were equally well recognised when they all had been learned, participants 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. Generalisation between views has also been shown to be significantly worse when faces are inverted (Moses, Ullman, & Edelman, 1996).

Lighting invariance has also been investigated. 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). An advantage for illumination from above was found, with better performance in a matching task under this condition. This finding is consistent with evidence that the visual system employs an assumption of illumination from above with a single light source, when interpreting simple shading patterns (Ramachandran, 1988a, 1988b). By illuminating faces from below, then inverting, Johnston, et al. demonstrated that the face inversion effect could be significantly reduced, or even eliminated (Johnston, Hill, & Carman, 1990). They took this as evidence for a surface based code for faces, but subsequent evidence illustrates that further information must be employed by the visual system. The **hollow face illusion**, in which a concave model of a face is perceived to be convex, demonstrates that this is not the only constraint in the recovery of shape from shading (Hill & Bruce, 1993, 1994). Familiarity with the three-dimensional structure of the face still seems to play some part in the process.

Psychophysical evidence thus seems to indicate that the visual system's representation of faces is not totally viewpoint and lighting invariant. It is view-dependent and lighting-dependent. Neurophysiological investigations of the primate brain have uncovered cells in the superior temporal sulcus tuned to specific facial orientations, particularly full-face and profile (Perrett et al., 1991; Perrett et al., 1985). Hietanen et al. found such view dependent cells to be lighting and position invariant (Hietanen, Perrett, Oram, Benson, & Dittrich, 1992). Hasselmo et al. found cells that respond to all views of a face (Hasselmo, Rolls, Baylis, & Nalwa, 1989). This seems to indicate 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 & Bülthoff, 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 (Bülthoff & Edelman, 1992; 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 (Wallis & Bülthoff, 1999). The importance of the temporal aspect in learning invariant models has been powerfully demonstrated in an experiment in which viewing position and identity of a face were simultaneously altered (Wallis & Bülthoff, 2001). In these circumstances, subjects showed a tendency to treat the views as though they were of the same person.

2.2.5 Features or configuration?

*Photofit* is a facial reconstruction tool used by the police, consisting of many variations of individual facial features that can be combined by a witness to form a composite face. Penry, the inventor of Photofit, wrote,

"because each facial part is the sum of its individual details and the whole face is the sum of its sections, the total assessment of it requires careful visual addition" (Penry, 1971).

The principal of photofit is that the face can be broken down into its constituent parts and any face can be made up from the components available. In normal circumstances, this does not appear to be the case and photofit, and similar facial reconstruction packages, are notoriously ineffective. Davies et al. reported that witnesses found it very difficult to produce an accurate copy of a face with this limited reconstruction tool, noting that they could produce considerably better likenesses just by sketching with the face in view (Davies, Ellis, & Shepherd, 1978). In testing the witnesses a week later, it was found that there was no difference in accuracy of reconstructions, demonstrating not so much the robustness of the kit, but the lack of sensitivity; witnesses could produce no better reconstruction with the face in view than by memory a week later.
Laughery found similarly disappointing results testing the American equivalent, *Identikit* (Laughery, Duval, & Fowler, 1977; Laughery & Fowler, 1980).

These reconstruction kits fail, not only because of their limited set of features, but because people do not appear to store faces solely in this way. One need only consider linguistic terms defining facial features, vague regions of the face with no clear boundaries, shrouded in fuzziness. Where exactly do the cheeks begin and end, for example? Whether faces are stored as a collection of features, or not, was investigated by Sergent (Sergent, 1984). She argued that, if faces are stored by individual features alone, it should not be possible to identify two faces as being different any faster than the time taken to identify a change in the most salient feature. She found that correct judgements for faces with several differing features were made faster than for changes in any single feature. The contrary was found for inverted faces, suggesting that inverted faces are processed on a feature-based level, whereas a different approach is undertaken for upright faces. Featural information is still clearly retained. When subjects were shown photo-fit pictures of unknown faces in a learning phase, Solso and McCarthy demonstrated that novel faces containing features present in the learning task rated high in familiarity, despite their novelty, not only in comparison to wholly novel faces, but in comparison to the learned set as well (Solso & McCarthy, 1981).

So when faces are not processed by their individual features, how are they processed? In a review, Maurer, et al. distinguish three types of configural processing:

> “detecting the 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)” (Maurer, Le Grand, & Mondloch, 2002).

Utilising a computerised image-processing tool that allowed him to place any feature anywhere on a face, Haig demonstrated that subjects are highly sensitive to quite minute changes in position of features (Haig, 1984; Haig, 1986a, 1986b). Such changes have been shown to be much harder to detect once the face is inverted, whereas alterations to the appearance of individual features are relatively unaffected by inversion (Freire, Lee, & Symons, 2000;
Kemp, McManus, & Pigott, 1990; Le Grand, Mondloch, Maurer, & Brent, 2001; Leder, Candrian, Huber, & Bruce, 2001). Extreme alterations in positioning of features also appear less grotesque when the face is inverted (Bartlett & Searcy, 1993). This suggests that it is this configural information that is key to the inversion effect. It seems that upright faces are encoded effectively in a configural manner, but this information is degraded when inverted and subjects have to resort to less efficient feature-based processing.

Leder et al. argue, however, that this configural information used in face processing consists at least partly of locally processed relations (Leder et al., 2001). They found that in inter-ocular distance judgements, inversion effects occurred strongly whether the eyes, nose and mouth were all presented, just the eyes and nose, or even just the eyes alone. Additionally, when testing these judgements with upright faces against inverted faces that had been altered by inverting the eye region, as in the Thatcher illusion (Thompson, 1980), they found no effect of inversion.

Powerful configurational effects have been found with faces. Interchanging upper and lower halves of faces between individuals produces composite faces with new identities independent of their components. Young and Hay found pronounced latencies in identification of top and bottom halves when aligned to form a new face (Young & Hay, 1986). Subjects found it much easier to identify each half when they were misaligned, suggesting that that the configurational effect of faces makes it difficult to consider each half individually. They did find, however, that subjects found recognition of top halves easier when the aligned composites were inverted.

Homa et al. discovered that subjects could recognise individual features easier in a test phase, if they were initially presented in context within a face in sensible positions, rather than in a scrambled non-face (Homa, Haver, & Schwartz, 1976). Another configural advantage has also been shown in that subjects are more accurate in recognition of features presented in context in a face, rather than in isolation (Tanaka & Farah, 1993; Tanaka & Sengco, 1997). It was found there was no benefit if the experiment was repeated with scrambled faces or houses.
Studies of babies present strong evidence for early processing of subtle configurational information. Infants of 5-6 months have been shown to be able to distinguish male from female and young from old, yet have not been able to distinguish between individuals until 7 months old (Fagan & Singer, 1979). Similar pairs were used for testing, such as twin brothers and sisters for testing discrimination of gender and bald men compared to baby faces for age, yet the infants were able to distinguish between them. Comparatively different faces were used within age and gender groups, yet the infants were not able to distinguish between these.

All this evidence strongly suggests that the visual system does not store a face just in terms of its individual features, but rather as a more general configuration, where spatial inter-relations play a crucial part.

2.2.6 Principal sources of variance in the face

Johnston suggested a parameterised representation of a face, described in terms of its major sources of variation (Johnston, 1992). This model could be analogous to common simple geometric solids such as spheres, cuboids, etc, which can be described with a small number of parameters, such as radius, and width and height. Parke’s 3D parameterised head model reduces the 3D surface geometry of the face to a comparatively small set of parameters, but Johnston’s parameterisation would seek to more efficiently encode the information with factors directly related to the principal variations in faces.

By flashing up individual features that made up a composite face in a random order and testing subjects on their ability to detect which was missing, Fraser and Parker found the outline of the face to be the most salient feature, followed by eyes, mouth, then nose (Fraser & Parker, 1986). Haig confirmed these findings by testing subjects on a face recognition task varying how much of the face was revealed through manipulation of apertures (Haig, 1986a). He verified the importance of the hair and eye regions, but also noted that salience of features for recognition depended on individual faces, leading him to question the whole issue of feature salience. Shepherd et al. found the principal sources of variation in a set of faces to be hairstyle, face shape and age (Shepherd, Davies, & Ellis, 1981).
Varying a prototype in terms of facial features affects only local structure and does not capture the essence of the parameterisation suggested. 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. It seems more appropriate to consider a holistic representation in which the face varies in terms of pseudo-features that affect the configuration of the face as a whole.

By considering faces to be parameterised by a set of features, regardless of their local or global nature, a face space can be envisaged in which dimensions are composed of the parameters and the average of all faces lies at the centre. This approach for representing faces has proved useful in accounting for several findings in the literature. The fact that caricaturing enhances recognition for both line drawings of faces (Rhodes, Brennan, & Carey, 1987) and of photographs of faces (Benson & Perrett, 1991), can be explained by considering distinctiveness as distance from the centre of face space and a caricature to be an extrapolation along the vector from the centre to the original, thus accentuating the distinctive parameters. Moving along that same vector towards the centre of the space instead, a process termed anti-caricaturing, would be expected then to reduce distinctiveness, thus impairing recognition and this is consistent with psychophysical findings (Stevenage, 1995). Leopold et al. recently demonstrated powerful after-effects in the context of the face-space paradigm (Leopold, O'Toole, Vetter, & Blanz, 2001). 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.

2.2.7 Encoding of Motion

Since experiments have rarely shown any advantage of dynamic sequences of faces over stills for recognition, it has been suggested that motion information is redundant. Johansson demonstrated with point light sources, however, that motion alone is sometimes sufficient for recognition of certain objects (Johansson, 1973). Even after adjusting the contrast on video footage of a walker dressed in black so that only point light sources on his joints are visible, the walker is still discernable from his motion. Bassili conducted similar
experiments with faces by blacking out the face and teeth with makeup and scattering white circular labels over the surface (Bassili, 1978, 1979). He found that naïve subjects were able to recognise the sequences 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. He found that subjects were also still able to identify emotions being expressed in these conditions, to some extent, and that errors in judgement correlated with errors encountered in full-face display conditions. Bruce and Valentine used this technique to investigate whether individuals could be recognised on the basis of their facial motion alone (Bruce & Valentine, 1988). They found that subjects achieved above chance results in recognition of emotions and of individuals from a small set, but performance was very poor.

Point light displays only offer motion information at the location of each point light source, so it is unsurprising that subjects find recognition tasks difficult in these conditions. Since recognition of faces, whether still or moving, is generally performed at ceiling, it is not easy to show a benefit of moving faces over still faces without somehow degrading the stimuli. Knight and Johnston did just this by degrading image sequences with photographic negation, arguing that this maintains the full motion field rather than offering a discretised sample of point light sources (Knight & Johnston, 1997). Their subjects found famous faces in moving sequences of negated images significantly easier to recognise than stills, reinforcing the hypothesis that motion cues do provide useful information in face processing tasks. Lander et al. furthered this work by countering the argument that a dynamic sequence simply provides more views of the face by comparing performance on similarly degraded dynamic sequences to performance on the same frames simultaneously presented (Lander, Christie, & Bruce, 1999). Subjects again found it significantly easier to recognise moving sequences.

Hill and Johnston isolated motion cues to test sex and identity judgements, by animating an androgynous three-dimensional face with the facial movements of various actors (Hill & Johnston, 2001). Subjects were found to be able to successfully discriminate sex and identity. Knappmeyer et al. also demonstrated how identity judgements could be biased by motion cues, by training subjects on two synthetic faces, animated with distinct characteristic sequences of
motion (Knappmeyer et al., 2001). Judgements of identity, whilst viewing morphs between the two faces, were shown to be influenced by the motion the face was undergoing.

2.3 - Biologically Inspired Facial Representation

In both psychology and computer graphics, practical representations for facial geometry have been carefully considered. Marr and Nishihara’s generalised cones, Pentland’s superquadrics and Biederman’s geons were proposed to account for the human visual system’s facility for identifying familiar objects, regardless of position or orientation (Biederman, 1987; Marr & Nishihara, 1978; Pentland, 1986). This also needed to be considered in computer graphics, where an efficient representation is paramount for fast rendering and storage. A single representation of an object is generally preferable over the alternative, storing a 2D image of every configuration of a scene that is likely to be encountered in the particular application.

It is thus not surprising that these volumetric primitives find an analogue in computer graphics’ CSG, and generalised cones form a subset of implicit surfaces. Evidence of view dependency in the human visual system suggests an alternative representation, however. Representations of faces other than three-dimensional have rarely been considered in computer graphics.

In perception, many have considered whether the visual system represents faces, locally as a set of individual features, or more globally, as a configuration. Psychophysical evidence, however, strongly suggests that faces are not processed merely as a set of building-block features. A more global configural approach is suggested.

2.3.1 Eigenfaces

A parameterised representation of the face was proposed in 2.2.6, where parameters are chosen to account for the fundamental sources of variance. Principal components analysis (PCA), a statistical technique that has been applied in computer graphics for numerous applications, provides a means of implementation. PCA involves a linear transformation of the axes of a high dimensional dataset in order to point them in the directions of the most variance
in the samples. This allows a more efficient coding system, since less important dimensions can be ignored, thus reducing the dimensionality of the space.

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). Figure 2.5 shows the first 12 principal components resulting from PCA on 80 photographs of faces. The images were taken from the Nottingham face database (Hancock: http://pics.psych.stir.ac.uk/cgi-bin/PICS/New/pics.cgi) and were first aligned with a scale, rotation and translation so that the eye positions were aligned, and then were cropped. 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. For large face databases, reasonable reconstructions can be made of faces outside of the training set (see Figure 2.6).

Figure 2.5- the first 12 eigenfaces of a database of 80 images
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. 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.

PCA has almost entirely been applied for encoding identity in faces, rather than facial movements. An exception is Calder, et al.’s work in analysis of facial expressions (Calder, Burton, Miller, Young, & Akamatsu, 2001). 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. This is consistent with the observations of the independence of expression and identity discussed in 2.2.2.

2.3.2 Conclusion

To conclude, computer-generated models of faces have previously generally been three-dimensional with complex hard-coded underlying muscle models employed for their control. Psychophysical and neurophysiological evidence, however, indicates that we store faces in a more two-dimensional manner. Two-dimensional, or image-based, approaches have rarely been considered in this context, but previous work with principal components analysis has demonstrated its effectiveness in efficiently encoding facial identity, with
dimensions tuned to the natural variations between faces. This motivates a novel approach to the generation of computer models of faces, where the model is image-based and dimensions of movement are extracted from experience of the face in motion using techniques such as PCA within, rather than between, facial identities. This evades the problem of hard-wiring complex muscle models, and it can be shown how this approach enables the automatic generation of an avatar from example footage of a face in motion by learning.
Chapter 3 - Related Work

The focus of this thesis is on the development of techniques for transferring motion from one face onto another in a convincing, photo-realistic manner. In the previous chapter it was shown how the problem of modelling faces has been approached generally with three-dimensional polygonal models, with movements constrained by underlying muscular models. An alternative image-based approach was suggested with movements constrained by example, rather than explicit knowledge of muscle structure.

Performance-driven animation requires tracking an actor's movements and relating those movements to the model. This chapter of related work begins by examining previous approaches to capturing facial motion. Previous two-dimensional image-based approaches in facial animation are then discussed, followed by past applications of principal components analysis in the context of faces.

3.1 - Motion Capture

Given a computer model of a face, a procedure is required for animating it. A mechanical analogue for computer-generated facial animation is animatronics. In situations where the puppeteer would otherwise be impossible to conceal, puppets are manipulated remotely using motors within their structure, usually via radio control. A puppeteer works with an interface, often just a wire frame skeleton of the puppet's structure known as a *Waldo*, and the target moves appropriately.

When animating a computer-generated face, abstract control devices with sufficiently many degrees of freedom have previously been applied, such as a musical keyboard in *Sony's System G*, however, if realistic facial motion is required, where better to get it from than a genuine face?

In order to drive a computer generated head model in accordance with an actor's facial movements, it is necessary to somehow track that movement in a
form suitable for projecting onto the model. A fundamental difficulty is the correspondence problem, registering the actor’s face with the geometric model, such that every moving point has a unique analogue in the model’s face. This section discusses several methods that have previously been applied in the tracking of facial motion and how they approach the correspondence problem.

3.1.1 Key-framing

As discussed in the previous chapter, Parke introduced the first computer generated facial model (Parke, 1972, 1974). A polygonal mesh was painted onto a face, then the face was photographed from two angles and the 3d location of each vertex calculated by measurement and geometry. This technique was applied for extracting meshes in a variety of face poses to provide key-frames. By taking two key-frames as extremes, the model could then be animated smoothly between these by calculating the linear path for each vertex and filling in intermediate frames by simply interpolating along these paths. This interpolation process is known as keyframing and is still often used to avoid having to calculate the 3D locations of the vertices for every frame of a facial sequence.

3.1.2 Dot tracking

First attempts at controlling computer-generated faces with real face movements were implemented by tracking markers on an actor’s face. Williams recorded himself with 22 retro reflective circles of tape positioned on his face, then manually tracked their positions in order to drive a model generated by laser-scanning the plaster cast of a human model’s head (Williams, 1990). Each point in the tracked sequence was then related to a single point chosen by the animator in the geometric model, thus providing a manual solution to the correspondence problem. Although this is a tedious process, it is perfectly tractable for short sequences with such a small set of discrete points. This technique only provides, however, displacements for this small set of points, clearly insufficient to animate a model head. Displacements in a predefined local area around each point (its zone of influence) needed to be estimated by weighting each point inversely proportional to its distance from the control point, a cosine window peaking at one in the centre and smoothly falling to zero at the edges (such as the Hanning window seen in Chapter 2).
Markers have been used often since for tracking facial motion. The introduction of automated dot tracking techniques and multiple cameras for 3-d motion estimation has led to high quality results (Guenter, Grimm, Wood, Malvar, & Pighin, 1998). Commercial packages are even available with automated dot tracking (eg. Famous Faces (www.famous3d.com)).

3.1.3 Deformable contours

Terzopoulos and Waters approached facial motion capture by using deformable contour models (also known as snakes) to track non-rigid motion of the face (Terzopoulos & Waters, 1993). These snakes are energy-minimising splines in the x-y plane that are attracted to and settle in valleys in 2D potential functions; the first derivative of the Gaussian-smoothed image in this case. They thus gravitate towards contours of high contrast, such as the eyebrows or lips, if sufficiently enhanced with make-up. Manual initialisation was necessary, but by incorporating visco-elasticity and rigidity properties, snakes proved to be robust contour trackers so long as displacements were not too extreme, since points that escape from the attractive zone of the valley are generally pulled back in by those that are still within.

3.1.4 Feature tracking

Dot tracking and deformable contour tracking can capture useful motion information, but these require the face to be highlighted in some way. Some of the more prominent facial features can be tracked without needing to do this.

Deformable templates have been applied for specific feature location (Hallinan, Gordon, Yuille, Giblin, & Mumford, 1999). For example, a template for eyes was described with a set of nine parameters, $\vec{g} = (x^e, x^c, r, a, b, c, \theta)$, where $x^e = (x^e, y^e)$ is the centre of the eye, $x^c = (x^c, y^c)$ is the centre of the iris and a, b, c and $\theta$ are defined as in Figure 3.1.
The upper and lower curves bounding the eye are approximated as parabolas and the iris is modelled as a circle. It is known that the whites of the eyes must be at high intensity peaks and the iris will be at a low intensity valley, so an iterative steepest descent process was used to steer the template parameters towards the target eye.

This procedure was principally designed for locating facial features in images and representing them in a form that can be used for geometrical comparison. It is easy to see, however, how it could be applied for feature tracking.

In working on performance-driven animation of cartoons, Buck et al. needed only the positions of the principal features for each frame and used non-intrusive methods to find them (Buck et al., 2000). A colour-based tracker was applied in the method described by the authors to extract the positions of features in a face. Ten parameters in total were extracted:

- The $x$ and $y$ values of the midpoint of the line segment $l$ connecting the two pupils (2)
- The angle of $l$ with respect to the horizontal axis (1)
- The distance between the upper and lower eyelids of each eye (2)
- The height of each eyebrow relative to the pupil (2)
- The distance between the left and right corners of the mouth (1)
- The height of the upper and lower lips, relative to the mouth centre (2)
The tracker found these features by comparing colour data in each frame with template data for each feature stored during a user calibration session. They found that sequences could not be tracked, however, if there was insufficient colour contrast between the user’s skin and lips or if there was too much rigid motion.

These techniques require the handcrafting of templates, but Cootes and Taylor demonstrated a method for generating a more general class of templates, which they refer to as active shape models or smart snakes (Cootes, Taylor, Cooper, & Graham, 1995). The user delineates feature points for objects in a set of example images and, given a novel case, their technique iteratively adapts the template for a best fit.

3.1.5 Optic flow

There are a variety of definitions of optic flow. In psychology, for example, it is often referred to as the retinal velocity field induced by a moving observer (Marr, 1982). Simoncelli, however, describes optic flow as the measurement of

"the apparent motion of local regions of the image brightness pattern from one frame to the next. In doing this, one is assuming that these intensity patterns are preserved from frame to frame." (Simoncelli, 1993)

This is the definition that will be used throughout this thesis. Optic flow algorithms can provide an estimate of speed and direction for locations in a frame of an image sequence. It should be noted, however, that optic flow and image motion, the true movement of objects within the scene, are not always the same (Verri & Poggio, 1989).

The methods previously described only provide movement information for discrete points or areas in the input sequence, requiring estimates to be made for all remaining points in order to drive the geometric model in an effective manner. Essa and Pentland improved on tracking by applying an optic flow algorithm (Simoncelli, 1993) to a sequence registered with a canonical mask in order to estimate displacement vectors for every pixel (Essa & Pentland, 1997). Their interest was in investigating dynamic expression classification as an extension to FACS.
Optic flow was calculated using Simoncelli's (Simoncelli, 1993) coarse-to-fine, Kalman filtering-based algorithm, then canonical points were extracted using view-based and modular eigenspace methods (Pentland, Moghaddam, & Starner, 1994). A geometric mesh was then adjusted in order to align it with these canonical points, and a mapping function applied to the motion data to compute an estimate of the velocities at each vertex of the mesh. A control loop was used to refine the input vector until it was consistent with a permissible muscle-based deformation of the geometric model.

Black and Yacoob used parameterised optic flow to track regions in face sequences (Black & Yacoob, 1995). They defined areas around the mouth, eyes, brows and the head, and tracked them with an optic flow algorithm to extract measures for each region for translation, divergence (isotropic expansion), curl (rotation) and deformation (squashing and stretching). They then showed how these could be applied to recognise expression. A method by Fleet, Black, Yacoob and Jepson for parameterising optic flow from examples, using PCA, is also discussed later in 3.3.6 (Fleet, Black, Yacoob, & Jepson, 2000).

### 3.1.6 3D tracking

There are fundamental problems with tracking motion in 2D and using this information to drive a 3D model. If 3D data could be captured, it would be much more appropriate to utilise this in these circumstances. Williams suggested using a mirror positioned next to the performer’s face in order to recover the 3D Cartesian co-ordinates of each point in every frame by calculating depth from stereo (Williams, 1990). Using a mirror or multiple cameras is equally as appropriate for all these methods, but another correspondence problem consequently arises that must be dealt with.

Guenter et al. implemented an elaborate system involving the tracking from six video cameras of 182 fluorescent dots, of six different colours, glued to an actor’s face in order to drive a laser scanned model (Guenter et al., 1998). The six views offered a more robust estimate of the three-dimensional position of each dot than from two views. Although tracking was automatically performed using a colour tracker, initialisation was required for every dot, as was its correspondence with the three-dimensional mesh’s analogue point. Some
packages on the market can track motion of 3D optical markers, such as Famous3D.

3.1.7 Muscle Tracking

It is also possible to drive a muscle model with muscle actuations from a real actor. This bypasses the problem of having to register the actor’s features with the model's. Muscle actuations have previously been captured for this purpose with intramuscular electromyography, or EMG, recordings (Lucero & Munhall, 1999). Once signals have been recorded from sensors on the actor’s face, the corresponding muscles can be activated on the model in order to mimic their movements. This is an effective solution, but requires expensive specialist equipment.

3.2 - Two-dimensional Facial Animation

Most work in realistic facial animation has previously involved three-dimensional models, but it is very difficult to produce a 3D model of sufficient quality to fool a human observer in believing it to be a real face. Although two-dimensional models have often been previously used for cartoon animation (Buck et al., 2000), they have rarely been applied in animations striving for realism. Presented here, are examples of realistic image-based animation techniques.

3.2.1 Example-based modelling

Beymer, Shashua and Poggio demonstrated how novel views of objects varying rigidly and non-rigidly could be generated from an image-based model by interpolating between example images, registered by application of an optic flow algorithm (Beymer, Shashua, & Poggio, 1993). By associating the example images with a handcrafted parameter space (for example, degrees of head pan and tilt) the mapping function from parameter space to image space could be smoothly approximated by using radial basis functions. Warping the examples towards the target view and blending enabled the generation of intermediate views. The causal parameters for a novel image could be estimated by solving the inverse mapping function.
Ezzat and Poggio employed this technique along with radial basis functions to model simple rigid and non-rigid facial motion by associating flow fields between example images and a reference image with user-defined high-level parameters (Ezzat & Poggio, 1996). Their approach for parameter estimation differed by basing an error-metric on flow fields rather than images, since flow fields are unaffected by lighting artefacts. Once the initial frame of a sequence was in correspondence with the reference, the high level parameters of the sequence could be extracted by iteratively perturbing the parameters and comparing the flow fields until a threshold was reached.

Fidaleo and Neumann recently introduced a methodology for generating an example-based virtual puppet from images (Fidaleo & Neumann, 2002). The face was split up into a small set of local regions, referred to as co-articulated regions (CRs), representative of small groups of muscles. A set of basic facial movements were chosen, which activated movements independently in each CR, and sequences were recorded in which an actor performed each of these movements individually. Markers were attached to an empty glasses frame in order to warp the sequences onto a standard position. Muscle actuations were separated using independent components analysis. Independent components analysis (ICA), is similar to PCA, but seeks to maximise independence between components, rather than variance (Bell & Sejnowski, 1995). The face could then be parameterised and new footage of the same actor could be analysed in real-time to extract the high-level parameters. Those same parameters could then be used to drive that model, or a cartoon character, handcrafted under the same parameterisation. This effectively enables photo-realistic performance-driven facial animation, but is limited to the one actor.

3.2.2 Prototyping

Tiddeman and Perrett demonstrated a technique, based on prototyping, which allows them to transform existing facial image sequences in dimensions such as age, race and sex (Tiddeman & Perrett, 2001). Prototypes are generated by averaging shape and texture information from a set of similar images (same race, for example). Each frame from the sequence can then be transformed towards this prototype. In each image, 179 points must be located and, although this can be automated using active shape models (Cootes et al.,
1995), a set of examples must first be delineated; so manual intervention cannot be avoided. Although this is not technically performance-driven facial animation, a new face is generated driven by the original motion and results are of realistic quality.

3.2.3 Videoconferencing

Compactly modelling the face at high quality with the minimal number of parameters is paramount for videoconferencing, where bandwidth is limited. Koufakis and Buxton presented an image-based method for encoding an individual’s face for this purpose (Koufakis & Buxton, 1999). Assuming that a set of landmark points could automatically be extracted from images of faces, they showed how static face pose could be modelled by linearly combining warps of three manually selected basis faces. By then extracting three sets of image patches, one for each of the eyes and one for the mouth, over a corpus of aligned example frames of that person, they then applied PCA to those patches to model their dynamics as a small number of the most important principal component basis images. Although they only tested their scheme on single static frames, so were unable to test for smoothness of transitions between frames and other temporal artefacts, they demonstrated how high quality reconstruction could be achieved in this manner at very low bit rates.

3.2.4 Lip-synching

There are some examples of photo-realistic image-based techniques used for synchronising the movement of the mouth with an audio track, or lip-synching. One such system is MikeTalk, developed by Ezzat and Poggio (Ezzat & Poggio, 1997, 1998, 2000). A set of basis images, or visemes, is manually extracted from example footage, one for each of 40-50 phonemes, and pixel-wise correspondences are pre-calculated between them all with an optic flow algorithm. Video footage accompanying either a phoneme-annotated audio track, or from the Festival speech synthesis system (Black & Taylor, 1997), which automatically performs the annotation from text, is automatically generated by placing the images associated with the respective phonemes in their place on the time-line and interpolating between by morphing (see 4.3.1) along the flow fields.
By considering a one-to-one mapping between phonemes and visemes, however, coarticulation effects are ignored; for example, the “t” in “beet” looks very different from in “boot”. *Video Rewrite*, a system developed by Bregler et al., also automatically lip-synchs existing footage to a new audio track, but deals with coarticulation effects by considering triplets of phonemes, or *triphones*, instead (Bregler, Covell, & Slaney, 1997). The movements associated with each triphone available in example footage are extracted and lips are tracked using *eigenpoints* (Covell & Bregler, 1996). Mouth movements are predicted for each triphone from the new audio track, with the best approximation used if the required triphone is not available, and these are incorporated into the existing sequence and merged by warping and blending. Since triplets of phonemes need to be stored, rather than single phonemes, an enormous database with considerable storage is consequently required. Cosatto and Graf demonstrated a similar image-based system, but split the face into features and modelled it as a small set of quadrilateral planar surfaces in 3D in order to deal better with rigid head movements (Cosatto & Graf, 2000). They handled coarticulation by using the generative footage to build a space relating phonemic information to their image-based model. For a novel timeline of phonemes, a cost function was minimised to calculate the best trajectory in this space through the target phonemes. A minimal amount of expressive detail could also be added by animating around the eyes and eyebrows for emphasis.

The latest contribution to date from Ezzat, Geiger and Poggio took a similar approach to coarticulation to Cosatto and Graf, but also improved on their own previous face model (Ezzat, Geiger, & Poggio, 2002). Instead of selecting a prototype image for each phoneme, frames in the centre of naturally occurring clusters were used. In order to find these, all frames were transformed to greyscale and PCA was performed to reduce the dimensionality of the data for computational efficiency purposes. Fifteen principal components were retained as a basis and all frames were projected into this PCA space. It was assumed that 46 multidimensional Gaussian clusters could model the scattered data in this space and k-means clustering was used to estimate their means and variances (Bishop, 1995). For each of the clusters, the image closest when projected into the PCA space was retained as a prototype. One of these prototypes was then arbitrarily chosen as a reference, and optic flow was used to find pixel-wise correspondences between this and the other 45 prototypes.
(the flow field onto itself was set to zero). The 46 flow fields and 46 prototype images were then used to parameterise a 92-dimensional space of mouth movements. Existing and novel frames could then be generated from this parameterised model. By linearly combining the flow-fields, new mouth shapes could be produced and best approximations of the resulting images could be synthesised by warping the prototype images onto these shapes and blending. These new mouth shapes could then be composited onto the original footage.

These image-based approaches are similar to the approach discussed in this thesis, in that existing known movements are used to synthesise new frames, but the voice of the other actor, or speech synthesiser, drives the mouth alone and the driver’s facial expressions and visible emotions are totally ignored. Only Cosatto and Graf alter anything other than the mouth and they merely add minor eyebrow movement for emphasis.

3.3 - Principal Components Analysis on Faces

PCA plays an important part in the work presented in this thesis, in the parameterisation of facial motion, so it is necessary to review its use in the context of faces. The technique has been applied widely on static images of faces of varying identity, but there are few examples of its application on moving faces.

3.3.1 Images of Faces

Sirovich and Kirby were the first to apply PCA to images of faces, principally as a means of data compression (Sirovich & Kirby, 1987). Face images were turned into vectors by concatenating rows of pixel-wise grey level intensity values and transposing. They demonstrated how the weighted sum of just a small number of principal components could be used to reconstruct recognisable faces, requiring only the storage of the weights. The principal components extracted from sets of facial images in this way are often termed eigenfaces and have been successfully applied since, particularly for facial recognition (Pentland et al., 1994; Turk & Pentland, 1991).

PCA has been applied much less often on images of the same face in motion, but examples of its use are mentioned earlier (see 3.2.3 and 3.2.4), particularly
for data compression in order to enable transmission of image patches at low bandwidth (Koufakis & Buxton, 1999) or just for computational efficiency (Ezzat et al., 2002).

### 3.3.2 Separated Shape and Texture

A problem with the application of PCA on intensity values of images, however, is blurring, since linearly combining images results in deterioration of sharp edges. By first aligning face images onto a standard shape, blurring can be dramatically reduced. Craw and Cameron introduced such a technique by manually marking a set of key points on faces and finding their average positions over the set (Craw & Cameron, 1991). The faces could then be warped such that the key points fell in the same place for all images. PCA then yielded much sharper results. Information is lost however, since all features in the components always remain in a fixed position. One way of avoiding this is by applying a second PCA on the key points and exploiting these together. The feature-aligned image is generally referred to as either texture or shape-free information and the positional data encoding feature locations is usually known as shape.

Hancock employed shape and shape-free PCAs in conjunction with a genetic algorithm in a prototype system for witness identification (Hancock, 2000). He demonstrated how the global nature of PCA could address the shortcomings of the feature-based Photofit and Identikit systems, as discussed in Chapter 2. Following PCA on a set of example faces from a database, novel sets of shape-free faces could be generated with a random linear combination of the texture components, and then warped with a random linear combination of the shape components. The user could then rate the similarity of each face in the set to the target and a genetic algorithm would generate a new set, evolved using this information, with the aim of converging at each step closer towards the desired target.

Shape and texture information do not necessarily need to be dealt with separately. The two vectors can also be extracted and then combined into a single long vector. Beymer, and Vetter and Troje presented such vectorisations using optic flow to find dense pixel-to-pixel correspondences between images (Beymer, 1995; Vetter & Troje, 1995). Once flow fields were extracted from
each face to a chosen reference face, these could be averaged to find the mean shape and, for each face, shape could be encoded as the flow field deviation from this mean. By then warping faces onto the average, shape was removed, leaving only texture.

### 3.3.3 Laser Scanned Heads

PCA has also been applied to laser-scanned heads for defining a low-dimensional space for 3D head shape. This has particularly been applied for constraining the space of three-dimensional face meshes in estimating face shape from individual photographs of faces (Atick, Griffin, & Redlich, 1996; Blanz & Vetter, 1999).

### 3.3.4 Dot Tracking Data

Although PCA has often been used to find axes of variation between people, variations within people have been considered less often. PCA has been applied to dot tracking data from facial sequences. Arslan et al. used it simply for dimensionality reduction in building codebooks relating acoustic data and phonemes to three-dimensional positions of dots for speech-driven facial animation (Arslan & Talkin, 1998). Kshirsagar et al. used PCA on these vectors of dot positions and mapped a configuration associated with each phoneme into the principal component space (Kshirsagar, Molet, & Magnenat-Thalmann, 2001). New facial sequences could then be generated from phonemes by fitting a smooth path in this space between co-ordinates of key-frames via cubic spline interpolation, thus dealing automatically with co-articulation effects in a manner similar to that of Cosatto and Graf, discussed earlier (Cosatto & Graf, 2000).

### 3.3.5 Dynamic Laser-scanned Heads

Laser-scanners are not currently able to capture moving faces, since the laser must rotate 360° around the head whilst the subject remains still. Kuratate et al. circumvented this difficulty by capturing laser scans of a face in eight different poses and using PCA to reduce the dimensionality of the data (Kuratate, Yehia, & Vatikiotis-Bateson, 1998). By relating the positions of a small number of points on the meshes to their principal component scores via a linear estimator,
they were able to drive the 3D mesh by tracking points positioned analogously on an actor.

### 3.3.6 Facial Motion from Optic Flow

PCA has also been applied to optic flow measures in video samples of mouths in speech. Fleet, Black, Yacoob and Jepson demonstrated how optic flow fields could be parameterised by linearly combining flow fields from a basis set (Fleet et al., 2000). An example they used was in collecting 3000 training images of the mouth of a speaker articulating the words “centre”, “track”, “print” and “release”. Optic flow was calculated between consecutive frames and submitted to a principal components analysis. The first seven components (accounting for 91.4% of the variance in the set) were then extracted as a basis for use in robust tracking, with the assumption that they spanned (or approximately spanned) the full space of permissible mouth motions. Novel flow fields around the mouth could be parameterised as a linear combination of the basis flow fields.

### 3.4 - Summary

A variety of techniques for tracking facial motion have been introduced. Dot, contour and feature tracking capture movement at a small set of locations on the face and require manual registration of these points with the model. Usually markers, or some form of highlighting is required for effective tracking. Optic flow techniques, however, require no markers or highlighting and some algorithms can produce accurate and robust estimates of motion at every pixel location. Registration of every pixel in the face with its corresponding location in the model is still required, but this can be estimated by warping the flow field based on a small number of registered features. Once movements are known and are registered with the model, the process of animation is still non-trivial. Most recent approaches use a complex, hand-coded underlying muscle model and activate these muscles using a control loop.

These tracking techniques have mainly been used to drive 3D models of the face, but 2D approaches can be used to simplify the problem and avoid coding complicated muscle models by morphing between real prototypical face images.
Such approaches, as summarised, have mainly been applied to drive faces from phoneme data and manipulate the lips alone, so any additional expressive information is lost. Most of these models require considerable manual intervention in their generation.

The work presented in this thesis employs techniques, such as principal components analysis, to automatically generate and parameterise 2D facial models from example footage of a face in motion. Previous applications of PCA on facial movement have been shown. Although this is a common technique applied in the coding of facial identity, it has rarely been applied to faces in motion. The examples presented, however, demonstrate its applicability to motion tracked with a variety of techniques.
In this chapter, a technique for generating example-based models of the face is presented. All that is required is an example sequence of the target face in motion. Each frame of the sequence is considered to be an example configuration and these can be provided in any vectorised format. Probably the most basic vectorisation relies simply on a list of the frame’s grey-level pixel values. Naturally, the richer the set of example movements, the more information there is to learn from. Also, the better the quality of the vectorisation, the better the quality of the resulting model. It is shown how PCA can be applied to these example vectors in order to extract a smaller set of orthonormal vectors forming a basis that closely spans the set. It is then demonstrated how a generative model of the target face can be produced, based on these principal components.

The approach is initially introduced in the simple context of a pixel-wise intensity vectorisation, and then more elaborate vectorisations are discussed along with their merits.

4.1 - An Example-based Approach to Modelling Faces

4.1.1 Facial motion as pixel-wise intensity variations

Consider an image of width $w$ and height $h$ to be an $h \times w$ matrix, $X$, of grey level intensity values, one value for each pixel of the image, where $X_{ij}$ 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 (Figure 4.1).
This vector (of length $N = w \times h$) can be thought of as representing a location in an $N$-dimensional space. Now consider a set of $M$ frames from a continuous recorded sequence of a face vectorised in this manner, $x_1, x_2, \ldots, x_M$. This method for vectorising images of faces has commonly been used for principal components analysis of facial identity (Sirovich & Kirby, 1987; Turk & Pentland, 1991).

### 4.1.2 Centring the Examples

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, $\mu = \frac{1}{M} \sum_{i=1}^{M} x_i$. Considering $\mu$ as a reference, each face, $x_i$, in the set can be considered as a linear translation, $\phi$, from this, $\phi = x - \mu$ (Figure 4.2).
4.1.3 Principal Components Analysis

In order to move around the subspace occupied by these particular vectors, a co-ordinate system can be set up that spans it using the examples as a basis. This space will necessarily have dimensionality of at most $M$, but it is unlikely that this will form a good description, since two or more example faces may be of a similar configuration and image noise will be responsible for most of the variance in those dimensions. By application of principal components analysis, a new improved orthonormal co-ordinate system centred on $\mu$ can be defined, which more efficiently spans this subspace, with axes chosen in order of descriptive importance. That is, basis vectors are defined sequentially, each chosen to point in the direction of maximum variance, unaccounted for so far by their predecessors, subject always to the constraint of orthonormality. Since noise tends to be uncorrelated, vectors describing it will be of low importance in the hierarchy and can be later discarded by truncation to a lower dimensionality.

Principal components analysis is a mathematical technique that seeks to linearly transform a set of correlated $N$-dimensional variables, $\{\phi_1, \phi_2, \ldots, \phi_M\}$ (assumed without loss of generality to have zero mean), into an uncorrelated set that better describes the data, termed principal components, $\{b_1, b_2, \ldots, b_M\}$ (Chatfield & Collins, 1980). $\Phi = [\phi_1, \phi_2, \ldots, \phi_M]$ is defined to be the matrix with columns composed of the $\phi_i$’s. 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 $\{\phi_1, \phi_2, \ldots, \phi_M\}$.
### 4.1.4 First principal component

Consider first $b_1$. This is the first principal component and so must point in the direction of maximum variance of the data set. It is thus necessary to choose the first basis vector, such that the magnitude of the projection of each member of the dataset onto $b_1$ is optimal,

$$
\sum_{i=1}^{M} (\phi_i b_1)^2 / b_1 \cdot b_1 = \sum_{i=1}^{M} (\phi_i b_1)^2
$$

(since orthonormality of the basis set dictates that $b_1$ must have a magnitude of one). This can be represented in matrix form as

$$
(b_1^T \Phi)(b_1^T \Phi)^T = b_1^T \Sigma b_1
$$

where $\Sigma = \Phi \Phi^T$. It should be noted that $\Sigma / M - 1$ is, by definition, the covariance matrix of the set of image vectors (recall that the $\phi_i$’s are centred on their mean) and $1 / (M - 1) b_1^T \Sigma b_1$ gives a measure of the variance in the set that $b_1$ accounts for. Orthonormality adds the constraint that $b_1^T b_1 = 1$. The problem is thus the maximisation of (4.2) subject to the constraint $b_1^T b_1 = 1$. Introducing a Lagrange multiplier (Bryson & Ho, 1975), $\lambda$, a new function, $L_i(b_1)$, can be defined,

$$
L_i(b_1) = b_1^T \Sigma b_1 - \lambda (b_1^T b_1 - 1)
$$

Employing the procedure of Lagrange multipliers, maximisation is now just a case of finding when $\frac{\partial L_i}{\partial b_1} = 0$,

$$
\frac{\partial L_i}{\partial b_1} = 2 \Sigma b_1 - 2 \lambda b_1
$$

Setting $\frac{\partial L_i}{\partial b_1} = 0$, we have
This leaves an eigenvalue problem, where candidate solutions are the eigenvectors of $\Sigma$. Pre-multiplying by $b_1^T$ yields,

$$b_1^T \Sigma b_1 = \lambda_1 b_1^T b_1 = \lambda_1$$

which is the very function to be maximised (4.2), so the optimal solution is necessarily the eigenvector associated with the largest eigenvalue of $\Sigma$.

**4.1.5 Second principal component**

To find the second principal component, $b_2$, it is necessary, similarly, to maximise

$$b_2^T \Sigma b_2$$

subject to orthonormality constraints

$$b_2^T b_2 = 1 \text{ and } b_1^T b_2 = 0$$

With two Lagrange multipliers, $\lambda_2$ and $\delta$, the new function $L_2(b_2)$, can be defined,

$$L_2(b_2) = b_2^T \Sigma b_2 - \lambda_2 (b_2^T b_2 - 1) - \delta b_2^T b_2$$

Again, maximisation is now just a case of finding when $\frac{\partial L_2}{\partial b_2} = 0$, which leaves,

$$\frac{\partial L_2}{\partial b_2} = 2 \Sigma b_2 - 2 \lambda_2 b_2 - \delta b_2 = 0$$

Pre-multiplying by $b_1^T$,

$$2 b_1^T \Sigma b_2 - 2 \lambda_2 b_1^T b_2 - \delta b_1^T b_1 = 0$$

this reduces to

$$2 b_1^T \Sigma b_2 = \delta$$
due to orthonormality constraints (4.8). Rearranging this, a combination of the symmetry of $\Sigma$, (4.5) and (4.8) can be exploited together, to show that $\delta = 0$:

$$\delta = 2b_1^T \Sigma b_2 = 2(\Sigma b_1)^T b_2 = 2(\lambda_1 b_1)^T b_2 = 2\lambda_1 b_1^T b_2 = 0 \quad (4.13)$$

This reduces (4.10) to

$$\frac{\partial L_2}{\partial b_2} = 2\Sigma b_2 - 2\lambda_2 b_2 = 0 \quad (4.14)$$

which leaves, again, the eigensystem,

$$\Sigma b_2 = \lambda_2 b_2 \quad (4.15)$$

Since $b_1$ is already the eigenvector associated with the largest eigenvalue, the next best solution will be the eigenvector associated with the second largest eigenvalue.

**4.1.6 The remaining principal components**

By continuing this process for each $j \in [1, M]$, with the constraints $b_j^T b_j = 1$ and $b_i^T b_j = 0$ for all $i < j$, it is apparent that the principal components are simply the eigenvectors of $\Sigma$ (or, equivalently, the eigenvectors of the set's covariance matrix) ordered by magnitude of their associated eigenvalues, $\lambda_j$,

$$\Sigma b_j = \lambda_j b_j \quad (4.16)$$

Pre-multiplying (4.16) by $b_j^T$ gives,

$$b_j^T \Sigma b_j = \lambda_j \quad (4.17)$$

Since the variance accounted for by $b_j$ is given by $\frac{1}{M-1} b_j^T \Sigma b_j$, the corresponding eigenvalues provide a measure of this, differing only by scaling. With consideration towards these variances, lower order components can be discarded as noise, thus reducing the dimensionality to some $P \leq M$. 

55
Figure 4.3 – First five principal components from a sequence of facial motion vectorised as pixel-wise intensity variations (267 frames at 160 x 240 pixels resolution). Columns show the components -2, 0 and +2 standard deviations from the sequence mean. See Appendix E.1.1 for animated version.

Figure 4.3 shows the first five principal components from an image sequence of Harry speaking, vectorised as described. Together, these mere five principal components account for 75% of the variance in the sequence of 267 frames. In each case, the central column always shows the mean image from the sequence. The left and right columns show the images two standard deviations, $2\sigma$, away from the mean in the negative and positive directions respectively for
each principal component. More explicitly, for row \( j \), from left to right, images are \( \mu - 2\sigma b_j \), \( \mu \) and \( \mu + 2\sigma b_j \), where standard deviation, \( \sigma = \sqrt{\frac{\lambda_j}{M-1}} \).

### 4.1.7 Reducing computation

Computationally, finding the eigenvalues and eigenvectors of the \( N \times N \) matrix, \( \Sigma = \Phi \Phi^T \), is difficult due to its large size. It is generally much easier to find the eigenvalues and eigenvectors of the \( M \times M \) matrix \( \Phi^T \Phi \) when \( M \ll N \) (Sirovich & Kirby, 1987),

\[
\Phi^T \Phi v = \lambda v
\]

This can be exploited, since, pre-multiplying each side by \( \Phi \),

\[
\Phi \Phi^T \Phi v = \lambda \Phi v
\]

and adding some parentheses,

\[
(\Phi \Phi^T)(\Phi v) = \lambda(\Phi v)
\]

it can be seen that \( \Phi^T \Phi \) and \( \Phi \Phi^T \) share the same eigenvalues, and that, if \( v \) is an eigenvector of \( \Phi^T \Phi \), then \( u = \Phi v \) will be an eigenvector of \( \Phi \Phi^T \). This provides a useful computational shortcut.

For particularly large values of \( M \) and \( N \), however, memory constraints sometimes make it impractical to store the matrix of outer products, whether it be \( \Phi \Phi^T \) or \( \Phi^T \Phi \). In such situations, there are techniques where the first \( P \) principal components can be learned by a neural network (Sanger, 1989), or can be extracted using a convergence algorithm (Roweis, 1998). Techniques such as these allow the continuous updating of principal components from a stream of examples and thus indicate some degree of biological plausibility.

### 4.2 - Vectorising Faces by Warping

Many alternative vectorisations could be employed to define the original space of an individual’s facial movement. By simply concatenating the three colour
planes, for example, RGB colour images can be vectorised and the procedure outlined above can be applied.

A clear and well-recognised problem with the examples presented previously, however, is the blur inherent in linearly combining images. Given a facial image sequence, one approach for evading this drawback is to choose an arbitrary frame to be a reference and define the remaining frames in terms of warps from this single frame. Figure 4.4 demonstrates this warping approach. Here each frame, $I$, is represented as a matrix containing the colour information for each pixel of the image, for example as an RGB triple. $I(x,y)$ is written to represent the colour information for the pixel at $(x,y)$. The image shown in (a) is chosen as a reference. Although this is a somewhat arbitrary choice, it is best to ensure that it is in a 'neutral' pose with eyes open and mouth slightly open. This is because, for example, an open mouth can be warped onto a closed mouth, but a closed mouth cannot be warped onto an open mouth. In practice, the frame closest to the luminance mean has been found to be an effective choice.

### 4.2.1 Warping a reference

Assume that the flow field $[U, V]$ can be found, relating each pixel in the target frame, $T$, to its source location in the reference, $R$ (shown in Figure 4.4(b)), where $U$ and $V$ are matrices containing the horizontal and vertical components of the field, respectively, for each location $(x,y)$. $U$ and $V$ effectively specify a set of vectors, one for each $(x,y)$, where the tail lies at $(x,y)$ and the head lies at that point's corresponding location in the reference image. Note that this means that, although $[U,V]$ provides a mapping to warp from $R$ to $T$, the vectors point from locations in $T$ to their sources in $R$. This choice is used because warps in this case are preferentially calculated by backward mapping, denoted by the function $\text{Warp}(R, [U, V])$ in Algorithm 4.1. This involves simply iterating through each location in the destination image and filling it with the colour information from its source location in the reference image:

$$\tilde{T}(x,y) = R(x + U(x,y), y + V(x,y))$$  \hspace{1cm} (4.21)
where $\tilde{T}$ is the reconstruction of $T$. Since $(x + U(x,y), y + V(x,y))$ will rarely correspond exactly to pixel locations in $R$, an interpolation technique is employed. Here bilinear interpolation is used (Appendix A.1).

$$\mathbf{Function: \tilde{T} = Warp(R,[U,V])}$$

$R$ is reference image, $[U,V]$ is flow field, $\tilde{T}$ is resulting warped image

Make a new destination image, $\tilde{T}$, same size as $R$

For every $(x,y)$ in $\tilde{T}$,

1. Find source location, $(x',y')$, in $R$

   $$x' = x + U(x,y)$$
   $$y' = y + V(x,y)$$

2. Calculate nearest integer position, $(x_0', y_0')$, to the lower left of $(x',y')$, and horizontal and vertical distances from it, $(\Delta x', \Delta y')$

   $$x_0' = \text{Floor}(x') \quad y_0' = \text{Floor}(y')$$
   $$\Delta x' = x' - x_0 \quad \Delta y' = y' - y_0$$

3. Apply bilinear interpolation formula (Appendix A.1)

   $$\tilde{T}(x,y) = (1 - \Delta y')[(1 - \Delta x')R(x_0', y_0') + \Delta x'R(x_0' + 1, y_0')] + \Delta y'[(1 - \Delta x')R(x_0', y_0' + 1) + \Delta x'R(x_0' + 1, y_0' + 1)]$$

End

Return $\tilde{T}$

Algorithm 4.1
Figure 4.4 – Representing an example frame, T, as a warp, [U, V], from a reference, R. The frame can be reconstructed by applying [U, V] to R. See Appendix E.2.1 for an animation of the entire sequence reconstructed by warping a single reference.

Mapping from source to destination, or forward mapping (as outlined in Appendix A.3) could be used to warp images instead, but this involves putting pixels from the source into locations in the destination that are not necessarily integer valued. This means that sub-pixel precision must be lost in implementation, filling the nearest pixel to the target location in each case. This will also result in some pixels in the target that are not filled and these must be corrected with a hole-filling algorithm. Backward mapping has neither of these drawbacks and is consequently more precise and much faster to implement.

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] 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.

It is important to note, however, that calculating flow fields between images is a non-trivial task, so it is necessary to outline the methods applied to this end. A variety of optic flow algorithms could be used, but in this work, an adaptation of the Multi-channel Gradient Model was applied.
4.2.2 The Multi-channel Gradient Model (McGM)

The Multi-channel Gradient Model (McGM) is an optic flow algorithm modelled on the processing of the human visual system (Johnston, McOwan, & Benton, 1999). A basis set of spatio-temporal derivatives is calculated by convolving the image sequence with derivative of Gaussian filters. These are then combined 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.

The model was initially applied to just two images for each frame, the reference and the target, since the algorithm provides only an estimate of image motion and it was assumed that errors would be disproportionately magnified for frames temporally further from the reference, were fields to be combined over time. Some adaptation was thus required for this application, since the McGM would usually have a large temporal buffer of images to work with and, in this case, there would be only two. This can be overcome by replacing the zeroth and first temporal derivatives with their average and difference, respectively, and discarding all those of higher order. A coarse-to-fine implementation of the McGM was applied in this manner at three spatial scales, 0.25, 0.5 and 1.0, progressively warping a reference facial image onto the target frame. This is presented in Algorithm 4.2.

```
Function: [U, V] = MultiscaleMcGM(T, R, k_1, ..., k_S)

T is target image, R is reference image, k_1, ..., k_S is list of scales, [U, V] is resulting flow field

Set initial estimate of flow field to zero: [U, V] = [0, 0]

Set initial reconstruction to the reference: \( \hat{T} = R \)

Iterate through scales: For \( s = 1 \rightarrow S \)

1. Resize \( T \) and \( \hat{T} \) by current scale, \( k_s \), yielding \( T_s \) and \( \hat{T}_s \)

\[ T_s = \text{Resize}(T, k_s) \quad \hat{T}_s = \text{Resize}(\hat{T}, k_s) \]
```
2. Calculate flow field from $T_s$ to $\hat{T}_s$ (Note that $McGM(\hat{T}_s, T_s)$ will be written to denote this function)

$$[U_s, V_s] = McGM(\hat{T}_s, T_s)$$

3. Rescale and resize flow field to full-size

$$[U', V'] = \frac{1}{k_s} \times Resize([U_s, V_s], \frac{1}{k_s})$$

4. Update estimate by appending ($\oplus$) onto previous estimate (see Appendix A.2 for implementation details)

$$[U, V] \rightarrow [U, V] \oplus [U', V']$$

5. Update reconstruction

$$\hat{T} = Warp(R, [U, V])$$

End

Return $[U, V]$

Algorithm 4.2

Using this technique, small movements from the reference were tracked well, but larger deviations often resulted in large errors. It was found subsequently that calculating flow fields sequentially from the reference and concatenating them improved estimates, since movements between temporally adjacent frames are generally small. In this algorithm, a reference frame is chosen then the flow field mapping onto the next frame is calculated (denoted $[U^i_0, V^i_0]$, where $[U^j_i, V^j_i]$ denotes the flow field for warping frame $i$ onto frame $j$, and the reference frame is set as frame $0$). From then on, the flow field mapping to the current frame, $n$, from the previous frame is calculated, $[U^n_{n-1}, V^n_{n-1}]$, then it is appended onto the previous full estimate, $[U^{n-1}_0, V^{n-1}_0]$, in order to estimate the flow field mapping from the reference to the current frame, $[U^n_0, V^n_0]_{cont}$, the continuous estimate. If the reference frame is not the first frame of the sequence, the fields before it can be estimated by applying the same process from the reference frame in reverse.
The initial concern that results would degrade in quality as the temporal
distance from the reference increased, was addressed by taking the weighted
sum of this new estimate and the estimate using the multi-scale approach, or
direct estimate, to avoid divergence. Weights were chosen by considering the
quality of reconstructions generated by the two flow fields against the target. If
the reconstruction for one of the flow fields was found to be comparably better
for a particular location in a neighbourhood, it would be weighted highly and the
other flow field would have low weighting at that point. A linear confidence
metric based on the two reconstruction errors was chosen. The details are
summarised in Algorithm 4.3. It was found that application of the McGM on a
single scale for the continuous estimate was sufficient for high quality results,
since movements between adjacent frames were small.

| Function: \{([U^0_0, V^0_0],...[U^M_0, V^M_0]) = EstimateFlowFields(T_0,...T_M) \}
| \( T_0,...T_M \) is an ordered list of target frames,
| \{([U^0_0, V^0_0],...[U^M_0, V^M_0]) \} is a list of respective flow fields
| mapping from the reference, \( T_0 \)
| Let \( n = 0 \). Set the flow field from reference frame, \( T_0 \), to itself to zero
| \([U^0_0, V^0_0] = [0, 0]\)
| For \( n = 1 \rightarrow M \)
| \1. Calculate flow field to warp previous target frame, \( T_{n-1} \),
| onto current, \( T_n \)
| \([U^n_{n-1}, V^n_{n-1}] = McGM(T_{n-1}, T_n)\)
| \2. Calculate continuous estimate of flow field from current
| frame to reference by appending to previous estimate
| continuous estimate: \([U^n_0, V^n_0]_{cont} = [U^n_{n-1}, V^n_{n-1}] \odot [U^n_{n-1}, V^n_{n-1}] \]
| \3. Calculate multi-scale direct estimate of flow field from
current frame to reference

direct estimate: \([U_0^n, V_0^n]_{\text{dir}} = \text{MultiscaleMcGM}(T_0, T_n)\)

4. Calculate squared reconstruction errors for continuous estimate and direct estimate

continuous error: \(E_{\text{cont}} = (\text{Warp}(R, [U_0^n, V_0^n]_{\text{cont}}) - T_n)^2\)

direct error: \(E_{\text{dir}} = (\text{Warp}(R, [U_0^n, V_0^n]_{\text{dir}}) - T_n)^2\)

5. Find weighted sum of estimates

\([U_0^n, V_0^n] = \left(\frac{E_{\text{dir}}}{E_{\text{cont}} + E_{\text{dir}}} [U_0^n, V_0^n]_{\text{cont}}\right) + \left(\frac{E_{\text{cont}}}{E_{\text{cont}} + E_{\text{dir}}} [U_0^n, V_0^n]_{\text{dir}}\right)\)

(if \(E_{\text{cont}}(x,y) + E_{\text{dir}}(x,y) = 0\) at any points, equal contribution is assumed, to avoid division by zero)

End

Return \{[U_0^0, V_0^0], \ldots, [U_0^M, V_0^M]\}\)

Algorithm 4.3

4.2.3 Results

Figure 4.5 shows the first five principal components from Harry’s sequence, vectorised as warps from a reference as described above. The middle column shows the chosen reference image and the left and right columns show the warp \(-2\) standard deviations and \(+2\) standard deviations respectively in the direction of each shown component. Together, these five components account for 85% of the variance in the whole set.
### Principal Component -2σ 0 +2σ Variance

<table>
<thead>
<tr>
<th>Component</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Pc 1:</td>
<td></td>
<td></td>
<td></td>
<td>62%</td>
</tr>
<tr>
<td>Pc 2:</td>
<td></td>
<td></td>
<td></td>
<td>14%</td>
</tr>
<tr>
<td>Pc 3:</td>
<td></td>
<td></td>
<td></td>
<td>4.6%</td>
</tr>
<tr>
<td>Pc 4:</td>
<td></td>
<td></td>
<td></td>
<td>2.7%</td>
</tr>
<tr>
<td>Pc 5:</td>
<td></td>
<td></td>
<td></td>
<td>2.0%</td>
</tr>
</tbody>
</table>

**Figure 4.5** - First five principal components from a sequence of facial motion vectorised as warps from a reference (267 frames at 160×240 pixels resolution). Columns show the component warps -2, 0 and +2 standard deviations from the reference. See Appendix E.2.2 for an animated version.
4.3 - Vectorising Faces by Morphing

A good approximation of a target face can be reconstructed by applying a flow field to warp a reference. However, warping images alone cannot capture some changes. For example, in Figure 4.6d, the failure to replicate the shadow under the eyebrows in the target image occurs because there is no shadow in the reference image to warp into position, so a correct mapping is impossible.

![Figure 4.6](image)

If the warp were to be attempted in the opposite direction, it would be necessary to transform from an image without teeth to an image with teeth, a clearly impossible task. Changes such as these will be referred to as *iconic* and occur whenever a feature present in the target image is not present in the reference image, or vice-versa. The similar changes in shadow or lighting, such as the...
change noted at the beginning of this paragraph, will be referred to as *lighting* changes.

Naturally, careful choice of the reference image will minimise the occurrence of such changes, but these can rarely be totally eliminated. It should be noted that the failure to totally close the mouth in Figure 4.6d occurs because of the disappearance of the teeth, rather than the emergence of a new feature. The presence of a feature in one of the images, that is not visible in the other, will always be problematic, since there is no correct mapping for those points.

These problems can be overcome by additionally encoding the image information for each frame. This motivates a vectorisation based on morphing, a combination of warping and image blending.

### 4.3.1 Image Morphing

Given the reference image, \( R \), and target image, \( T \), from Figure 4.6, suppose a smooth transition is desired between the two in \( N \) steps (in this case, \( N = 5 \)). Starting with the reference image, an increasing amount of the target image could be added, while simultaneously reducing the contribution of the reference.

\[
I_k = \frac{N - 1 - k}{N - 1} R + \frac{k}{N - 1} T
\]  

(4.22)

This is illustrated in Figure 4.7a. Effectively, as \( k \) increases from 0 to \( N - 1 \), the weight on the first image decreases from one (full contribution) to zero (no contribution), whilst the weight on the second image increases from zero to one, with the two weights always summing to 1 for each \( k \).

This approach will yield blurred results, since none of the features will be aligned and the effect is a fade from one image to the other. Note that in the central frame of Figure 4.7a two superimposed mouths are clearly visible.

An alternative could be to progressively warp the first image towards the second as shown in Figure 4.7b, where the flow field is \([U, V]\) and each result is denoted \( F_k \).

\[
F_k = Warp(R, \frac{k}{N - 1}[U, V])
\]  

(4.23)
Conversely, the second image could be progressively warped towards the first, as shown in Figure 4.7c. The previous flow field can be reversed by applying the function given in Appendix A.4 (Algorithm A.3), \([P, Q] = \text{Reverse}(U, V)\). Results of this warp are denoted \(B_k\).

\[
\mathbf{B}_k = \text{Warp}(T, \frac{N-1-k}{N-1}[P, Q])
\]  

(4.24)

Comparing the two rows, the features will be aligned for each column, with the quality of the warps decreasing as they progress farther away from their respective sources. By warping and simultaneously fading from the reference to the target, a process referred to as morphing, a better quality transition, \(M_k\) (Figure 4.7d), will result.

\[
\mathbf{M}_k = \frac{N-1-k}{N-1} \mathbf{F}_k + \frac{k}{N-1} \mathbf{B}_k
\]  

(4.25)

Linear addition of images (Figure 4.7a) and warping (Figure 4.7b and c) can thus be combined to eliminate their weaknesses and capitalise on their individual advantages. A further improved vectorisation can thus be envisaged, based on morphing.
Figure 4.7 – Morphing: a) blending from image $R$ to image $T$ by weighted image addition; b) warping from $R$ to $T$ (left to right); c) warping from $T$ to $R$ (right to left); d) morphing – a combination of warping and blending
4.3.2 Morph Vectorisation Procedure

The morph vectorisation described in this section enables the combination of image warping and image blending for realistic synthesis of facial movement without blur, and without losing iconic or lighting changes.

As in the case of the warp vectorisation, a reference frame must be chosen, although it is now demonstrated how the arbitrariness of selection can be reduced by generating a mean morphed frame for the sequence and adjusting flow fields to register instead to this. The same procedure is followed as before, first choosing an arbitrary frame from the sequence and finding the correspondences between this and all other frames. By then averaging these flow fields, the mean warp can be calculated, \([\bar{U}, \bar{V}]\). Warping the reference with this flow field gives the mean frame for the sequence for the warp vectorisation, but morphing involves taking image information into account also. The morph mean frame can thus be calculated by adjusting each flow field, \([U, V]\), to register to the warp mean instead of the original reference. This involves reversing the mean flow field, using the process outlined in Appendix A.4, so it maps from the (currently non-existent) mean to the reference, then appending the original flow field for each frame onto this. The operation of concatenation is denoted by the symbol: \(\oplus\) (see Appendix A.2). This yields \([U', V']\), the adjusted flow field allowing one to warp from the mean frame onto the target frame.

\[
[U', V'] = Reverse([\bar{U}, \bar{V}]) \oplus [U, V]
\] (4.26)

All frames can then be reverse warped using these adjusted flow fields to align all features, and averaged to actually visualise the morph mean frame. Figure 4.8a shows a selection of frames from a sequence, deliberately chosen for iconic differences. The flow fields relating each frame to the mean shape are superimposed (calculated using (4.26)) and the resulting reverse warps onto the mean shape are illustrated in Figure 4.8b. It should be evident that all features are aligned in this row and averaging all warp aligned frames from the sequence results in a sharp image, Figure 4.9a. If a movie was played of perfect warp aligned images, only iconic and lighting changes would be visible. The images in Figure 4.8b seem quite different since they were selected to have a variety of iconic features, but normally these frames will look quite
similar. The flow fields encode the positions of features in the face. This information will now be referred to as shape.

Figure 4.8 – some example frames from a sequence: (a) with flow fields onto mean shape superimposed; (b) reverse warped onto mean shape

Using this morph mean as a reference and warping can itself provide a variation on the original warp vectorisation (Figure 4.9). The mean perhaps provides a more sensible reference, since it is more representative of the sequence and is most probably closer to each frame on average. By its nature, however, some of the texture will be blurred in the averaging process, so results will not be as sharp as when a genuine frame is used.
These adjusted flow fields are then stored to encode shape information, as a list of $2N$ numbers, consisting of the $x$ component and $y$ component for the vector associated with each pixel of the image. The changes not captured by warping are then encoded by taking each frame and its shape information and reverse warping it onto the mean face shape, as in Figure 4.8. This step aligns all features and remaining variations should then only be attributable to lighting, occlusions and image noise. This data is referred to as texture and is stored as a list of $3N$ numbers in this case, one value for each of the red, green and blue channels for each pixel.

Once shape and texture have been separated, each frame is finally then encoded by concatenating the shape information and texture information into one long vector (Figure 4.10).
Figure 4.10 – Vectorising faces as morph vectors: (a) texture information; (b) shape information; (c) texture and shape concatenated together into one long vector; (d) original target frame; (e) reconstruction generated by warping (a) with flow field (b). See Appendix E.3.1 for an animation of the entire sequence reconstructed from the morph vectorisation.

The original frame can be reconstructed simply by separating the vector into its shape and texture components and warping the texture with the shape information as shown in Figure 4.10e.

It should be noted that the shape and texture components are not adjusted to balance their contributions to the vectorisation as some authors suggest (Cootes & Taylor, 2001). It could be expected that texture information would overwhelm the shape components in variance, but it is important to note that the textures are shape aligned, so the variance within these aligned image frames is minimised. To illustrate: in a perfect scenario with veridical motion fields and no iconic or lighting changes, texture information would not change at all throughout the sequence. In addition to this, the respective contributions of shape and texture are irrelevant in the context of model generation and any weighting would have to be reversed for reconstruction of image frames, adding unnecessary overhead.
4.3.3 Results

Figure 4.11 shows the first five principal components from Jason's sequence, vectorised using the morph basis described above. The middle column shows the texture image with the shape information (flow field) superimposed. The left and right columns show the morph $-2$ standard deviations and $+2$ standard deviations respectively in the direction of each shown component. Together, these five components account for 35% of the variance in the whole set.
Figure 4.11- First five principal components from a sequence of facial motion vectorised as morphs. Left and right columns show the component morphs $-2\sigma$ and $+2\sigma$ standard deviations from the sequence mean, respectively. The central column shows components separated into texture information (image) and shape (superimposed flow fields). See Appendix E.3.2 for an animated version.
4.4 - Summary

It has thus been shown how example footage of a face in motion can be mathematically analysed and broken down into a small set of basis movements, specifically using principal components analysis, although other techniques could also be used. The resulting basis vectors effectively capture constituent facial movements from the example set. A small set of examples will generate a heavily biased model, but as the number of examples increases, these biases will reduce.

This process was initially presented in the context of a simple vectorisation, in the form of a list of intensity values for the pixels in each frame. This method of vectorising frames, however, is blurry, since feature boundaries are not aligned over images and linear combinations blend them away.

An alternative vectorisation was suggested, in which an arbitrary reference frame is chosen and all other frames are encoded as vector flow fields, which can be applied to the reference in order to re-synthesise the original frame. It was shown how the resulting basis evaded the problem of blurring, but other weaknesses were uncovered. Iconic changes, that is, image changes due to occlusions, and changes due to lighting could not be encoded by simply warping a single image.

A third vectorisation was finally presented, in which the respective advantages of both previous vectorisations could be exploited, without their weaknesses. This involves morphing, combining warping and blending. It was shown that resulting basis vectors are able to capture iconic and lighting changes, whilst retaining sharpness.
Chapter 5- Facial Movement Analysis

In the preceding chapter, techniques were presented for encoding the movement of the face by vectorising a set of examples in a logical fashion and extracting a basis set from these in order to parameterise a model of the face. Principal components analysis was employed to calculate these basis vectors. In this chapter, the resulting principal components are discussed and compared to existing facial motion encoding schemes.

This chapter begins by summarising existing manmade schemes and the previous applications of PCA in the context of facial motion. It then continues to compare the principal components from chapter 4 to these hand-crafted schemes.

5.1 - Existing Facial Action Coding Schemes

5.1.1 The Facial Action Coding Scheme (FACS)

In behavioural studies, there are occasions when facial expressions need to be recorded in a comparable and interpretable manner. The task of schematically classifying individual facial expressions was addressed by Ekman with the introduction of the Facial Action Coding Scheme (FACS), a system used widely since by psychologists (Ekman & Friesen, 1978). Facial expressions are attributed to manually defined action units, such as "inner brow raiser" (AU1) and "lip stretcher" (AU20), chosen on an anatomical basis to break down facial actions into the smallest possible atomic units.

5.1.2 FACS+

Essa and Pentland noted that FACS was an inherently static system and sought to improve on it, by using an optic flow algorithm to automatically track the movement of the face in video sequences (Essa & Pentland, 1997). Motion was
coupled with a hard-coded geometric muscle-based model of the face and movements were interpreted as actuations of those underlying muscles. Temporal profiles of these actuations could then be acquired and examined and interactions between action units could be observed as expressions develop. This system was named FACS+.

5.1.3 MPEG-4

A similar scheme to FACS was introduced into computer graphics, as part of the MPEG-4 compression standard, to represent face pose information as facial action parameters (FAPs) for the animation of synthetic faces (Ostermann, 1998). By standardising facial action parameters, animations of the face could be massively compressed by reducing the complex movements of polygonal face models into a small set of parameters.

5.2 - PCA for Coding Facial Action

The above examples are coding schemes created by hand. The PCA-based approach discussed in this thesis instead extracts an encoding system independent of a presumed model of the face, and based solely on the variety present in provided exemplars.

5.2.1 Previous Applications of PCA in Facial Motion Analysis

As mentioned earlier, although PCA has often been used in the encoding of identity, it has rarely been considered as a tool for encoding facial 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). Since their results were used to parameterise the movements of the
mouth, the principal components were used in a similar manner to the warp vectorisation presented earlier in 4.2, 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 analysed 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 in 4.3 and their results will be discussed.

To summarise, the principal components of natural facial motion have not previously been extracted and analysed. 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 techniques discussed in Chapter 4 allow principal components to be obtained from natural sequences of facial motion.

5.2.2 Methods

In order to make results more generalisable, ten individuals were filmed. To encourage expressiveness, sequences of the volunteers telling jokes were captured on video, positioned facing the camera. All participants told different jokes. These were then transferred onto computer and cropped to 160x240 pixels to contain mainly the face.

The sequences were then vectorised as morph vectors as illustrated previously in 4.3, and a basis set was extracted for each participant using principal components analysis. The resulting principal components can be seen in Appendix B, figures B.1 – B.10.

5.2.3 Comparison of results

These principal components of facial movement are very different from Ekman's Facial Action Coding System. FACS encodes facial action as individual muscle activations, but the principal components found here tend to be much more
global in nature, thus fitting the observation that facial muscles rarely function independently.

The principal components capture the correlations between various muscles, and serve to encode high-level facial movements, such as the constituent mouth shapes for speech. Early components quickly account for rigid movement, leaving later components to fill in more subtle movements. It should be noted, however, that rigid and non-rigid motion do not appear to be separated in the principal components. There is strong correlation between rigid and non-rigid motion, perhaps because of the input. Since the input was sequences of speech, the lips were moving almost continuously, whereas in genuine conditions faces are not always experienced talking, often just moving rigidly or communicating emotion. Perhaps with an alternative form of input, or even just more input, a separation may occur.

PCA on different sequences yields different sets of components. An individual component from one sequence may be comparable with a component from another, but these rarely occur at the same point in the hierarchy.

**5.3 - PCA and the human visual system**

The principal components of natural image patches have already been shown to closely resemble receptive fields of cells in the visual cortex (Hancock, Baddeley, & Smith, 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 principal components analysis on static images of faces and encouraging parallels were found (Hancock, Bruce, & Burton, 1995; Hancock, Bruce, & Burton, 1998; Hancock, Burton, & Bruce, 1996; O'Toole, Deffenbacher, Valentin, & Abdi, 1994). O'Toole et al. found that how well a face could be reconstructed using eigenfaces could predict how memorable it was for human subjects (O'Toole et al., 1994). 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). Hancock et al. showed that separating faces into shape and texture and
performing PCA on these components improved further correlations between the quality of reconstruction and how memorable it is in recognition trials for human subjects (Hancock et al., 1995; Hancock et al., 1998; Hancock et al., 1996).

Calder et al. applied PCA to static images of faces of varying identity and expression and found that components naturally separated these channels, coding largely either identity information, or expression (Calder et al., 2001). It should be noted that in none of this research is it claimed that principal components analysis is applied by the human visual system; a linear encoding of an analogous nature is instead suggested. Indeed, a technique similar to PCA, independent components analysis (ICA) on natural image patches has been found to produce filters which also closely resemble the receptive fields of simple cells (van Hateran & van der Shaaf, 1998). ICA is a technique similar to PCA, which focuses on maximising statistical independence between components, rather than the variance they each account for.

5.3.1 PCA on a corpus of sequences

Calder et al.’s result, that principal components automatically separate identity and expression, can be tested on natural motion of the face, by applying principal components analysis to the corpus of sequences vectorised as previously discussed. They first need to be altered, however, so that the vectors are suitably comparable. The ten sets of faces were previously vectorised such that any frame could be reconstructed, by separating the respective vector into shape and texture components then warping the resulting texture with the flow field described by the resulting shape. The shape and texture information associated with each frame was generated by registering the original frame to a reference: that sequence’s mean frame.

Collating all ten sets of vectors together in this form, and applying PCA would suffer the same problems of blurring associated with misalignment as the original pixel-wise intensity vectorisation, since the references have different face shapes. It is thus necessary to standardise the references and adjust their associated encoded sequences accordingly. This process will now be explained.
Consider first the standardisation of the references, recalling that the sequence mean frame was previously used. The ten references (sequence mean textures), $T_1, T_2, \ldots, T_{10}$, were aligned in order to find the global mean texture, $\bar{T}$. This alignment was attained by manually locating a set of corresponding points on the ten reference faces, as illustrated in Figure 5.1. A global mean shape was then calculated by averaging their positions (Figure 5.2a). For each reference, $T_i$, a flow field, $[P_i, Q_i]$, could be found to warp it onto this average shape, by calculating the necessary flow vectors at each of the manually delineated points and filling in the missing data by means of Matlab 12.1's `griddata` function, interpolation based on the Delaunay triangulation. Before filling in the missing data, a set of points around the image border (five equally spaced points for each edge, the same for all references) were added to the delineated points, in order to ensure that the filling algorithm was continuous over the outline of the face. Without this step, it was found that warped faces had ragged outlines, where vectors occasionally fell outside the delineated points.

Figure 5.1 – Ten sequence mean textures, $T_1, T_2, \ldots, T_{10}$, with manually located landmark points superimposed
Each reference, $T_j$, could then be adjusted onto the global mean shape, by simply warping with the flow field, $[P_j, Q_j]$. Since hairstyles and the background would introduce irrelevant changes, pixels outside the outline of the face were masked out. See Algorithm 5.1 for the implementation details of the masking procedure. The resulting warped and masked references, $\tilde{T}_1, \tilde{T}_2, \ldots, \tilde{T}_{10}$, are illustrated in Figure 5.3. Averaging these produces the global mean texture, $\bar{T}$, shown in Figure 5.2b. It should be noted that the calculation of the adjusted references is an unnecessary step in the standardisation of the sequences and is performed here purely for illustrative purposes. Only the flow field, $[P_j, Q_j]$, and mask, $M$ (the matrix, containing 1's for all pixels within the global mean
face area and 0’s for those outside), are needed in order to adjust the vectorised sequences. This process is now explained.

Consider now the task of adjusting a vectorised sequence, so that the shape and texture information are encoded with respect to a new reference, $\tilde{T}'$. The adjusted texture information for each frame should be stored on a new shape, such that all pixels are spatially aligned, feature-wise, with the pixels of $\tilde{T}'$. The shape information should then be adjusted, so that it correctly warps the modified texture to still produce the original target frame.

Given the flow field, $[P_i, Q_i]$, and mask, $M$, the texture information, $T$, and shape information, $[U, V]$, were adjusted as follows. Texture information was directly warped onto the global mean shape using $[P_i, Q_i]$ and values outside of the face area were masked to white (pixel level 255), using $M$, to produce $T'$.

$$T' = \text{Mask}(\text{Warp}(T, [P_i, Q_i]), M)$$ (5.1)

For aesthetic reasons, $M$ was convolved with a Gaussian of space constant 1.5, in order to smooth the edges of the mask.

---

<table>
<thead>
<tr>
<th>Function: $T' = \text{Mask}(T, M)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$T$ is an image, $M$ is a mask (containing 1’s for pixels to be kept and 0’s for pixels to be masked), $T'$ is resulting masked image</td>
</tr>
<tr>
<td>Make a new destination image, $T'$, same size as $T$</td>
</tr>
<tr>
<td>For every pixel location, $(x, y)$, and colour channel, $c$, in $T$,</td>
</tr>
<tr>
<td>$T'(x, y, c) = 255 - M(x, y) (255 - T(x, y, c))$</td>
</tr>
<tr>
<td>End</td>
</tr>
<tr>
<td>Return $T'$</td>
</tr>
</tbody>
</table>

---

Shape information was adjusted by simply appending $[P_i, Q_i]$ onto $[U, V]$.
\[ [U', V'] = [U, V] \oplus [P_t, Q_t] \]  \hspace{1cm} (5.2)

The first twenty principal components can be seen in Appendix B, figure B.11.

5.3.2 Results

It is evident when viewing the resulting components that identity indeed naturally separates from facial movement. The first nine principal components seem to capture changes between identities, with most of these strongly coding changes between pairs of people from the training set. The fact that individuals are still visible in these components most probably occurs because there is little redundancy in the set of identities due to the small sample used. With a larger sample, more generic exemplars of identity would be expected.

All principal components from the tenth onwards seem to be of the average identity and exhibit facial movements similar to those extracted in the previous principal components analyses. The 14th is one exception to this, coding mainly variations in the shape of the hairline and is better grouped conceptually with the first nine.

Again, rigid head movements are accounted for early in the hierarchy, starting immediately after the encoding of identity and still do not seem to be independent of non-rigid movements. It can be seen, for example, however, that the 12th component encodes a nodding rigid movement with the mouth transitioning from closed to open, whilst the 13th component exhibits a similar nodding movement with the mouth performing the reverse action. A linear combination of just these two should thus capture the rigid nodding action alone.

Components lower in the hierarchy tend to capture more individualistic movements. Movements that correlate across all sequences are encoded early, since these can account for a large amount of variance, leaving movements that correlate strongly for individuals to be encoded later. This is consistent with previous findings in PCA on discriminating familiar from unfamiliar static faces (O'Toole, Abdi, Deffenbacher, & Valentin, 1993). O’Toole et al. found that more useful information for this task was present in a subset of components in the higher dimensions of PCA-defined face-space. This can be explained by considering that early components encode information common to all faces,
which is thus shared with the unfamiliar faces. Later components encode information unique to the learned set.

5.4 - Conclusions

By examination of the principal components presented in this chapter, it can be seen that a set of movement vectors can be automatically extracted to encode facial action for individuals, or for a generic model based on a corpus of facial image sequences. Application of PCA to the whole corpus of face sequences still extracts a strong set of correlated facial movements, clearly interpretable as generic facial actions.

The results demonstrate that an encoding system can be extracted from facial image input alone, without the need for a hand-crafted heuristic model. The result is an image-based model of the face.

An image-based approach such as this will be inherently viewpoint-dependent, but it is known that the human visual system performs poorly when viewing angles are altered (Hill et al., 1997) and there is neurophysiological evidence of cells responsive to particular views of the face in the primate brain, particularly full-face and profile (Perrett et al., 1991; Perrett et al., 1985).

The resulting encoding thus naturally arises from the example sequences and mirrors some of the known properties of human perception of faces, but movements tend to be global in nature. FACS and MPEG-4 coding schemes are based on localised movements, initiated by single muscles or small groups of muscles. They are thus much easier for human interpretation, but coordinated activity of the face or head may be missed.
Chapter 6- Facial Mimicry

A novel approach to modelling the face was introduced in Chapter 4, involving the study of example footage of the face to be modelled. Example frames are represented as points in a high dimensional space, in a form that allows nearby linear combinations of these examples still to appear as plausible poses of that face. Mathematical tools, such as Principal Components Analysis, are then applied to extract a compact basis set, which spans the examples, and forms the set of virtual strings for the puppet.

Since the extracted basis, certainly in the case of PCA, generally describes global correlated changes, the components will not usually be easy for a puppeteer to control. The goal of this work, however, is performance-driven animation, not to create computer-based puppets for manual manipulation. This chapter now presents methods for transferring an actor’s movements onto the computer-generated model.

6.1 - Mimicry by Projection

6.1.1 The Topography of the Face

Although faces vary widely in appearance, their overall structure is highly standardised. The nose appears roughly in the centre, with a mouth positioned below and two eyes positioned level above, equidistant from the centre. The dimensions of these features are variable, as are their spatial interrelationships, but such variations are generally small, which is why faces can be overlaid so successfully to extract averages or principal components, for example. Even though there is only a small degree of variation in positioning of features, we still are able to discriminate those familiar to us, from others. It is clear that we need to be sensitive to small alterations in the positioning of facial features (Haig, 1984; Haig, 1986a, 1986b).

Faces similarly move in a structured way, controlled by the underlying muscles, which are anatomically organised in a regular fashion for all faces (Parke &
Waters, 1996). These similarities in feature locations and constraints on the movement of the face can be exploited.

If two faces are filmed from a similar viewpoint, then are sufficiently aligned, with a rigid transform consisting of a translation, scale and rotation, in order to align the eyes, for example, features will, in most cases, overlap fairly well. Consider the original pixel-wise intensity vectorisation. Figure 6.1 demonstrates the perhaps naïve approach of performing this vectorisation, then applying deviations from the mean of one sequence to the mean of the second. This produces a synthetic sequence mimicking the original, but imperfections in the alignment, such as in the outline of the face, particularly around the hair, lead to a transparent, ghostly result. There is also a disturbing effect in which the features of the first sequence appear to fade in and out of the second.

![Figure 6.1](image)

Figure 6.1 – a) Selected frames from a sequence of facial motion (185 frames at 160x240 pixels resolution); b) frames from (a) encoded as pixel-wise luminance deviations from the mean (rescaled so mid-level grey is zero); c) frames from (b) added to a second person’s mean (Harry). See Appendix E.1.2 for an animated version.

These problems arise, because the target face does not have the versatility to transform in exactly the same way that the original face could. It is constrained
by the fixed appearance of features, the geometry of the face and the underlying muscles. The facial model developed in Chapter 4, however, learns these constraints, and it will now be shown how this can be used to transform these unrealistic movements into permissible actions in context of the target face.

6.1.2 Projecting actions into face space

Having found a new co-ordinate system representing an individual's face space, any facial movement, \( \xi \), from a sequence of any individual can be projected onto this basis provided it is vectorised in the same manner and centred on its own sequence mean.

Given a set of \( M_{\text{train}} \) training vectors from individual one (the face we wish to drive), \( x_1, x_2, \ldots, x_{M_{\text{train}}} \), and a set of \( M_{\text{drive}} \) driving vectors from individual two (the face that will be doing the driving), \( y_1, y_2, \ldots, y_{M_{\text{drive}}} \), both sets are centred on their means and put into matrices \( \Phi \) and \( \Psi \), such that

\[
\Phi = \begin{bmatrix} \phi_1, & \phi_2, & \ldots, & \phi_{M_{\text{train}}} \end{bmatrix},
\]

and

\[
\Psi = \begin{bmatrix} \psi_1, & \psi_2, & \ldots, & \psi_{M_{\text{drive}}} \end{bmatrix},
\]

where

\[
\phi_i = x_i - \mu_{\text{train}}, \quad \text{and} \quad \psi_i = y_i - \mu_{\text{drive}}.
\]

Principal components analysis, or some other basis extraction technique, provides a set of basis vectors, \( b_1, b_2, \ldots, b_P \), where \( P \leq M_{\text{train}} \). It is easy to project into the new lower dimensional co-ordinate frame provided by the principal components by simply employing the basis transformation matrix, \( B = \{b_1, b_2, \ldots, b_P\} \), where the basis vector columns are normalised to unit length. For example, to project the \( N \)-dimensional vector, \( \psi_i \), into the \( P \)-dimensional subspace described by the principal components basis, apply

\[
c_i = B^T \psi_i
\]

Each element of \( c_i \) represents a weighting on the respective basis vector. In order to transform the projection, \( c_i \), 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 is orthonormal, so \( B^TB = I \), which implies that \( B \) is the inverse transformation, so
In the case of the pixel-wise intensity vectorisation, the new $N \times 1$ vector, $z_i$, is then rearranged into $h$ rows of $w$ elements, to form a $w \times h$ image.

It is important to note that only the one principal components analysis is performed in this procedure and that is on the training vectors alone.

6.1.3 Summary of procedure

1. Load training vectors $x_1, x_2, \ldots, x_{M_{\text{train}}}$

2. Find training mean $\mu_{\text{train}} = \frac{1}{M_{\text{train}}} \sum_{i=1}^{M_{\text{max}}} x_i$

3. Form matrix of centralised training vectors $\Phi = \{\phi_1, \phi_2, \ldots, \phi_{M_{\text{max}}}\}$, where $\phi_i = x_i - \mu_{\text{train}}$

4. Define basis set for face space eigenvector associated with the $i^{th}$ largest eigenvalue of $\Sigma = \Phi \Phi^T$

5. Load driving vectors $y_1, y_2, \ldots, y_{M_{\text{drive}}}$

6. Find drive mean $\mu_{\text{drive}} = \frac{1}{M_{\text{drive}}} \sum_{i=1}^{M_{\text{max}}} y_i$

7. Form matrix of centralised driving vectors $\Psi = \{\psi_1, \psi_2, \ldots, \psi_{M_{\text{drive}}}\}$, where $\psi_i = y_i - \mu_{\text{drive}}$

8. Project vectors into face space $C = B^T \psi$

9. Project back into original space $\Omega = BC$
It could be noted that steps 8 and 9 can be combined to give a direct mapping from the mean-centred driving vectors, the columns of $\Psi$, onto the mean-centred projections, the columns of $\Omega$. This mapping is expressed as $\Omega = BB^T \Psi$. Due to the orthonormality of the basis vectors, the product $B^T B = I$. The reverse, $BB^T$, however, will not give the identity unless the number of components equals the dimensionality of the vectorisation. The number of components must be less than or equal to the number of example frames, so must necessarily be much smaller than the vectorisation's dimensionality in practice.

Although this direct mapping matrix $BB^T$ exists, it will be $N \times N$, where $N$ is the dimensionality of the vectorisation. For most sufficiently informative vectorisations, this will be prohibitively large, so its calculation and use would be impractical.

6.1.4 Results

The next figure demonstrates typical results from this process for the pixel-wise intensity vectorisation defined previously. A 20-dimensional face space was defined for Harry (the first five dimensions of which were shown previously in Figure 4.3). The dimensionality can be truncated at any level, but these first 20 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. Figure 6.2a shows five frames from a real image sequence of the author telling a joke. The sequence was first transformed with a translation, scale and rotation, in order to ensure that the centres of the eyes were aligned with the centres of Harry's eyes. These frames were then vectorised in the same manner as Harry's, then projected into Harry's face space using the procedure defined. The resulting vectors were then transformed back to image space and are shown in Figure 6.2b, below their corresponding frames.
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 Harry has been experienced to be capable of making.

![Image](image_url)

**Figure 6.2** – Pixel-wise intensity vectorisation results: a) Selected frames from a sequence of facial motion (185 frames at 160x240 pixels resolution); b) frames from (a) projected onto the first 20 dimensions of Harry’s face space. See Appendix E.1.3 for an animated version.

These steps can be applied in the context of any vectorisation. Results from the procedure applied to the warp vectorisation are shown below, in Figure 6.3. Five frames from the resulting synthetic sequence are shown in row (b), below their corresponding frames in row (a) (using only 20 principal components, which account for 94% of the variance in Harry’s sequence).
Frame: 20 40 83 104 164

Figure 6.3 - Warp vectorisation results: a) Selected frames from a sequence of facial motion (185 frames at 160 x 240 pixels resolution); b) frames from (a) projected onto the first 20 dimensions of Harry's face space. See Appendix E.2.3 for an animated version.

The same steps can be applied when morphing, although, in the example used here, the features did not align sufficiently well using the eyes alone, so 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. The results of this transform are illustrated in Figure 6.4.

Following this transformation, the novel morph vectors are projected onto the target face space as before (Figure 6.5).
6.2 - Caricaturing and Rescaling

As well as improving alignment of features, projections onto the bases can be additionally enhanced, or exaggerated, or even made subtler, if required, by scaling the coefficients.

6.2.1 Exaggerating Vectors

Since sequences of facial motion 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 just
to make facial motion more coherent in situations where subtle movements would otherwise be missed.

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 vectorisation captures changes in a more realistic manner, so exaggeration will result in more extreme positional changes away from the reference. Exaggerations in the morph vectorisation also benefit from amplified positional changes, but colour changes in the three colour planes will also be accentuated.

6.2.2 Rescaling the Distribution

Coefficients corresponding to different basis vectors do not necessarily need to be rescaled equally. A variety of magnification factors could be applied arbitrarily to the dimensions of face space. A basis vector encoding an unwanted change, for example, could be weighted by zero.

Rescaling dimensions in such an arbitrary manner introduces manual intervention into the process, which until now, could be entirely automated. This can be avoided, however, by considering the statistical properties of the coefficients.

Having projected the frames from a novel sequence into a face-space, the resulting distribution of weights can additionally be modified to match the distribution of the weights of the original sequence. This involves a simple statistical rescaling in which the novel set of weights, \( \mathbf{C} \), is first normalised by subtracting its mean, \( \bar{c} \), for every frame, then dividing by its standard deviation, \( \sigma_c \). It is then transformed to have the mean, \( \bar{d} \), and standard deviation, \( \sigma_d \), of the training set, by multiplying by the required standard deviation, then adding the required mean. This can be written,

\[
\mathbf{C}' = \frac{(\mathbf{C} - \bar{C})\sigma_d}{\sigma_c} + \bar{D}
\]  

(6.3)
where $\bar{D}$ and $\bar{C}$ are the matrices of same size as $C$, with columns of $\bar{d}$ and $\bar{c}$ respectively. The rescaled weights are given by $C'$. This process can be thus applied to adjust the statistical distribution of the weights to be consistent with the original footage. This avoids manual intervention and can be useful in situations where certain idiosyncratic movements are common in the original footage, but projections of another’s movements onto the basis set result in low weightings on the relevant components. This rescaling step can correct for this.

6.3 - Mimicry Across Vectorisations

6.3.1 Creating a Related Basis Set

The method of principal components analysis for generating basis vectors representative of the set of images was discussed in Chapter 4. Since the $P$ principal components (the columns of $B$) are eigenvectors of the covariance matrix $(\Phi\Phi^T)$, it can now be shown that they are thus composed of a linear combination of the original mean-centred dataset (the columns of $\Phi$). The eigensystem can be written as

$$BA_p = \Phi\Phi^TB \tag{6.4}$$

where $\Lambda_p$ is the $P \times P$ diagonal matrix of eigenvalues corresponding to the $P$ eigenvectors in the columns of $B$. Post multiplication by $\Lambda_p^{-1}$ yields

$$B = \Phi\Phi^TB\Lambda_p^{-1} \tag{6.5}$$

Simply denoting $W = \Phi^TB\Lambda_p^{-1}$, it is clear that the principal components are indeed a linear sum of the original dataset,

$$B = \Phi W \tag{6.6}$$

where the columns of $W$ give the weights for each component.

Consider now a second vectorisation of the same frames, $\Phi'$, centred on its own mean, $\mu'_{\text{train}}$. Since the principal components in $B$ were generated from a
linear combination of the columns of $\Phi$, it could be imagined that the same linear combination of the columns of $\Phi'$ could provide an analogous basis set,

$$B' = \Phi'W$$

(6.7)

It is highly unlikely that this basis set would ever be orthonormal itself, but it is generated from the same linear combination of movements as the former, just represented in a different form.

Using these two bases in conjunction, it is now possible to project novel vectors of the first vectorisation into the target face space, then project back out into the second vectorisation. This involves a minor modification of steps 9 and 10.

9. Project back into second vector space

$$\Omega = B'C$$

10. Translate to true origin

$$z_i = \omega_i + \mu_{\text{train}}^t \omega_i$$

denotes the $i$th column of $\Omega$
Figure 6.6 – Relating basis sets. Central columns show first five principal components (-3 and +3 standard deviations from the mean) from an image sequence (267 frames at 160×240 pixels resolution) vectorised as pixel-wise intensity values. Leftmost and rightmost columns show the related warp basis vectors, generated using the approach outlined above. See Appendix E.4.1 for an animated version.
Figure 6.7 – Mapping between vectorisations: a) Selected frames from a sequence of facial motion (185 frames at 160×240 pixels resolution); b) frames from (a) vectorised as pixel-wise intensity values and projected into Harry’s face space; c) same coefficients applied to the related warp basis vectors of Figure 6.6. See Appendix E.4.2 for an animated version.

The facility to map between different vectorisations has important consequences. Low quality, low-resolution images can be used to drive high quality, high-resolution avatars. Although the Multi-channel Gradient Model has been successfully implemented in real-time (Dale, 2002), the modifications discussed in order to extract high quality flow fields for the purposes of warp and morph vectorisations, are computationally intensive and could not currently be processed at frame rate on a standard personal computer. The computation of flow fields could, however, be completely evaded at the driving stage.

The driving frames could be encoded using a fast, low-resolution vectorisation, such as the pixel-wise intensities at half image size, for example. They could then be projected into the thus encoded target face space. This yields a set of weights, which could then be applied to a better quality, higher resolution vectorisation, such as the morph-based vectorisation at full image size, and the
resulting frames could be created from this. This facilitates high quality animations in real-time.

6.3.2 Hand-coding a Related Basis Set

Related basis sets can also be hand crafted. By examining the first few principal component motions, the key movements can be synthesised with a simple parameter set, in order to drive a computer-generated model. Figure 6.8 demonstrates how parameters describing the features on a simplistic cartoon face can be artificially manipulated in tandem with the principal components from a simple pixel-wise intensity vectorisation. Here, a simple set of four parameters describes the cartoon face:

- \( p_1 \) = Mouth width
- \( p_2 \) = Mouth height
- \( p_3 \) = Mouth vertical location (\( y \)-co-ordinate)
- \( p_4 \) = Brow height (distance from eyes)

Columns b and c show the first five principal components from the original sequence \(-2\) and \(+2\) standard deviations from the mean, with a and d showing cartoon faces with the respective parameter changes.
Figure 6.8 – (a) hand crafted and (b) pixel-wise intensity principal component basis sets shown −2 standard deviations from the mean; (c) principal component and (d) hand crafted basis sets shown +2 standard deviations from the mean. See Appendix E.4.3 for an animated version.

By projecting real frames onto the five dimensional basis set, the same weights can be applied to the hand crafted set, in the same way as before, in order to generate performance-driven animation, as illustrated in Figure 6.9.
6.4 - Discussion

In chapter 4, a method for automatically creating computer-generated puppets was presented, by constraining motion into a subspace of permissible actions. The mechanisms of control are not usually intuitive, but the focus of this work is performance-driven animation. In this chapter it was demonstrated how this can be achieved. 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 puppet mimicking the actor’s movements.

Examples were presented, illustrating the technique applied in the case of all three example vectorisations from chapter 4. 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. It would be rather disturbing, for example, to see someone winking, or raising their eyebrows, were they not normally capable of doing so. Provided a sufficiently rich set of motion is captured for the generation of the model, these constraints do not pose a problem.
It has also been shown how 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.

Finally, it was shown how a second analogous basis set can be automatically created, or hand crafted, allowing novel sequences to be encoded and decoded by two different bases. The first basis set can be used to capture the weights required on that basis set in order to imitate the sequence in that vectorisation. Those same weights can then be applied to the second basis set in order to imitate the sequence in the second vectorisation. This enables the generation of high quality animations from low quality footage and the performance-driven animation of hand crafted artificial characters.

The generative phase of the process, in which the avatar is created, is computationally intensive due to the extraction of principal components. For sequences of around 300 frames, 160 by 240 pixels, represented in the morph vectorisation, this can take up to two hours on a 1GHz Pentium III 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 $P \times N$ matrix by an $N$-dimensional vector, followed by the multiplication of an $N \times P$ matrix by a $P$-dimensional vector (recalling that $N$ is the dimensionality of the vectorisation and $P$ 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 achieved at frame rate (under 40 ms) using the vectorisation and configuration above.
Chapter 7- Conclusions

A set of novel techniques for generating and driving photo-realistic computer-generated faces have been presented in the preceding chapters. A summary of these methods is now collated and reviewed in consideration of the current state of the art, bringing forth the relevant advantages and disadvantages. In closing, future opportunities for extending the work are discussed.

7.1 - Summary of Contributions

The contributions contained within this thesis are now summarised. The current typical approach to the problem is first recapped, and then the contributions are reviewed in that context; beginning first by considering the modelling of the face, followed by performance-driven animation and concluding with contributions to facial motion analysis.

7.1.1 The Typical Approach to Performance-driven Facial Animation

Previous approaches to realistically animating the face from an actor's movements have mainly focussed on modelling the face as a three dimensional polygonal surface. Realism is typically attained through many hours of work by a talented artist, or through capturing the geometry of a real face with expensive equipment, such as a laser scanner.

A complex underlying muscle model is then usually added, in order to make the model moveable. This is usually a simplification of the real anatomy of the face, but is still a complex task and much skill is required in order to realistically synthesise muscle movements. Despite high degrees of realism in static views, synthetic faces are usually betrayed by their movements.

The actor's actions are then recorded by tracking the motion of individual features, often highlighted with make-up, or by tracking coloured markers, physically attached to the actor's face. Driving the model then consists of
calculating the artificial muscle movements consistent with the actor’s recorded motion.

7.1.2 Contributions to Modelling the Face

In this thesis, an alternative approach to modelling moveable faces has been presented. The complex three-dimensional surface is discarded for a simpler 2D representation. The handcrafting of an underlying muscle model can then be avoided by using real footage of a target actor to parameterise its movement. It was shown how this could be achieved by encoding each frame from a sequence as a vector and calculating a basis set from these, such that each frame can be reconstructed as a linear combination of the basis vectors. It was demonstrated that principal component analysis could be applied to automatically extract such an orthonormal basis set.

Three vectorisations were presented, beginning with a simple scheme, encoding each frame as a vector of pixel-wise intensity values, extending earlier work on PCA of identity (Sirovich & Kirby, 1987; Turk & Pentland, 1991) to a single face in motion.

Linearly combining images, however, blurs out image features, so a second vectorisation was introduced. An optic flow algorithm was applied in order to register the frames with a single reference image for that sequence. Each frame was then represented as the flow field, which, when applied to warp the reference image, results in the reconstruction of the original frame. PCA has previously been applied to optic flow fields for faces in motion, but has been calculated around the mouth alone and purely for motion parameterisation (Fleet et al., 2000; Yacoob & Black, 1999).

Since occluded features and lighting changes could not be encoded with the warping approach, a third vectorisation was presented, based on morphing, to combine the advantages of both previous methods. This extended previous work on PCA of identity, with separated shape and texture (Beymer, 1995; Craw & Cameron, 1991; Vetter & Troje, 1995), to natural facial movement.

In Chapter 4 it was shown how principal components analysis could be applied to a sequence of frames vectorised in any of these forms to generate a moveable puppet.
7.1.3 Contributions to Performance-driven Facial Animation

Once a model has been generated as above, it was shown how performance-driven animation could be achieved from the movements of an actor without markers attached to the face, or highlighted features. The novel approach involves aligning the driving sequence so that the features are positioned in close spatial locations to the model’s respective features. The driving sequence is then encoded in the same manner as the training sequence and the resulting vectors are projected onto the model with the basis transformation matrices. It was shown how this technique converts an individual’s movements into actions that are permissible in the context of the model’s experience.

It was further demonstrated how a model could be created with a second basis in a different vectorisation, such that each basis vector captures an analogous movement to its twin. It was then shown that movements from an actor could be encoded in the same form as the vectorisation of the first basis set and projected onto it, then reconstructed using the second basis set.

7.1.4 Contributions to the Analysis of Facial Movement

PCA was applied to a variety of sequences of facial motion and it was found that components naturally draw out an encoding scheme, which is different for each individual. Components tend to encode holistic facial motion, rather than local muscle movements, such as those described by handcrafted coding schemes such as FACS and MPEG-4 (Ekman & Friesen, 1978; Essa & Pentland, 1997; Ostermann, 1998).

Calder et al. previously found that PCA could be used to encode facial identity and expression by applying it to faces of various expression and identity, vectorised in a similar manner to the morph vectorisation. The database they used was of static photographs of fixed posed expression. They noted a natural separation of expression and identity in the components. It was shown in Chapter 5, using a set of ten facial sequences encoded in an adapted morph vectorisation, that the same separation effect could be found for natural facial motion. The first nine principal components varied virtually solely in identity, while the remaining components demonstrated facial movements on a constant global mean identity.
7.2 - Critical Assessment

It is now considered how the techniques presented in this thesis compare with the current state of the art. The approach introduced for modelling and parameterising the face is first considered, followed by the method for producing performance-driven animations.

7.2.1 Modelling the face

The two-dimensional approach to modelling the face is much simpler than using three-dimensional polygonal models, with less storage requirements, faster rendering and easier generation. Expensive equipment, such as laser scanners, or the skills of talented artists are required to capture the 3D structure of the face.

Two-dimensional models, however, are less versatile. A 3D model can easily be re-rendered from any viewpoint, whereas a 2D model would need to be completely re-generated from new footage recorded from that viewpoint. The 2D technique presented here is also sensitive to large rigid movements. The optic flow algorithm used to track the face will fail under large head rotations, for example, because of extreme occlusions and reappearances of facial features. Models can thus only be generated from footage in which the head does not undergo large rigid movements and will consequently be unable to perform such movements.

The facility to automatically parameterise the model, however, is an important advantage over 3D approaches. The model is parameterised from experience of real facial movements, so its synthetic actions are consistent with real facial actions, allowing for compelling animation. Three-dimensional models are difficult to parameterise and require plenty of skill and time and ultimately produce less convincing results at the current time.

7.2.2 Performance-driven Animation

The techniques presented for projecting an actor's movements onto a computer-generated puppet remove the need for explicitly tracking dots or features. The face needs simply to be approximately aligned with the computer-
generated puppet, and then global movements can be translated into analogues on the puppet. This is advantageous over previous approaches, since dot and feature tracking procedures record motion only at discrete locations, leaving much of the detail to be filled in and the inevitable loss of information.

Animations are limited by the model, however. Only movements that can be made by combinations of those experienced in the training phase can be projected onto another face, since only those are represented by the basis set. This may seem to be a limitation, but it can equally be considered advantageous, since, only movements faithful to the target’s repertoire can be made. As mentioned previously, it would be unnatural, for example, to see someone wink, or raise their eyebrows, were they not normally able to. It should also be noted, though, that all animation procedures share this limitation, since animators are always constrained by the model they use.

Animations resulting from the methods presented are compelling and highly realistic, but it could be argued that such realism has previously been achieved with other image-based techniques. Most image-based techniques have been applied in the context of generating animations from phoneme input to accompany text-to-speech synthesisers (Cosatto & Graf, 2000; Ezzat et al., 2002; Ezzat & Poggio, 1997, 1998, 2000) or to synchronise lip movements in existing footage with a new annotated audio track (Bregler et al., 1997). Facial poses are associated with phonemes and morphing is used to transition smoothly between these. Results are indeed realistic, but strings of phonemes communicate only information that can be associated with mouth movements and animations are devoid of any emotional content. Indeed, Ezzat, Geiger and Poggio comment on the “zombie-like” appearance of the face animated by phonemes alone (Ezzat et al., 2002). They make their animations more lifelike by superimposing the novel mouth movements onto the original footage, as do Bregler, et al. (Bregler et al., 1997).

In contrast, the approach presented in this thesis has the power to capture and transfer all subtleties of the face’s motion, not just those of the lips. It does, however, require video footage and does not enjoy the advantage of being driveable from the phoneme information alone, a much more compact signal than video.
Generation of the model is, nonetheless, more of an issue in the aforementioned techniques. The training footage must first be annotated with the phoneme information in temporal synchrony with the mouth movements, a highly non-trivial problem in itself. A diverse range of phoneme lists must be uttered in order to generate an effective model, since the movements associated with phonemes are also affected by neighbouring phonemes, resulting in coarticulation effects. Some of the procedures mentioned require the model to be created from footage in which the target is recorded saying a pre-designated phoneme-rich passage of text (Ezzat et al., 2002; Ezzat & Poggio, 1997, 1998, 2000). Models generally end up consisting of enormous databases of phonemes and their associated morphs. Models generated using the techniques contained within this thesis, in contrast, are compact and can be automatically extracted from footage of an actor performing any range of facial movements, provided that they are sufficiently diverse for the purposes for which the model is ultimately required.

7.3 - Future Research

A currently unexplored area for further research concerns the evaluation of the quality of resulting animations. Two fundamental aspects must be addressed in the measurement of quality. The first regards how accurately the synthetic sequence mimics the movements of the original. The second is the general realism of the animation, an important issue considering that the goal of this work is photo-realism.

Consider first a measure for the accuracy of the mimic. The target sequence and its synthetic imitation could be compared frame by frame, using an image-based or image-difference-based error metric, but these will not be suitably informative, since an overwhelming component of image differences will be due to identity, and the evaluation of mimicry must be independent of this by necessity.

The optic flow techniques discussed in this thesis could be applied to compare the motion of the faces in the target and synthetic sequences directly. For each sequence, a neutral reference frame could be chosen and the flow fields registering that frame with all of the others could be calculated. One or both of
the sequences of flow fields could then be warped in order to align the features for fair comparison. Flow fields would then be compared frame by frame in order to gauge their similarity. A shortfall of this approach, however, would be its failure to account for differing underlying muscle structures: in addition to the positioning and general shape of facial features, their freedom of movement across individuals can vary considerably. A mimicked facial expression will thus not result in an identical motion field, an issue central to the approach of projection onto a basis set of permissible learnt actions. In addition, the optic flow techniques discussed provide only estimates and not veridical motion fields, so the inevitable variance due to error cannot be ignored.

A manual evaluation of the mimic could be applied using a coding system such as FACS. This seems more appropriate, since FACS was designed to encode facial actions, rather than facial motion. Both sequences could be independently annotated and then compared. FACS, however, can encode only onset and offset of facial actions, with no notion of their temporal signature or extremity and this is an important limitation.

Although human ratings of similarity would be too subjective to be informative, individual aspects of the mimicked sequence can be evaluated separately in a more objective manner using naïve observers. Mouth movements due to speech can be tested by evaluating the quality of lip synchronization: it could be asked, for example, how often can naïve observers spot that the lips were not recorded concurrently with the audio track?

The testing of photo-realism is a much simpler problem and approaches akin to the previous test can be envisaged. The goal of photo-realism is to produce a synthetic sequence so compelling that it could conceivably be mistaken for genuine footage. This could be tested by showing naïve observers a set of sequences, some genuine and some synthetic, and testing how well they can categorise the sequences as real or fake.

Three image-based methods for vectorising faces were discussed in the preceding chapters, but the modelling techniques presented are generalisable to any vectorisation. A variety of opportunities is available for testing more powerful and versatile vectorisations. One particular example would be an encoding system generated from multiple views of a facial sequence, captured
simultaneously and synchronised. By encoding all views in a single vector, it would be possible to see projected animations from any of the camera angles and intermediate views could potentially be generated by morphing their neighbours (Koufakis & Buxton, 1999). The dual basis encoding system could also be exploited to investigate whether a model of multiple views could be generated from a single view.

Vectorisations need not necessarily be image-based. If a technique for capturing the three-dimensional surface information and corresponding texture of a face in motion was available, this information could be encoded and the same generative and driving techniques could be applied. An opportunity for obtaining this information could be investigated with one of the tools already developed in this work. Following calibration of a pair of camera views (or more, if available), the three-dimensional surface geometry of an object can be extracted if pixel-wise disparities are known between views. The adapted two-frame version of the Multi-channel Gradient Model could be applied in order to perform this registration. The three-dimensional surface of the face could subsequently be extracted and texture mapped. This information could then be collated into a vector for each frame and animation techniques could be applied.

Finally, performance-driven animations, produced using these techniques, currently retain the voice of the driving actor, thus reducing the persuasiveness of the result. Techniques for additionally transforming the voice could further be incorporated into the process in order to correct for this (Arslan & Talkin, 1997).
References


Fraser, I., & Parker, D. (1986). Reaction time measures of feature saliency in a perceptual integration task. In H. D. Ellis & M. A. Jeeves & F. Newcombe
& A. Young (Eds.), *Aspects of Face Processing*. Dordrecht: Martinus Nijhoff.


123


Appendix A - Flow Field
Operations

A.1 - Bilinear Interpolation

With reference to Figure A.1, suppose the pixel value in image $R$ at $(x, y)$ is required, where $x_0$ and $y_0$ are integer-valued and $x$ and $y$ are not ($x_0$ and $y_0$ are found simply by truncating $x$ and $y$ beyond the decimal point, or flooring). $R(x, y_0)$ can be set as the sum of $R(x_0, y_0)$ and $R(x_0 + 1, y_0)$, inversely weighted by the magnitude of their respective horizontal distances from $R(x, y_0)$, since the closer location should exert more influence.

$$R(x, y_0) = (1 - \Delta x)R(x_0, y_0) + \Delta xR(x_0 + 1, y_0) \quad (A.1)$$

Similarly, $R(x, y_0 + 1)$ can be set as a weighted sum:

$$R(x, y_0 + 1) = (1 - \Delta y)R(x_0, y_0 + 1) + \Delta yR(x_0 + 1, y_0 + 1) \quad (A.2)$$
\( \mathbf{R}(x, y) \) can then simply be defined to be the sum of \( \mathbf{R}(x, y_0) \) and \( \mathbf{R}(x, y_0 + 1) \), weighted inversely by their respective vertical distances from \((x, y)\).

\[
\mathbf{R}(x, y) = (1 - \Delta y)\mathbf{R}(x, y_0) + \Delta y\mathbf{R}(x, y_0 + 1)
\]

(A.3)

Substituting (A.1) and (A.2) into (A.3) yields the bilinear interpolation formula:

\[
\mathbf{R}(x, y) = (1 - \Delta y)[(1 - \Delta x)\mathbf{R}(x, y_0) + \Delta x\mathbf{R}(x_0 + 1, y_0)] + \Delta x[(1 - \Delta y)\mathbf{R}(x_0, y_0 + 1) + \Delta y\mathbf{R}(x_0 + 1, y_0 + 1)]
\]

(A.4)

A.2 - Concatenating flow fields

Given three frames, such as those indicated in Figure A.2, suppose the flow field mapping from frame 2 to 1, \([U_2^1, V_2^1]\) (Figure A.3a), and the flow field mapping from 3 to 2, \([U_3^2, V_3^2]\) (Figure A.3b), are known, but the flow field mapping from 3 to 1, \([U_3^1, V_3^1]\), is required.

![Figure A.2](attachment:image.png)
Simply adding the elements of the two flow fields will not work, since the heads of the vectors in \([U_2^3, V_2^3]\) refer to spatial locations in frame 2, which originate from elsewhere in frame 1. This erroneous approach is illustrated in Figure A.4. It is clear that the vectors only occasionally relate the source, at the tip, to the destination, at the tail, correctly.

Instead the flow fields must be correctly concatenated. The first stage involves aligning the vectors from \([U_1^2, V_1^2]\) to the correct new spatial locations in frame 3 by warping them with \([U_2^3, V_2^3]\) (Figure A.5a). Now that the vectors are aligned, the concatenated flow field can be calculated by simply adding on \([U_2^3, V_2^3]\) (Figure A.5a). It can be seen in Figure A.5c that this yields the correct result.
If the operation of appending \([U_j, V_j]\) onto \([U^i, V^i]\) is denoted as \([U_j, V_j] \oplus [U^i, V^i]\), the formula for correctly calculating \([U^i, V^i]\) is:

\[
[U^i, V^i] = [U_j, V_j] \oplus [U^i, V^i] = [U^i + \text{Warp}([U_j, V_j]), V^i + \text{Warp}([V_j, [U_j, V_j]])]
\]  \(\text{(A.5)}\)

### A.3 - Warping by Forward Mapping

Instead of backward mapping, it is possible to warp images by *forward mapping*. This is necessary when flow vectors are known for the integer valued source locations, rather than destination locations. If the mapping is isomorphic, that is, for every source pixel there is a unique destination pixel and vice-versa, then this is a simple case of iterating through the source pixels and placing them in their corresponding destinations. This, however, is rarely the case in the circumstances of image warping. Flow vectors normally point to non-integer locations and some pixels are not filled, leaving holes in the warped image.

An algorithm for performing forward warping is given in Algorithm A.1. For each source pixel, the value is put into the nearest integer-valued location to the desired location. If multiple values are placed into the same location, a weighted sum is taken, with nearer true locations weighted higher. Holes can be filled using Algorithm A.2. This involves filling each hole with an average of its non-empty neighbours, if they exist. This proceeds iteratively until no empty pixels remain.
Function: \( \tilde{T} = FwdMap(R, [U, V]) \)

\( R \) is reference image, \([U, V]\) is flow field, \( \tilde{T} \) is resulting warped image.

Make a new destination image, \( \tilde{T} \), and weight storage matrix, \( W \), both same size as \( R \).

Set all values in \( \tilde{T} \) and \( W \) to zero: \( \tilde{T} = 0, W = 0 \).

For every \((x, y)\) in \( R \),

1. Find nearest integer location, \((x', y')\), in \( \tilde{T} \), and distance of true value from it, \( d \):
   
   \[
   x' = \text{Round}(x + U(x, y))
   \]
   
   \[
   y' = \text{Round}(y + V(x, y))
   \]
   
   \[
   d = \sqrt{(x' - x - U(x, y))^2 + (y' - y - V(x, y))^2}
   \]

2. Inversely weight \( R(x, y) \) by \( d \) and add to value in \( \tilde{T}(x', y') \)
   (note that max possible value of \( d \) is \( \sqrt{\frac{1}{2}} \) due to rounding).

Accumulate weight in \( W(x', y') \):

\[
\tilde{T}(x', y') = \tilde{T}(x', y') + (\sqrt{\frac{1}{2}} - d)R(x, y)
\]

\[
W(x', y') = W(x', y') + \sqrt{\frac{1}{2}} - d
\]

End

For every \((x, y)\) in \( \tilde{T} \),

If \( W(x, y) \neq 0 \), correct for weights

\[
\tilde{T}(x, y) = \frac{\tilde{T}(x, y)}{W(x, y)}
\]

End
Store list of holes, \( h \) (these occur where \( W(x,y) = 0 \)).

Fill the holes: \( \text{FillHoles}(\tilde{T},h) \)

Return \( \tilde{T} \)

**Algorithm A.1**

**Function: FillHoles(\( T,h \))**

\( T \) is image with holes, \( h \) is list of hole locations

Make copy of image, \( T' = T \),

While \( h \) is not empty,

For every \( (x,y) \) in hole list, \( h \),

1. Average all filled immediately adjacent pixels and store in \( T'(x,y) \)

\[
T'(x,y) = \frac{T(x-1,y) + T(x+1,y) + T(x,y-1) + T(x,y+1)}{n}
\]

(\( n \) is number of filled pixels in summation)

2. If \( n > 0 \), remove \( (x,y) \) from \( h \)

End

Update image: \( T = T' \)

End

**Algorithm A.2**

**A.4 - Reversing flow fields**

In order to reverse a flow field it is not sufficient to merely multiply its vectors by \(-1\). The results of this for the flow field in Figure A.6a can be seen in Figure A.6b. It is necessary instead to interchange the head and the tail of each vector.
The negative $x$ and $y$ components of each vector must be moved from their sources to their destination locations as indicated in Figure A.6c.

(a) Original flow field, $[U, V]$

(b) $[-U, -V]$

(c) $[U, V]$ correctly reversed

$\{[P, Q]\}$

Figure A.6

In order to do this, it is necessary to reverse warp the field using forward warping and negate. The resulting function is given in Algorithm A.3.

\[
\text{Function: } [P, Q] = Reverse([U, V])
\]

\[
P = FwdMap(-U, [-U, -V])
\]

\[
Q = FwdMap(-V, [-U, -V])
\]

Algorithm A.3
Appendix B - Results: PCA of Morph Vectors

B.1 - Explanation

The following figures, B.1-B.10 show the first ten principal components from individual facial sequences vectorised as described in 4.3. In each case, components 1-5 appear in the left image and components 6-10 appear in the right image. The central column in each image always shows the average pose with the shape information superimposed (the flow field). Left and right columns show morphs -3 and +3 standard deviations from the mean, respectively. See Appendix E.3.2 for animated versions.

Figures B.11-B.13 show the first thirty principal components from a corpus of sequences (2735 frames at 160×240 pixels resolution), vectorised as described in 5.3.1. In Figure B.11, components 1-5 appear in the left image and components 6-10 appear in the right image, as before. This continues similarly with Figure B.12 and Figure B.13. In each case, the central column shows the average pose with the shape information superimposed (the flow field). Left and right columns in each case show morphs a specified number of standard deviations from the mean. See Appendix E.5 for animated versions.
Figure B.1 – Alison: (282 frames at 160×240 pixels resolution) see Explanation (B.1) for details
Figure B.2 – Glyn: (418 frames at 160x240 pixels resolution) see Explanation (B.1) for details
Figure B.3 – Jason: (287 frames at 160×240 pixels resolution) see Explanation (B.1) for details
Figure B.4 – Joanne: (356 frames at 160×240 pixels resolution) see Explanation (B.1) for details
Figure B.5 – Matt: (246 frames at 160x240 pixels resolution) see Explanation (B.1) for details
Figure B.6 – Polly: (105 frames at 160x240 pixels resolution) see Explanation (B.1) for details
Figure B.7 – Robin: (205 frames at 160×240 pixels resolution) see Explanation (B.1) for details
Figure B.8 – Sovira: (219 frames at 160x240 pixels resolution) see Explanation (B.1) for details
Figure B.9 – Szonya: (407 frames at 160×240 pixels resolution) see Explanation (B.1) for details
Figure B.10 – Tamara: (210 frames at 160×240 pixels resolution) see Explanation (B.1) for details
Figure B.11 – Corpus of facial sequences: The first ten principal components from a corpus of sequences (2735 frames at 160×240 pixels resolution), vectorised as described in 5.3.1. Components 1-5 appear in the left image and components 6-10 appear in the right image. In each case, the central column shows the average pose with the shape information superimposed (the flow field). Left and right columns show morphs –2 and +2 standard deviations from the mean, respectively. See Appendix E.5 for animated versions.
Figure B.12 – Corpus of facial sequences: principal components 11-20 from a corpus of sequences (2735 frames at 160×240 pixels resolution), vectorised as described in 5.3.1. Components 11-15 appear in the left image and components 16-20 appear in the right image. In each case, the central column shows the average pose with the shape information superimposed (the flow field). Left and right columns show morphs –4 and +4 standard deviations from the mean, respectively. See Appendix E.5 for animated versions.
Figure B.13 – Corpus of facial sequences: principal components 21-30 from a corpus of sequences (2735 frames at 160x240 pixels resolution), vectorised as described in 5.3.1. Components 21-25 appear in the left image and components 26-30 appear in the right image. In each case, the central column shows the average pose with the shape information superimposed (the flow field). Left and right columns show morphs –6 and +6 standard deviations from the mean, respectively. See Appendix E.5 for animated versions.
Appendix C - Commercialisation

C.1 - Introduction

The performance-driven facial animation technology presented in the preceding chapters has potential to be of commercial value. In pursuit of this, several avenues for exploitation have been investigated and are subsequently discussed in this appendix. This section begins by giving an overview of the technology concerned and the nature of the exploratory work thus far undertaken. Current providers of similar technologies are then considered in order to frame its position in the market. Three target industries are then individually discussed, examining the merits of developing the technology into a product that would meet their needs. In closing, this is all brought together, motivating the final decision for future developmental work in partnership with a major developer in the telecommunications industry, BTExact.

C.1.1 Commercial Applications

The technology, in its current form, enables the automatic generation of a photo-realistic virtual puppet from existing video footage of an individual’s face in motion. That individual can then be animated to mimic the facial movements of a second actor, without the need for highlighting facial features or physically attaching markers to the face. The tracking and animation processes are also automatic at every stage, except for an alignment step, applied once to ensure that the eyes and mouths of the two faces appear in the same approximate image locations. This step could be automated, however, if required.

Since the resulting virtual puppet is image-based and generated by learning the dynamic properties of a particular individual from example footage, it is highly realistic, even in motion. These properties suggest an application for special effects in television, film and advertisement production.

A virtual puppet generated in this manner can be driven in real-time, so a consumer would be able to see his or her own facial movements appear
simultaneously on another’s face. This could be exploited by video game makers, in order to control the faces of characters within the game with the player’s movements, for example.

Animation of the computer-generated face is achieved by applying weights to a set of basis movements, somewhat similar to pulling on the strings of a real puppet. Individual frames can thus be described with this small list of numerical weights, enabling high quality animations with as few as twenty numbers per frame. It can be seen how this could be advantageous in the telecommunications industry, for example, where there are often limitations on bandwidth.

C.1.2 Steps Taken towards Commercialisation at Present

*UCL Ventures* is the arm of UCL responsible for the management, protection and exploitation of the University’s intellectual property. In December 2001, UCL Ventures applied for a patent in order to protect the technology presented in this thesis. A study was also undertaken at London Business School as part of Dr Susanna Khavul’s New Technology Ventures course in Spring 2002, to assess its commercial viability, focussing mainly on the television, film and advertising industry. The study concluded that the technology was unsuitable, in its current form, for marketing to this industry. A meeting with *BTExact*, the research and innovation arm behind a large telecommunications company, however, was organised through a contact made during the course. Following this meeting, it was concluded that the telecommunications industry was the most appropriate target. A grant proposal was submitted, in collaboration with *BTExact*, to develop the technology further towards this goal and a decision on its approval will be made in March 2003.

C.2 - Current Providers of Similar Technology

There are already a small number of companies providing products that this new technology could compete with or supplement. In this section these companies are considered, beginning with the most directly related and moving on to those of associated application.
C.2.1 Performance-driven Facial Animation

Famous3D\textsuperscript{4} is a subsidiary of Blaze International and is the market leader in the provision of technology for producing automated performance-driven facial animation of comparably high quality. They boast Sony, Midway and Foundation as satisfied customers and their system was used in the recent Lord of the Rings movie. Their product enables the animation of a 3D computer-generated face from motion-capture data, the recorded movement of markers physically attached to an actor’s face. The system relies on an artist “painting” the influence zone on the face model for each marker (see 3.1.2) and thus requires much manual intervention. Famous3D provides software capable of two-dimensional tracking of coloured markers on the actor’s face. A professional 3D dot-tracking system, such as a Vicon\textsuperscript{5} rig, is recommended, however, for high quality animation. The current list price for the V6, a Vicon product sold specifically for face tracking, is £75,000.

Face2Face\textsuperscript{6} is a spin-off company from Lucent Technologies New Ventures Group. It was created in 2000 and focuses on performance-driven animation of cartoon characters without the use of markers, make up, or specialist cameras. They rely principally on the nostrils for locating face positions, so these must always be visible. Although they have not previously targeted high quality photo-realistic animation, their technique avoids markers and thus competes with the performance-driven facial animation technology presented here. Discussions are ongoing with Face2Face regarding the possibility of licensing the technology to them, or starting a joint venture.

In both cases mentioned above, software is provided as a plug-in for a variety of 3D modelling platforms. Famous3D supports 3D Studio Max\textsuperscript{7}, Maya\textsuperscript{8}, LightWave\textsuperscript{9} and SoftImage\textsuperscript{10}. Face2Face also supports all of these except for LightWave.

C.2.2 Text-driven Facial Animation

Some companies are targeting the Internet and mobile communications market, by providing the technology to animate a computer-generated face in tandem with a text-to-speech synthesiser. Such applications involve reading text messages on mobile phones, or information provision on the Internet, to give the Web a more human face. The techniques within this thesis could be
adapted for these purposes, so it is worthwhile considering the current marketplace.

Famous3D offer a product called Impersona, which is a free extension for MSN Messenger, a text-based Internet chat engine. Users are represented by virtual characters that read out their messages accompanied by a text-to-speech synthesiser. In the three weeks following its launch, Impersona claimed to have more than 70,000 active users. Discussions are ongoing for the use of Impersona as a communications tool for web-based tutoring systems. As mentioned previously, Famous3D have the technology to animate 3D photorealistic heads, although they do not specialise in creating them. Anthropics was founded in 1998, to develop and provide technology for text-driven facial animation of cartoon and photo-realistic quality. Their software has recently been announced to be compatible with the new Sony Ericsson P800 handset. Face2Face also provide technology with text-driven capabilities, but currently only for cartoon faces.

C.2.3 Speech-driven Facial Animation

Some companies provide software for driving virtual faces from speech. This is currently not possible for the technology presented here, but further development work, could enable this. Again, Famous3D offer software for this application, with photorealistic capabilities. Another company, Lipsinc also provides such software, which converts audio information into parameters describing the mouth. These products are provided as plug-ins for the 3D modelling and animation packages mentioned earlier; Maya, 3D Studio Max, Lightwave and SoftImage.

C.3 - Television, Film and Advertising

A study carried out at London Business School as part of Dr Susanna Khavul's New Technology Ventures course concentrated on the commercial viability of the technology in the television, film and advertising industry. In this section, the findings are summarised.
C.3.1 Introduction

The technology's place in this industry is as a tool for special effects. During an interview\textsuperscript{12}, David Muir, a managing director at Ogilvy and Mather, an advertising agency, highlighted the dynamics of the production process. He explained the entire process of creating an advertisement in the context of a recent Levi’s commercial. The process begins with a pre-production stage in which a concept (in this case, “Freedom”) is created and a preparatory project is designed and presented to the client. This stage typically takes three to four days. Following the client’s approval, shooting begins, followed by a post-production stage where all the special effects and animation processes take place. The post-production stage is usually the longest part (in this case, six days). The final cost of the Levi’s commercial was £1.4M, and most of this cost was sunk in this post-production stage.

The post-production work in television, film and advertising is contracted out to specialist post-production houses and special effects studios. These companies should thus be the target segment of the industry.

C.3.2 The Post-Production Industry

Post-production companies tend to be small operators with a high level of competition. Soho is the local epicentre of post-production, with about six large post-production houses and about twelve small ones. Amongst the Soho-based companies are The Moving Picture Company, Smoke and Mirrors, The Mill, Gearbox, VTR and The Hive. Work on one movie is rarely contracted out to a single post-production house; usually it is outsourced to many, each working on different computer-generated scenes. There is also a lot of specialisation. Many companies do not even get involved in facial animation, for example.

There is also a hierarchy of production qualities. At the top end of the scale are big budget movies. These tend to be laboriously tweaked to perfection with huge numbers of man-hours employed in order to finally produce results of the highest quality. At the opposite end of the scale are low-budget children’s television productions, which can be produced quickly, often with cost-saving short cuts.
David Muir pointed out that the industry is very price competitive, with shrinking margins and post-production houses are subject to a continuous push for the next generation of products. It is a fragmented industry with a low level of concentration. Their own customers have taken the low bargaining power resulting from this low concentration to their own advantage by compressing their margins.

C.3.3 Performance-Driven Facial Animation in the Post-Production Industry

Following interviews at several post-production houses, it soon became clear that synthetic faces are currently animated in the industry almost entirely by hand. Talented artists are employed to manually manipulate computer-generated models, usually charging approximately £2,000 per second of animation for movie work. Dot-tracking systems are considered to be of insufficient quality, except for lower budget television productions, where facial animation can be invoiced comparatively lower at around £100 to £200 per second. However, even in such situations, dot tracking is not commonplace, in fact, few of the artists interviewed had even heard of Famous3D, for example.

Any time saving steps would give a competitive advantage, but it is crucial, particularly for high budget productions, that quality is not compromised. Current time saving steps involve key framing. The dialogue is broken down into phonemes along a timeline, the animator quickly goes through, adjusting a single frame for each phoneme into the necessary shape, and the computer interpolates between for a first pass animation. This is a tedious task and generally takes one hour for every minute of animated footage. The animator can then adjust the results until he or she is satisfied with the lip-synchronisation. It can be seen how automated techniques could be applied at this first pass stage, where quality is not so much of an issue.

C.3.4 Critical Assessment of the Technology

The techniques presented in this thesis enable automatic photo-realistic facial animation of extremely high quality, something that no other product can currently offer. This is an important advantage and interviewees were particularly interested in the capability to generate models from existing footage and subsequently bring dead celebrities back to life, for example.
It was also noted that dense motion information is captured for the whole of the actor’s face, rather than at discrete locations where features are present, or markers are positioned. This provides much higher quality motion fields than other tracking techniques.

Facial animation in this industry, however, is an art and animators are wary of full automation. An automated first-pass animation could provide timesavings, but it would not be easy introducing such a product into an industry with such a strong culture of manual production.

Facial animation in the industry is most commonly applied for non-human characters where a real actor’s face cannot be used. The generative technology is unsuitable for such applications in its current form, since it relies on being able to learn, from example footage, how the target face moves. Such footage would not often be available in such circumstances. With some manual intervention, however, a handcrafted set of basis movements can be generated for a handcrafted character (in 3D if necessary) to correspond with a driving actor’s basis movements (see 6.3.2). The face can then automatically be animated using the marker-less performance-driven techniques.

The critical disadvantage of the technology in its current form is its image-based approach. It is currently essentially a two dimensional technique. Almost without exception, all animation work in the post-production industry is carried out in 3D. It was noted that if a director was to change his or her mind on a camera angle for a scene, the camera can easily be repositioned and the scene can be re-rendered with minimal effort. This is a common occurrence and would not be practical with a 2D model. The model would have to be totally regenerated from new footage at a different viewpoint.

### C.3.5 Commercial Attractiveness

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 industry. Two-dimensional models may not be appropriate for their needs, but the performance-driven techniques could be used to animate handcrafted models. Dave Child\(^{13}\), from *The Hive*, noted that they can currently output three seconds of lip-synched footage per day, but, with the capabilities that this technology
could offer, he estimated that this could be increased to up to twenty minutes per day.

The post-production industry is highly competitive and this has the potential to give companies a valuable edge, allowing rapid adoption if advantages were indeed seen.

Since the technology is mainly software-based and relies on equipment that is standard in the post-production industry, it is extremely easy to adopt. It would still need to be developed, however, to interface with all the major 3D modelling platforms.

C.3.6 Market Challenges

As seen in C.2, automated facial animation software has existed for some time, although such techniques have certainly not been embraced in the Soho post-production houses visited. A challenge this raises is entry into an already over-stretched market. Since software packages offering facial animation facilities are already available, this technology will not be able to draw from any first mover advantage.

Competitors have a high level of competency and are always striving to improve their technologies. To this end, they may draw from a technology related to this. This system is currently protected by a patent application, but this may be rejected for any number of reasons and there is always the risk that a cunning competitor may find a way around it.

A major challenge in entering the market is the culture of animation. Much work, particularly in facial animation, is painstakingly prepared by hand, frame by frame. Several animators interviewed commented that the technology would never be able to compete with the quality of their work. As mentioned, of those interviewed, most had never used or even seen existing software tools for performance-driven facial animation. The challenge of convincing potential clients of the benefits of an automated or semi-automated solution should certainly not be underestimated.

C.3.7 Summary

The generative technology is not suited for photo-realistic facial animation in the post-production industry in its current form. Further development work is
needed in order to enable the generation of three-dimensional photo-realistic avatars. A two-dimensional approach is too limited in an industry where everything else is 3D.

The marker-less driving aspect, however, could be exploited to speed-up the animation process. The likelihood of adoption, nonetheless, seems a little bleak, given the culture of manual animation and the limited acceptance experienced of similar products. A successful product would need to interface with all major platforms used in the industry and would have to be easy to use, in order to show a strong advantage over existing competitors.

In conclusion, considerable further development work is needed before the performance-driven facial animation technology presented in this thesis can be brought to market in the television, film and advertising industry.

C.4 - Computer Games

As well as in television, film and advertising, facial animation is commonplace in computer games. In this section, the applicability of the technology will be considered in this industry.

C.4.1 Introduction

Pre-generated facial animation is often used in computer games, in a similar fashion to television, film and advertising for introduction sequences, or for story development purposes. Games are often translated into many different languages, so dubbing is a common requirement. However, there are additional applications that this particular technology could enable. Since computer-generated puppets of any individual can automatically be created from video footage, the face of the gamer could be captured and put into the game and animated. This is a feature already offered by a company called Digimask\textsuperscript{14}, who provide software enabling the capture of a 3D head model from photographs. This can then be used in games such as Quake I\textit{liii}\textsuperscript{15}. The performance-driven aspect of the technology could also be exploited, by capturing the gamer’s facial expressions from a standard web-cam, for example, and transferring these onto a character in the game.
C.4.2 The Computer Game Production Industry

Computer game production is a big industry, now grossing more annually than the movie industry, worldwide. Due to the nature of the industry, a constant awareness of the state of the art is necessary to keep up with competitors. Games production is incremental, with each game having to look better than its predecessors in order to sell. There is a constant pressure to get the most out of hardware, and results must be rendered in real-time. There is not the luxury of the supercomputers and render-farms used by the post-production companies, where results do not need to be seen instantaneously.

C.4.3 Commercial Attractiveness

Speeding up processes is again an important advantage that this technology could offer, but this is additionally useful in the case of computer games at rendering time. The image-based nature of the technology at present allows the generation of new poses much faster than using standard polygonal models, draining away little of the precious processing power.

The marker-less real-time performance-driven aspect of the technology could add a new dimension to video gaming and, since web cams are becoming more common, this would be without the need for non-standard PC equipment. Performance-driven animation is already much better accepted in the video games production industry than in movies, with Famous3D, for example, boasting many of the major video games producers as users of its software. This awareness of existing technology may be advantageous, since the benefits of the new technology will be easier to see.

C.4.4 Market Challenges

The largest challenge in bringing this technology to market is again its current two-dimensional nature. Modern video games are almost entirely 3D and are highly dynamic, with fast changes in view that are not well suited to a 2D approach. Moving from the current 3D procedures to a two-dimensional model could seem to be a technological step backwards, rather than forwards.

There are also established competitors in the industry. Famous3D is a powerful market leader and its software is already established in the industry. One of the advantages that this new technology can offer above Famous3D’s capabilities,
is already offered by Digimask. They already offer a product to bring the player into the game and this is already 3D.

C.4.5 Summary
As for the television, film and advertising industry, the performance-driven facial animation technology presented in this thesis is not currently suited for this application in its present form. Further developmental work is needed to offer three-dimensional capabilities, before it can be brought to market.

C.5 - Telecommunications

The final industry to be considered is the telecommunications industry. This is the current target of future development work and the reasons for this will now be discussed.

C.5.1 Introduction
The capability of the technology for transmitting high quality facial animation at low-bandwidth is an important advantage for the telecommunications industry, particularly for mobile communications. As mentioned earlier, the fact that just a small number of numerical weights need to be transmitted per frame massively reduces the amount of data required for high quality animation.

The technology also offers the facility for surrogate animation, the visualisation of a speaker’s face through a virtual identity. This can be advantageous for corporate branding of information delivery services, where one identity is representative of a company. Once a celebrity model is created, for example, that person could then be animated by anyone, without the need for the original celebrity. Since this can also be achieved at real-time, it is conceivable that videophone users could also choose a representative character for themselves to speak through. This could be particularly desirable for operators who wish to maintain their anonymity. As mentioned, Famous3D’s Impersona has already been used for such purposes for text-based Internet messaging. Text and voice-driven technologies only have the capability to animate the lips, thus failing to represent any accompanying facial expressions. The technology presented here can capture the emotional content of the entire face in addition to mouth movements, giving it an important technological advantage. It is able
to capture the information that would otherwise only be available in a face to face meeting.

C.5.2 The Telecommunications Industry

The telecommunications industry is currently in need of new ideas and new applications, particularly for the third generation (3G) mobile network. The auction of 3G licenses raised more than £22.4 billion in the UK alone and, two years later, network operators have little to show for their investments. As noted in the Financial Times:

“All of these companies therefore need mobile data services such as email, picture messaging and even video conferencing – to take off quickly in Europe if they are to establish themselves or restore their battered reputation as growth stocks”

Estimated sizes of the 3G mobile market vary by source, but Ovum, for example, suggest a global rise from an estimated 11.5 million connections in 2003 to 251 million in 2007, with a European rise from 0 million in 2003 to almost 34 million in 2007.

Personalisation has already proved extremely popular in the mobile phone arena, with millions of phone users downloading custom ring-tones. Estimates put an increase in the market of £50 million per year from £2.5 million in 2000.

The telecommunications industry is thus a highly appropriate target for an innovative technology such as this.

C.5.3 MEDUSA

Mobile Environments for Dialogue and User Surrogate Animation (MEDUSA) is a PACCIT (People at the Centre of Communications and Information Technology) grant application. The PACCIT scheme encourages links between academics and industry in the development of technology for commercial application. The partners on the application are UCL, BTExact and Simian Industries (a mobile entertainment provider). The grant application is currently awaiting a decision on its approval in March.
The main objective of the grant is the eventual exploitation of joint IP for surrogate animation in the mobile market. UCL obviously offers IP and technological knowledge in marker-less performance-driven facial animation, whilst BTExact have technological experience of their own in marker-less face tracking, as well as an advantageous position for downstreaming resulting IP through existing lines of business or for licensing the resulting IP to UK and European third parties. Simian Industries are well placed to realise exploitation of resulting IP, and will also be providing IP and software modules. They also have immense market knowledge of mobile entertainment and mobile services, invaluable for the qualification of user scenarios.

Developmental work is proposed for extending the technology to enable multiple-view tracking and animation, as well as investigations into generating and driving entirely three-dimensional models. Further developments towards categorisation of expression will be looked into. The conversion of speech to fit with a surrogate identity will also be investigated.

C.5.4 Commercial Attractiveness

The technology is set for considerable further development if MEDUSA is approved, but its present and future commercial attractiveness can still be considered. It is a software-based technology and the processing power of state of the art mobile phones should be sufficient to ensure its portability. Cameras capable of capturing still images are already present on many new phones and video capable cameras are starting to appear on newer models, so no extra hardware should be necessary, allowing ease of adoption.

There is also the technological edge. No other system is known to be capable of real-time, marker-less and photo-realistic performance-driven facial animation. The intellectual property is protected by patent.

The two-dimensional nature of the current implementation, that posed such a limitation for the television, film and advertising industry and the computer games industry, is not such a problem in telecommunications, where faces are generally seen from a single view. This reduces risk, since the technology is not reliant on its extension to 3D, but offers further opportunities in these other industries given its success.
As discussed earlier, the climate is also currently good for introducing new applications into the mobile communications industry.

C.5.5 Market Challenges

One of the greatest challenges in developing a product for the market is predicting the customer’s needs before 3G has even started to become established. Industrial experience and market knowledge will be highly beneficial, but continuous market research will be required and this consequently plays an important part in MEDUSA. Several important questions need to be answered. The whole concept of user surrogacy brings up issues of trust and the psychological effects of disembodied interaction.

The technology is reliant on the success of 3G in the mobile market. If 3G fails to gain acceptance, there is no platform for mobile application.

The good climate for new applications in the mobile market will also attract developmental work in similar arenas. A necessary challenge will be to remain technologically ahead of the competition. This will not be a trivial task in such a high-tech industry.

C.5.6 Summary

The telecommunications industry is ideally placed to exploit the technology, particularly given the current climate and need for new ideas. MEDUSA offers an opportunity to develop the technology further towards this goal, whilst bringing it closer to a marketable product for the television, film and advertising industry and computer games industry.

C.6 - Conclusions

It has been seen that the performance-driven facial animation technology presented in this thesis cannot currently constitute a product of commercial viability in the television, film, advertising and computer games industries. The post-production industry is engrained with a culture of hand-animation and poses too much of a challenge to justify risking a product offering automation. In addition, until the technology can offer photo-realism in 3D, it will be inferior to current solutions in the eyes of both industries.
The issue of three-dimensionality is not so limiting in the telecommunications industry and the current climate is ideal for the introduction of new applications. The proposed PACCIT grant, MEDUSA seeks to develop the technology into a marketable 3G application in a low-risk environment with the backing of a major industry player and the market knowledge of an experienced mobile entertainment provider. Further developments can only serve to also increase the marketability of the technology in the other industries.

C.7 - References

1 BTExact Technologies, 81 Newgate Street, London, EC1A 7AJ;  
   http://www.btexact.com

2 UCL Ventures, University College London, Brook House 2-16, Torrington Place, London WC1E 7HN;  
   http://www.ucl.com/uclbusiness/uclv/uclvhome/uclvintro.html


4 Famous3D, 18 Macquarie Street, Prahan, Victoria, Australia, 3181;  
   http://www.famous3d.com

5 Vicon Motion Systems, 14 Minns Business Park, West Way, Oxford OX2 0JB, UK; http://www.vicon.com

6 Face2Face, Inc., 2 Kent Place Boulevard, Summit, New Jersey 07929, USA; 
   http://www.f2f-inc.com

7 3D Studio Max, Discreet, Montreal, Quebec, Canada; http://www.discreet.com

8 Maya, Alias|Wavefront, Toronto, Canada; http://www.aliaswavefront.com

9 Lightwave, NewTek, 5131 Beckwith Blvd, San Antonio, Texas 78249, USA;  
   http://www.lightwave3d.com

10 Softimage, 3510 St-Laurent Blvd., Montreal, H2X 2V2, Canada;  
    http://www.softimage.com
11 Anthropics Technology Ltd., Unit 1, 1 Wallpole Court, Ealing Studios, Ealing Green, London, W5 5ED; http://www.anthropics.com

12 Interview with Muir, D., Ogilvy and Mather, March 2002

13 Interview with Child, D., Head of the 3D Department at The Hive, March 2002

14 Digimask Ltd., 179 Lower Richmond Road, Richmond, Surrey, TW9 4LN; http://www.digimask.com

15 Quake III, id Software; http://www.idsoftware.com

16 Financial Times: "Understanding wireless: can wireless deliver?" Financial Times Survey, 16 October 2002


18 BBC News article: "Mobile tone tunes withdrawn";
   http://news.bbc.co.uk/1/hi/technology/1691574.stm

19 PACCIT Research Programme: http://www.paccit.gla.ac.uk

20 Simian Industries Ltd., Level 9, City House, The Overgate Centre, Dundee, Scotland, DD1 1UH; http://simianindustries.com
## Appendix D - Glossary

### D.1 - Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>2D</td>
<td>Two dimensional</td>
</tr>
<tr>
<td>3D</td>
<td>Three dimensional</td>
</tr>
<tr>
<td>AU</td>
<td>Action Unit (one of the 50 constituent muscles/muscle groups that make up FACS)</td>
</tr>
<tr>
<td>CAT</td>
<td>Computerised Axial Tomography</td>
</tr>
<tr>
<td>CSG</td>
<td>Constructive Solid Geometry</td>
</tr>
<tr>
<td>EMG</td>
<td>Electromyography</td>
</tr>
<tr>
<td>FACS</td>
<td>Facial Action Coding System[^1]</td>
</tr>
<tr>
<td>FACS+</td>
<td>Essa and Pentland's modification of FACS[^2]</td>
</tr>
<tr>
<td>ICA</td>
<td>Independent Components Analysis[^3]</td>
</tr>
<tr>
<td>McGM</td>
<td>Multi-channel Gradient Model[^4]</td>
</tr>
<tr>
<td>MPEG-4</td>
<td>Motion Picture Experts Group v4[^5]</td>
</tr>
<tr>
<td>MRI</td>
<td>Magnetic Resonance Imaging</td>
</tr>
<tr>
<td>PCA</td>
<td>Principal Components Analysis</td>
</tr>
</tbody>
</table>
# D.2 - Glossary of Terms

<table>
<thead>
<tr>
<th>Term</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Animatronics</td>
<td>The application of robotics in puppetry</td>
</tr>
<tr>
<td>Avatar</td>
<td>Computer-generated virtual face model</td>
</tr>
<tr>
<td>Correspondence problem</td>
<td>Given an image, the task of locating each pixel’s new location in a subsequent image</td>
</tr>
<tr>
<td>Dot-tracking</td>
<td>The dynamic tracking of markers physically attached to an actor</td>
</tr>
<tr>
<td>Drive</td>
<td>To manipulate an avatar through an actor’s movements</td>
</tr>
<tr>
<td>Eigenfaces</td>
<td>The principal components extracted from a set of images vectorised by their pixel-wise intensities</td>
</tr>
<tr>
<td>Face space</td>
<td>A multi-dimensional space in which axes define facial characteristics or movements</td>
</tr>
<tr>
<td>Flow-field</td>
<td>An array of motion vectors</td>
</tr>
<tr>
<td>Hanning window</td>
<td>A 2D window based on the cosine function between $-\pi/2$ and $\pi/2$, with its maximum centrally positioned (see Figure 2.3)</td>
</tr>
<tr>
<td>Hard-coded</td>
<td>Manually implemented</td>
</tr>
<tr>
<td>Hard-wired</td>
<td>Determined by neurological or physiological mechanisms</td>
</tr>
<tr>
<td>Hollow face illusion</td>
<td>A visual illusion in which a concave facial surface appears to be convex</td>
</tr>
<tr>
<td>Term</td>
<td>Definition</td>
</tr>
<tr>
<td>----------------------------------</td>
<td>---------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Iconic changes</td>
<td>Image changes due to the appearance or occlusion of features</td>
</tr>
<tr>
<td>Inversion effect</td>
<td>The well-documented impairment in processing of faces when viewed upside-down</td>
</tr>
<tr>
<td>Kalman filtering</td>
<td>A computational algorithm that calculates an optimal estimate of a system’s current state based on prediction and measurement</td>
</tr>
<tr>
<td>Lagrange multipliers</td>
<td>A technique for maximisation or minimisation of a function, given a set of constraints</td>
</tr>
<tr>
<td>Laser scanning</td>
<td>The acquisition of 3D surface information using lasers</td>
</tr>
<tr>
<td>Lighting changes</td>
<td>Image changes due to lighting</td>
</tr>
<tr>
<td>Lip-synching</td>
<td>The synchronisation of mouth movements with audio</td>
</tr>
<tr>
<td>Morphing</td>
<td>The smooth transition between two images with a combination of warping and blending</td>
</tr>
<tr>
<td>Optic flow</td>
<td>See 3.1.5</td>
</tr>
<tr>
<td>Performance-driven animation</td>
<td>The process in which an actor’s motion is tracked and then transferred onto a computer-generated model, to imitate movement</td>
</tr>
<tr>
<td>Photogrammetry</td>
<td>The extraction of three dimensional information from calibrated photographs</td>
</tr>
<tr>
<td>Photo-realistic</td>
<td>A sufficient level of realism such that a synthetic image could feasibly be mistaken for a photograph</td>
</tr>
<tr>
<td>Term</td>
<td>Definition</td>
</tr>
<tr>
<td>---------------------------</td>
<td>-------------------------------------------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Pixel</td>
<td>Picture element. The smallest resolvable rectangular area of an image on screen or in memory</td>
</tr>
<tr>
<td>Pixel-wise intensity</td>
<td>A vectorisation resulting from a list of the image pixel grey levels</td>
</tr>
<tr>
<td>Polygonal mesh</td>
<td>A computer-generated surface consisting of polygonal (usually triangular) elements</td>
</tr>
<tr>
<td>Prosopagnosia</td>
<td>A deficit in which all faces appear unfamiliar</td>
</tr>
<tr>
<td>Rendering</td>
<td>The projection of a three dimensional scene onto a 2D image plane</td>
</tr>
<tr>
<td>Shape</td>
<td>The information describing how the features of a facial image differ in their position from a standard</td>
</tr>
<tr>
<td>Shape-free</td>
<td>See Texture</td>
</tr>
<tr>
<td>Snake</td>
<td>A deformable contour model</td>
</tr>
<tr>
<td>Spline</td>
<td>A curve defined by a mathematical equation</td>
</tr>
<tr>
<td>Texture</td>
<td>The image information once a face has been warped onto a standard shape</td>
</tr>
<tr>
<td>Thatcher illusion</td>
<td>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</td>
</tr>
<tr>
<td>Videoconference</td>
<td>A conference linking participants at different locations using telecommunications and video technology</td>
</tr>
</tbody>
</table>
Voxel  Volume element. The smallest resolvable box-shaped part of a three dimensional space

Waldo  The interface used by a puppeteer to remotely animate an animatronic puppet

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

D.3 - Glossary of Mathematical Terms and Symbols

\( w \)  Image width

\( h \)  Image height

\( x_i \)  Training vectors

\( X \)  Matrix constructed with \( x_i \) s as its columns

\( N \)  Dimensionality of vectorisation

\( M \)  Number of training vectors

\( P \)  The number of vectors in a basis set

\( \mu \)  Mean of training vectors

\( \phi_i \)  Mean-centred training vector (\( \phi_i = x_i - \mu \))

\( \Phi \)  Matrix constructed with \( \phi_i \) s as its columns
\( b_i \)  The \( i^{th} \) principal component of \( X \)

\( B \)  Matrix constructed with \( b_i \)s as its columns

\( \Sigma \)  \( \Sigma = \Phi \Phi^T \)

\( \lambda_i \)  The \( i^{th} \) eigenvalue of \( \Sigma \) (the measure \( \lambda_i/(M-1) \) gives the variance accounted for by the \( i^{th} \) principal component of \( X \))

\([U, V]\)  A vector flow field where \( U \) is the matrix of \( x \) components and \( V \) is the matrix of \( y \) components

\([U^i, V^i]\)  The flow field that maps frame \( i \) onto frame \( j \)

\( \oplus \)  Flow field concatenation operator

\( y_i \)  Driving vectors

\( \psi_i \)  Mean-centred driving vector \( \psi_i = y_i - \mu_{\text{drive}} \)

\( \Psi \)  Matrix constructed with \( \psi_i \)s as its columns

\( c_i \)  Coefficient vectors (weights on basis vectors)

\( C \)  Matrix constructed with \( c_i \)s as its columns (coordinates in target face space)

\( z_i \)  Output vectors

\( \Omega \)  Matrix with the mean-centred projections as columns
\[ \Lambda_p \] \[ P \times P \text{ diagonal matrix composed of } \lambda_i s \]

D.4 - References


