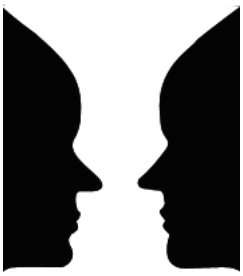
Second-order PCA of dynamic facial expressions for cognitive research

Fintan S. Nagle

March 22, 2011

Supervisors: Alan Johnston, Peter McOwan and Harry Griffin
5,000 words as counted by the TeXcount script at
`http://app.uio.no/ifi/texcount/index.html`.



Acknowledgements
The author would like to thank Harry Griffin for his principal components of cheer and helpfulness throughout the project.

Contents

1 Introduction

The human face is host to one of the most important channels of communication that can exist between two people: the observation and control of facial expressions. During much of our interaction with others our attention is fixed on their faces, either directly or peripherally. We observe features such as gaze direction[1], head orientation[2] and expression[3]; our brains use this information to make conscious or unconscious guesses about the emotional state of another person. In the opposite direction, our own emotional state is reflected in our faces through neural control of the facial muscles, of which there are over 20[4].

Social interaction, along with the facial expressions which underlie communication, are a very important part of social and conflict-based interaction. For this reason, humans have evolved complex recognition and decoding responses, both at the high psychological level (for example, detecting an incoming gaze can make us self-conscious) and the low neuroarchitectural level (for example, the machinery of the fusiform gyrus, an area of the brain specialised for face perception[5]). Of course, networks of neurons implement and form a substrate for the more abstract psychological processes; we must examine both levels in order to build up an accurate picture.

This report describes the investigation of a new facial expression modelling technique, second-order principal component analysis, which is useful both as a video generation tool and as a comparative model for cognitive research.

We begin by summarising existing work on modelling the human facial processing system and its cognitive properties. A description of the techniques introduced by Sirovich *et al* [6] (later extended by the UCL Vision Lab) to model our internal representations of faces is then given. These techniques rely on principal component analysis (PCA), a process which is then re-applied to build a second-order PCA model of dynamic expressions, in this case smiles. The second-order PCA process is discussed and evaluated; finally, potential further work is put forward.

The report is accompanied by a CD containing reconstructed videos produced during the project.

2 Literature Review

This section begins by summarising and comparing major models of human face processing. The technique of PCA is then explained and its applications to expression modelling are shown.

2.1 Existing efforts to model face processing

Explaining the neurological mechanisms underlying face processing is a facet of the most significant problem in psychology and neuroscience: determining how the brain links low-level processing units (neurons) to form larger integrative and problem-solving units. The major obstacle to its resolution is the small physical scale of the brain's neuroarchitecture and the phenomenal complexity thereof (the brain contains on the order of 10^{13} neurons, each with thousands of incoming connections[7]).

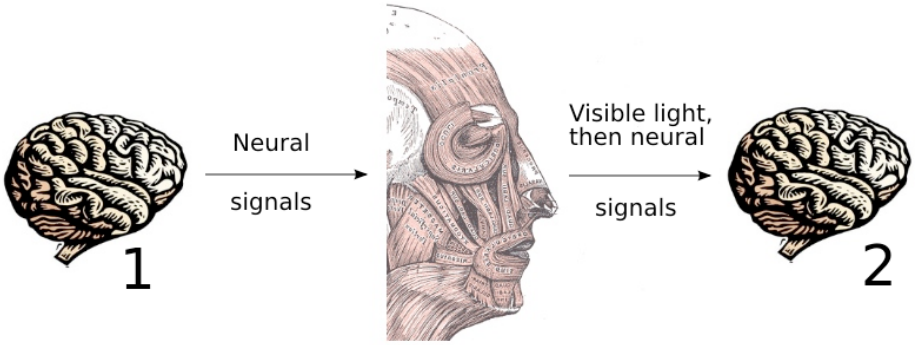


Figure 1: How brain 1 communicates with brain 2 via expressions; the signal is first neural (as brain 1 controls its facial muscles), then visual (as light passes from the face of person 1 to the eyes of person 2), then neural again as it passes from the observer's eyes to their brain. This link is present bidirectionally between two conversing people. (Public domain images.)

Our only method of elucidating brain function, therefore, is to make high-level (either behavioural, experimental, or broadly neural with fMRI scanning) or local (such as electrically monitoring a single neuron) observations and use these to make likely assumptions or definite deductions about the unknown features of brain operation. Until we develop imaging techniques capable of directly observing low-level, broad-scale neural function (and the processing capability required to store and analyse such a huge amount of data), we are restricted to building models of cognition and refining them by comparing their predictions with reality.

One important distinction we can make between models is their purpose.

- *Analogue models* aim to *accurately replicate* (form an *analogue*[1]) of a real system. They aim to encapsulate all our knowledge of a system and provide as much predictive and simulatory power as possible. We often build analogue models when we understand a system well; examples include wind-tunnel airflow simulators, flight simulators, electronic circuit simulation programs, and modern computer physics engines.
- *Summary models* are not required to be similar in any way to the actual mechanisms of a real system. They can simplify the real system or leave out major parts. We often build summary models of systems we do not thoroughly understand; examples include weather prediction, neural network models of cognition, the Lotka-Volterra population equations[8], and early computer physics engines.

The analogue-summary distinction is more of an informal axis than a binary classification; we can refer to models as being "more analogue" or "more summary" than others. The most analogue model possible is a direct copy of the target system itself, with exactly the same behaviours. The more summary a model is, the less faithful it is to the real system, and the fewer the predictions it can make about the target system.

[1]This is the original etymology of the term *analogue computer*.

Although it may at first seem impossible to deduce much about the unknown mechanisms of the brain from high-level behavioural activity, there are many key observations which imply properties of the face processing system, including:

- *Repetition priming.* A familiar face which has been seen recently is responded to faster than one which has not been encountered for a longer time.
- *Double dissociations.* If two neurological features can each be absent or nonfunctional in different impaired individuals (while the other feature is still present), it is highly likely that these features are independent (precisely, that they are implemented by separate neurological systems). For example, some patients can recognise identity but not emotional expression from faces; in some patients, the converse is true.
- *Distinctiveness.* In experiments testing recall and recognition of faces, exemplars which are distinctive are more likely to be accurately recognised. Conversely, typical faces are more likely to be confused with each other or incorrectly identified[9, 10]. We can deduce from this that a face's distinctiveness plays a part in its cognitive representation; distinctive faces are processed differently, in some respect, to non-distinctive faces.
- *Race.* Adults recognise faces belonging to people of other races with lower accuracy than they do own-race faces. However, this effect is absent in children.
- *Inversion.* Adults also find it more difficult to recognise inverted faces, and this effect is virtually absent in children[11]. This could be explained by the same maturing-schema hypothesis proposed to explain race-based recognition differences, and Occam's razor (otherwise known as the principle of parsimony[12]- a valuable deductive tool in cognitive science, as evolved processes often tend towards efficiency and simplicity) decreases the credibility of an explicit foreign-face-recognition interference mechanism.

Overall, these deductions are very vague; they are all expressed in natural language and lack any kind of logical formalism. This, of course, is because we have no such formalism in which to express cognitive concepts; we are forced to use terms like "memory" to express encodings and structures we know next-to-nothing about.

For example, Light *et al* proposed that distinctive faces are easier to recognise as they access "specific memories," whereas unseen, but typical faces are falsely recognised as they activate "schematic memories"[10]. This hypothesis is based around an apparent distinction between two types of memory, specific and schematic. There is much inferred evidence for a schema-based structure of memory[13, 14], but this distinction still lacks much detailed neurological or architectural support.

Hypotheses in cognitive science are often built on top of other theories in this way. Although this is the only way to proceed in deconstructing the brain deductively rather than empirically through actual mapping, we must be careful to appreciate the distance between terms like "memory" and the phenomenally complex implementation which memories doubtless exhibit. Nevertheless, such deductions can still lead to sensible, useful and experimentally confirmable predictions (such as Valentine's guess that distinctive faces should take longer to be classified than typical faces, later confirmed[15, p.163]).

We make use of models in cognitive science by following a meta-process of:

1. Building a model and observing its behaviour.
2. Observing, under similar circumstances, the behaviour of the brain.
3. Studying the differences and similarities between the two behaviours; making deductions therefrom about the differences and similarities between the model's machinery and the brain's machinery.

One of the first modern models of the human face processing system was the Bruce and Young model[16](figure 2). It uses the observation that recognition of identity, expression and speech can each be impaired separately by neurospychological conditions [17] to support the deduction that the brain handles these processes more or less separately.

Bruce and Young go on to posit seven different types of "information codes representing the interpretation and storage of a particular type of information in the brain, although they do not go into any further detail about their precise neurocognitive representation. Leaving this information undefined seems wise, since we are still very far from understanding how the brain stores information on a low level.

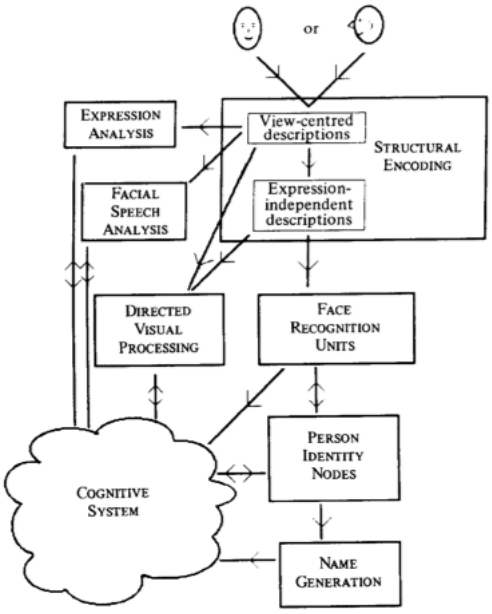


Figure 2: Bruce and Young's view of the face processing system. From [16].

The Bruce and Young model was later extended to the Interactive Action and Competition (IAC) model[18], which formalised the components of its predecessor into a computational model in which units compete to activate other units. This model did well by retaining much of the nondeterminism necessary when modelling such an ill-known system.

Later, Valentine's Multidimensional Face Space model used the observations of distinctiveness, race and inversion noted on page 5 as support for "a framework... in which faces are assumed to be encoded as points in a multidimensional

space."[15, p.165]. Formally, this conceptualisation takes the form of a hyperspace of unspecified dimensionality. "Example axes" such as hair length or hair colour are given, but no actual axes are specified. Each face is represented as a point (with a certain degree of error) in this space, and measurements such as the distance (Euclidean or otherwise) between two points or the distance between a point and the centre of the space (the "norm") can be made. Face recognition (determining whether an exemplar is a face or not) is represented as some kind of "decision process;" face identification (determining whether an examplar is a known face, and if so which) is represented by the distance between the point corresponding to that examplar and either *a)* the norm or average face or *b)* a set of other faces to which the exemplar is compared. This difference separates the *norm-based* and *exemplar-based* versions of the framework.

This framework is both surprisingly formal (including concepts such as points, vectors and space) and surprisingly lacking (leaving concepts such as the distance metric, dimensionality, and details of the decision process or comparison set in the exemplar-based version) unspecified.

Valentine is also unclear about the exact relationship between the face-space framework and the real processes of cognition. He terms it variously "an heuristic framework," "a... metaphor for the mental representation of a face," and a "model." It is unclear whether he intends it to be a summary model or an analogue one.

On the one hand, it is buillt from components which are very unlikely to be formally expressed in the brain. Concepts such as geometry, distance, and specifically placed points are human creations; we have no reason to suppose that the brain's neural architecture relies on such artificial concepts. Furthermore, the face-space model is hierarchical; it is built from concepts which rely upon each other (for example, "similarity" can be decomposed into "distance," which can be decomposed into, for example, Euclidean distance). Other evolved processing systems, such as the immune system[19] or the glucose-insulin system (which performs a complex regulatory task[20]) have been found to operate in an emergent manner; their behaviours simply *happen* in a non-designed way, rather than being back-traceable through a series of hierarchical components, as occurs in a designed system such as an electronic computer. Face recognition (and cognitive processes in general) could occur in a similar way. This makes face space unlikely to be a good analogue model of brain function, as it is based around so many needless high-level concepts.

On the other hand, Valentine describes some unspecified parts of the model, such as the decision process, as "yet to be filled in," which implies that further research will refine it and reveal the ways in which it is closer to reality. Furthermore, he describes auto-associating neural networks as "an implementation of the norm-based [face space] coding model," implying that the brain is another such implementation.

However close to reality Valentine considers it, the face space model has certainly proved "a potential link between many aspects of face recognition research," as Valentine claims. It has provided a clear framework for discussion and research which is still in use today, and is a natural metaphor for one potential way of representing faces. It is also especially well-suited to a major tool of modern facial modelling: principal component analysis, which is discussed shortly.

Given our limited neuroarchitectural knowledge, our current models must be

much more summary than analogue. Their purpose is more result-comparison with empirical data about real cognitive behaviour (in order to make deductions about the systems underlying real cognitive behaviour) than replicating the brain's architecture. The second-order PCA model presented next is one such model; it may only superficially resemble real cognitive processes, but it allows us to make deductions about them.

Details of the brain's machinery will likely be finally revealed by a combination of deductive research (with summary modelling) and investigatory research (using imaging).

2.2 The expression modelling pipeline

This section concisely describes the UCL Vision Research Lab's pipeline for expression deconstruction, modelling and reconstruction from a series of coefficients. Substantially more detail may be found in Appendix A.

The pipeline is based upon the technique of principal component analysis (PCA)[21], a technique which highlights the axes of largest variance in a set of multivariate point data. Each point is defined by d coordinates along the normal axes of the space. PCA imposes b *new axes* on the space (the choice of b is down to the operator, subject to $b < d$) and expresses each point by b coordinates along each of these new axes. Each new axis is chosen so that it spans the maximum possible variance among the points. The technique's usefulness lies in its ability to compress and summarise important data; it can also be used to separate meaningful data (important principal components) from noise (smaller principal components).

The first research on the application of PCA to faces was done by Sirovich *et al* in 1987[6], who evaluated the feasibility of the technique on a set of example faces. Taking each face (a 128×128 pixel greyscale picture) as a point in D-dimensional space, with $D = 128^2 = 2^{14}$, they performed PCA to extract a series of B basis vectors spanning the subspace in which the example faces were situated. They noted that each face can be expressed by adding together a series of "eigenfaces" (points in B-space, each one generated by extending a basis vector by a certain coefficient). To make their method work, Sirovich *et al* had to perform a substantial degree of cropping and normalisation before applying PCA; their test set was also composed exclusively of young Caucasian males.

Work over the intervening years has extended this simplistic application of PCA into a more mature method capable of dealing with a wider variety of faces and expressions in full colour. The Vision Lab uses advanced warping techniques (see Appendix A) to represent an image of a face as a high-dimensional vector containing *a)* serialised bitmap information describing the colour of the face *b)* serialised warp field information describing how to reshape the colour information into the shape of the face. This process is summarised in figure 3.

3 Modelling dynamic expressions

This section describes the design and execution of a feasibility study into performing second-order PCA on video data using the Matlab programming environment.

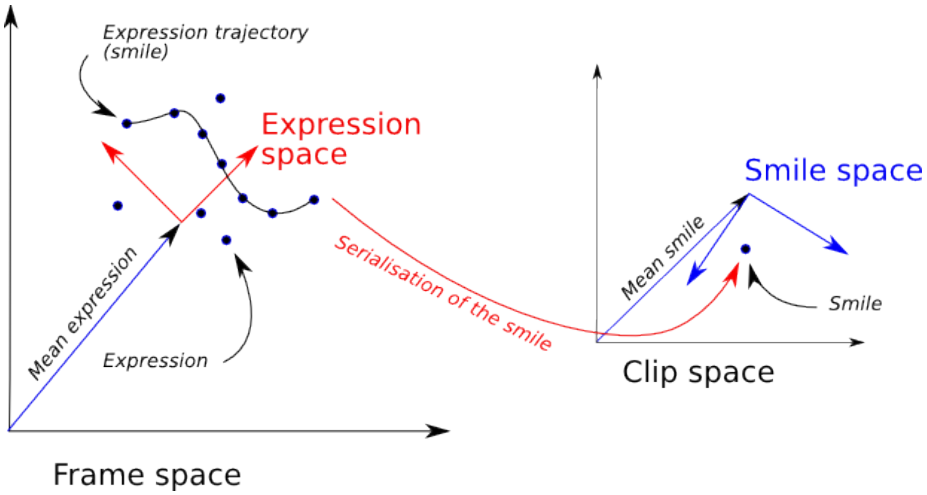


Figure 3: How smiles (sequences of expressions in frame space) are serialised, then represented in clip space and its PCA-obtained subspace, smile space. See Appendix A.

The term "second-order PCA", *per se*, is not very informative. "Second-order" usually implies some kind of extension, repetition or re-application of a procedure (see second-order logic, arithmetic and differential equations). In this case, the actual process of PCA is unaltered. We take "second-order" to mean simply that we apply PCA once to generate some data, and then take these data as input to a second application of PCA. In short,

1. Take a succession of still pictures, generate a description vector for each one (composed of RGB values and {x,y}-warp fields) and generate a PCA space using the methods described earlier.
2. Take a video sequence, separate it into frames, and obtain the PCA axis loadings for each frame.
3. Concatenate the PCA axis loadings for each frame into a description vector for each video sequence.
4. Build a PCA space around these description vectors; each video sequence is now a point in this second-order PCA space (hereafter, PCA^2 space). Different video sequences can be extracted by changing this point's coordinates in PCA^2 space (the loadings on the basis vectors of that space).
5. Take points in the PCA^2 space (either points corresponding to initial video clips, or completely new points) and use the inverse of the above process to generate sequences of points in the first PCA space, then assemble video clips.

PCA^2 can be useful for expression modelling for several reasons. First of all, it is an extension to first-order expression modelling allowing dynamic expressions to be represented; this model could allow interesting and informative behaviour comparisons with real cognitive behaviour. Secondly, it provides us with an easy way of generating artificial video sequences based on certain parameters; such sequences are often used in psychological experiments, and a

consistent, computerised and parametrisable generation method is desirable. Finally, PCA^2 has many envisageable non-scientific but nonetheless very useful applications in such areas as communication, video compression, security, animation, and entertainment. The technique's main purpose, however, is to aid the furthering of our understanding of face processing and other cognitive processes.

Figure 3.2 shows some parts of the pipeline.

The aim was to generate PCA^2 points representing video clips of smiles, in order to investigate the effectiveness of PCA^2 in modelling a the different forms of specific expression. The "model" here consists of the first-order PCA space, the PCA^2 space, and the code required to generate these spaces and to rebuild images and video clips from more detatched representations. It is very much a summary model, as it completely ignores all cognitive, neuromuscular and psychological aspects of smiling.

The following convention is used in what follows:

F	f-dimensional *frame space* whose points represent serialised frames. $f = 60000$ throughout.
E	e-dimensional 1$^{\text{st}}$ order PCA space: *expression space*. $e = 75$ throughout.
C	c-dimensional *clip space* whose points represent serialised clips. $f = 11250$ (75 coefficients $\times$ 150 frames) throughout.
S	s-dimensional 2$^{\text{nd}}$ order PCA space: *dynamic expression* or *smile space*.

These terms are appropriate as each point in E captures an expression (static frame), while each point in S captures a motion sequence (dynamic expression). As the only dynamic expressions studied in this project were smiles, "smile space" seems appropriate.

Note that F contains all possible images. C, on the other hand, contains all possible clips which can be built out of frames generated by points in E; its shape is determined by the set of frames used to generate E.

M_E and M_S symbolise the matrices used to change a point's coordinate system: M_E is an f-by-e matrix and M_S a c-by-s matrix.

$$M_E = \overbrace{\begin{bmatrix} \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \end{bmatrix}}^{e} \Big\} f \qquad M_S = \overbrace{\begin{bmatrix} \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \end{bmatrix}}^{s} \Big\} c$$

3.1 Data collection and post-processing

As this project was a feasibility study, it was deemed best to start with video sequences from only one subject. This subject was recruited informally from the author's friends and the standard ethics and data protection procedures were applied, including full briefing and signature of a consent form.

The goal of the data collection process was to obtain good quality video of a diverse range of different smiles. After being informed that they were going to be filmed as part of an experiment into facial expressions, the subjects were asked to respond naturally to a series of jokes told to them by the experimenter. The only other advice they were given was to keep a relatively straight face until the punch line. More details of the experiment can be found in Appendix .

Efforts were made to keep extraneous features (such as lighting, hairstyle and the subject's gross distance from the camera) constant.

To generate an initial first-order PCA space, a selection of frames covering a wide range of the subject's expressivity was required (in order to let the space cover a wide range of possible expressions). A limit of 3000 frames was put on the input to the initial PCA space (due to memory constraints of the hardware used to run the generation code). This corresponded to 20 clips of 3 seconds each at 50 fps.

An alternative approach would have been to subsample at a much lower frame rate, allowing a wider range of expressions to be used to build the PCA space.

3.2 The PCA^2 pipeline

The next step was the choice of algorithm used to build the PCA spaces; two existing implementations were already present in the form of Matlab code written at the Vision Research Lab:

1. A version of PCA employing the expectation-maximisation (EM) algorithm[22] originally formalised by Dempster *et al* [23].
2. A version of PCA using singular value decomposition (SVD)[24].

The SVD version was selected as it had been in use for longer; the EM version still required some debugging and had not been formally tested. Morph vectors corresponding to each of the 3000 source frames were generated; this had to be done by Harry Griffin at the Vision Lab, as the MCGM code was proprietary. This resulted in a 60,000-D vector (RGB and $\{x, y\}$ warp components for each pixel in a 120×100 image) for each frame.

The SVD implementation of PCA was then used to generate a 60-D expression space E encoded by the transformation matrix M_E. This computation took under an hour on 1 dual-core[2] 3.06 GHz iMac with 4GB of RAM. The processing times and post-processing storage times are shown in figure 5.

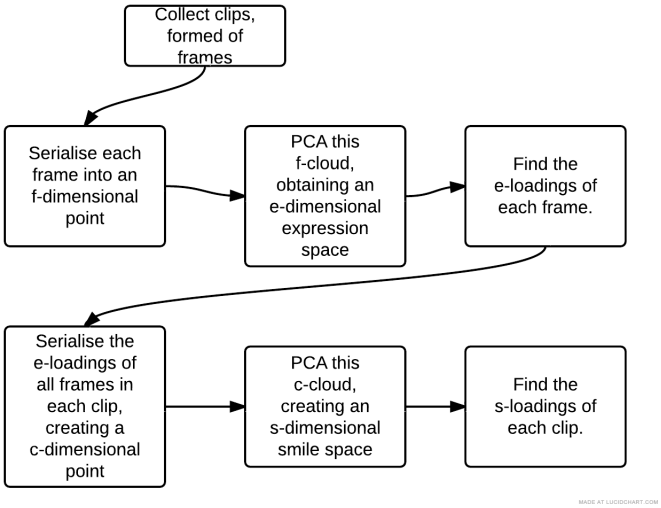
Prewritten Matlab scripts were used to observe the effect of varying the first 12 principal components of E. It was noticed that in one of the initial clips the lighting rig had not been set up correctly, leading to the first and largest principal component mostly capturing this variance in lighting. The first 3 PCs also largely accounted for a variance in the subject's hairstyle, which they had changed twice during the capture process.

It was desirable to have the PCs account for sensible facial expression features rather than lighting and hairstyle changes, so it was decided to split the initial sequence of 20 video clips, V, into 3 sets:

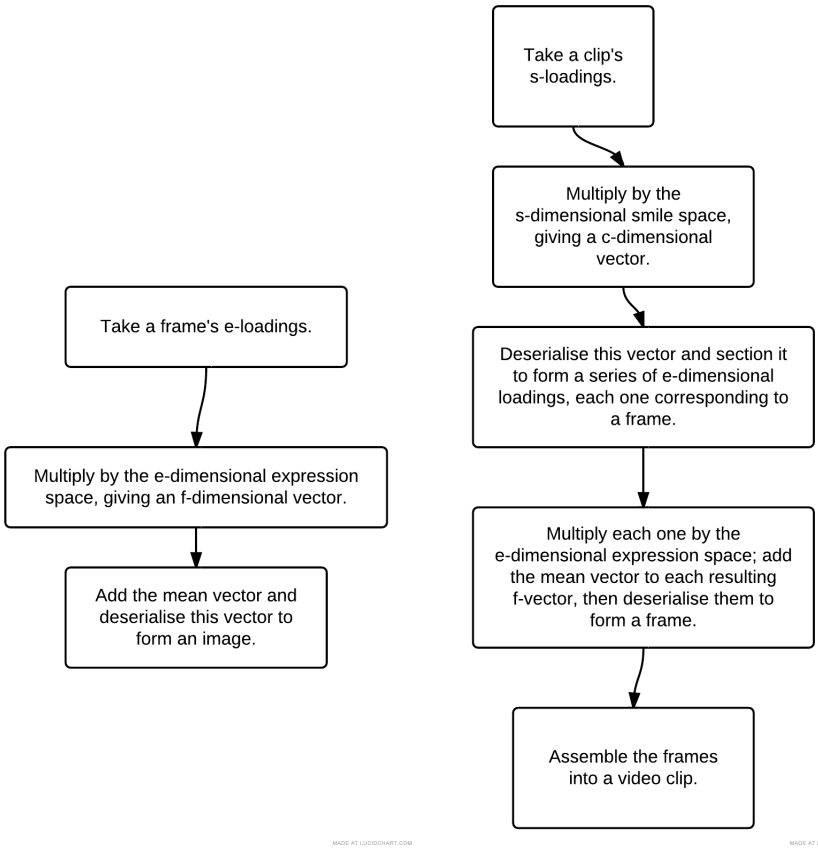
V_1: 14 clips (2100 frames). Constant lighting, 3 different hairstyles.
V_2: 6 clips (900 frames). Constant lighting, constant hairstyle. $V_2 \in V_1$.

Next, Matlab functions (see Appendix B were written to accomplish the following tasks:

[2]Intel Core i3.



(a) Building first- and second-order PCA spaces.



(b) Reconstructing an image from its b-loadings.

(c) Reconstructing a clip from its s-loadings.

Figure 4: Flowcharts explaining parts of the pipeline. *Note:* the expression space and the smile space are not actually represented by e- and s-dimensional vectors; it is the *spaces* which are e- and s-dimensional. Their representations are matrices M_E and M_S as described on page 10.

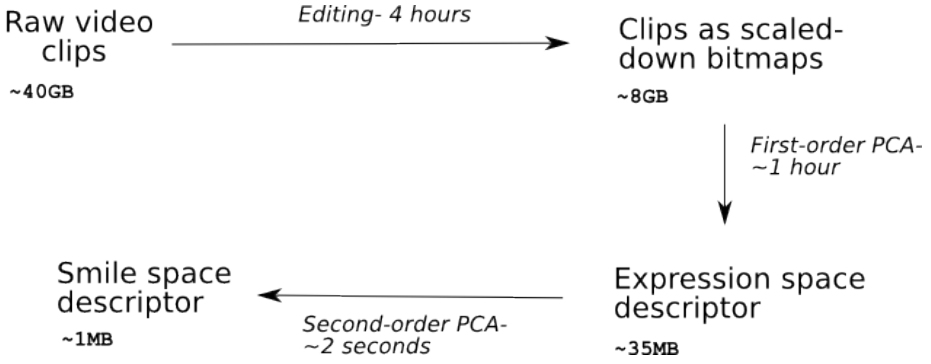


Figure 5: Data storage (not RAM) requirements and processing times for key parts of the pipeline.

- Build the second-order PCA space.
- Reconstruct a clip from an s-vector of loadings in smile space.
- Reconstruct a clip from an s-vector, then produce a video showing *a)* the reconstructed video *b)* the original video *c)* the difference between the two videos, side-by-side. For each pair of original and reconstructed frames f_o and f_r, a "difference frame" was constructed by subtracting the value of each pixel on each channel in f_o from the corresponding pixel value in f_r, taking the absolute value (so that the results were in the interval $[0, 1]$) and building a new frame from these values.
- Generate a random smile video clip from an s-vector with appropriately distributed coefficients. Each element in the random s-vector should have zero mean and the same variance as the corresponding element in the set of s-vectors describing the set of real video clips the PCA space was generated from. This was done by finding the variance of these real-clip s-vectors, then choosing random coefficients from a normal distribution with zero mean and the appropriate variance.

Attempts were made to implement a GUI allowing live control and reconstruction of frames and clips from PC loadings controlled by moving sliders. Unfortunately, the idea had to be abandoned due to the time (0.3s) taken to reconstruct frames.

3.3 Effect of the number of PCs

It was desirable to study the effect of s, the dimensionality of smile space, on the fidelity of reconstructed video clips c_r to original clips c_o. Two error metrics were defined to check consistency:

1. *Absolute error.* For each frame in a clip, an error frame was built as described on page 11. A scalar error was then calculated for each frame f by summing the errors for each pixel p and channel c; finally, a scalar clip error was obtained by summing all the frame errors.

$$\text{error}_{\text{absolute}} = \sum_f \sum_p \sum_c |c_o(f, p, c) - c_r(f, p, c)| \tag{1}$$

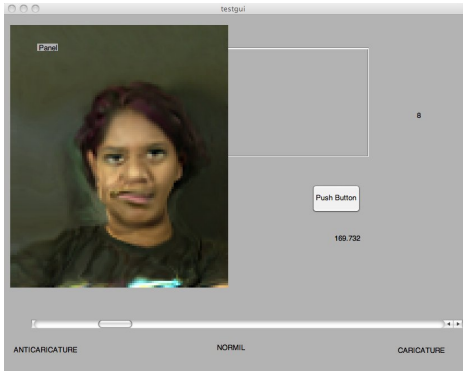


Figure 6: Prototype image reconstruction GUI (showing a slightly caricatured expression); abandoned due to responsiveness issues.

2. *Euclidean error.* Considering each frame as a point in n-D space ($n = h \times w \times 3$)[3] allowed a scalar error to be calculated as the Euclidean distance between two frames. A scalar clip error was constructed by summing the error for each frame.

$$\text{error}_{\text{euclidean}} = \sum_{f} \sqrt{\sum_{p} \sum_{c} (c_o(f,p,c) - c_r(f,p,c))^2} \tag{2}$$

Next, PCA spaces were built with different values of s; video set V_2 was chosen as it had the largest number of clips. Values of s were $\{1, 2, 3, 4, 5, 8, 10, 14\} (s < \#(V_2)$ as the dimensionality of a PCA space must be smaller than the number of observations in the source data set). As shown in figure 7, the error measures were consistent (they are mathematically very similar) and monotonically decreased as s went up. This was to be expected, as more PCs account for more of the original variance and allow better reconstruction.

This effect was confirmed by informally viewing a selection of clips across different values of s. Reconstruction was noticeably better for higher s, especially for detailed facial features like eye movement and exact facial angle (blinking is only reconstructed realistically from around $s = 10$). Low s led to the mouth and nose seeming "disconnected" from the rest of the face.

3.4 Generating artificial smiles

To determine which aspects of expression were captured by each principal component, a selection of videos were generated from s-vectors with all loadings set to zero except one, which was offset from zero by 2 standard deviations of that component in the positive or negative direction. In this case the smile space had dimensionality 7, leading to 14 videos showing the "extremities" of each principal component.

Viewing all the videos side-by-side in sync allowed PCs to be compared. PCs were not intuitively attributable to particular features (such as level of teeth shown), but they did appear significantly opposite (the positive and negative

[3] $h \times w$ pixels and 3 colour channels.

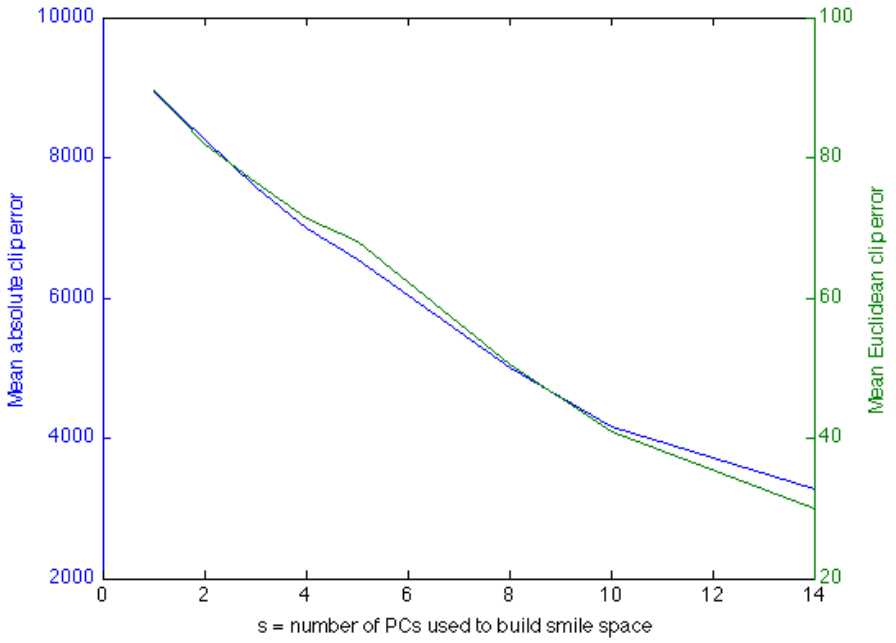


Figure 7: Effect of varying s on the absolute and Euclidean error.

loadings of one PC showed the eyes and mouth opening slightly and extremely, for instance; one of the PCs seemed to capture blinking). The difficulty in picking out the characteristics the PCs captured probably stems from the high similarity of all the smiles captured; using a wider range of emotions might give better results.

It is interesting to imagine a modification of PCA (minus PCA's orthogonality constraint) in which loading axes are *chosen* explicitly to capture particular variations in the target data; for example, we could explicitly search for an axis of mouth dimpling or blink speed.

3.5 Caricaturing and temporal dynamics

Caricaturing a dynamic expression is done by extending or reducing the length of its s-vector while maintaining its direction constant. An interesting question was whether caricaturing would affect the temporal as well as spatial dynamics of the resulting video clip. If so, a dynamic expression's spatial and temporal components would be in a way separable; PCA^2 could then be used to perform experiments investigating whether the same was true in the brain's cognitive representation.

To answer this question, an expression was chosen and several videos generated at various levels of caricaturing, from 0.5 to 2. The videos were compared side-by-side. Spatial dynamics were definitely affected; the amplitude of various facial features such as mouth opening were increased with the level of caricaturing, and the expressions looked convincingly more "intense." However, temporal

dynamics did not seem to be affected; the smiles had the same speed of onset and duration as without caricaturing. Attempting to time-caricature dynamic expressions would be a very interesting task.

A targeted experiment could explore this area further by building a PCA space from similar smiles which varied in speed and duration (perhaps by artificially speeding up and slowing down video before it entered the pipeline).

4 Conclusion

We conclude that second-order PCA (PCA^2) appears a promising technique for modelling facial expression and emotion processing, and conducting research thereon. It is capable of high-fidelity reconstruction of 50 fps video of expressions from as little as 5 real coefficients per clip, plus 30 MB of additional data (far less than would be required to store the original clips; this data is also constant for increasing numbers of clips).

In this case, the implementation of PCA^2 was limited by the single emotion (smiling and laughing with various degrees of intensity) that it was used to capture. The subspaces of frame space F and clip space C spanned by the subspaces generated[4] from this particular expression space E and smile space S therefore captured only smiles, not the full range of possible dynamic expressions. This meant that most points in S generated clips of smiles.

Further work would focus firstly on capturing a wider range of emotions, allowing the corresponding dynamic expression space D to contain more emotions. Deductions could then be made about whether the brain's internal representation of emotion and expression is similar to the PCA^2 pipeline, or not.

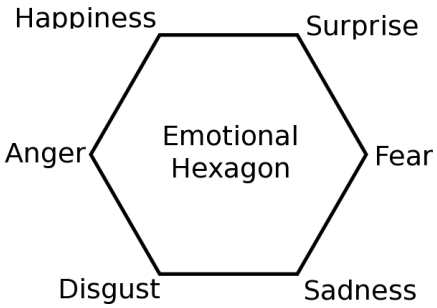


Figure 8: The emotional hexagon.

For example, a common diagram of human emotion is the *emotional hexagon* (see figure 8)[25, 26, 27]. It arranges six "primary" emotions into a hexagon so that adjacent emotions are more "similar" or "confusable", while diametrically opposite emotions are easily differentiable.

A dynamic expression space containing sufficient emotions could be used to generate artificial emotion-clips midway between two of the primaries (or at other selected points in D); these could then be classified by experimental

[4]Multiplying all points in E by M_E generates points forming a subspace of F; respectively for S, M_S and C.

subjects to support or disprove the relationship embedded in the emotional hexagon.

Subjects could also be adapted to emotions and their corresponding antiemotions (points in D and their opposites across the origin). For example, consider possible results in which "anti-sadness" was processed in the same way as "happiness;" this would be evidence that the cognitive representations of happiness and sadness are somehow symmetric and opposite. Another key question is whether such representations are relative or absolute.

Encoding temporal dynamics is a very interesting question, although hard to formalise given our lack of knowledge about cognitive representations of time. At the moment, PCA^2 can only encode expressions of identical length, as their vectors in clip space must be the same length (include the same number of frames). Expression space could conceivably be extended by adding another dimension and storing the *time offset* of a frame as well as its image data; this would allow expressions of differing length to be stored, as long as the frame counts were identical. Exploring the consequences of representing time in the same way as image data (or other, alternative representations of time) would be very interesting.

Experiments with clips generated by PCA^2 would only ever be able to draw high-level conclusions about cognitive function; however, such observations could still afford a great deal of insight into brain function, and at a much lower cost than functional neuroimaging[28] or neural circuit reconstruction[29]. PCA^2 is in a way an extension of the multidimensional, relative coding scheme of Valentine's face space; experiments may confirm that dynamic expressions as well as static expressions are encoded like this, which would increase the likelihood that this type of encoding is widespread across other cognitive processes or the representation of other ideas and concepts.

References

[1] M. Castelhano, M. Wieth, and J. Henderson. I see what you see: Eye movements in real-world scenes are affected by perceived direction of gaze. *Attention in Cognitive Systems. Theories and Systems from an Interdisciplinary Viewpoint*, pages 251–262, 2007.

[2] J.K. Hietanen. Does your gaze direction and head orientation shift my visual attention? *NeuroReport*, 10(16):3443, 1999.

[3] Y.I. Tian, T. Kanade, and J.F. Cohn. Recognizing action units for facial expression analysis. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 23(2):97–115, 2002.

[4] R.F. Gasser. The development of the facial muscles in man. *American Journal of Anatomy*, 120(2):357–375, 1967.

[5] N. Kanwisher, J. McDermott, and M.M. Chun. The fusiform face area: a module in human extrastriate cortex specialized for face perception. *Journal of Neuroscience*, 17(11):4302, 1997.

[6] L. Sirovich and M. Kirby. Low-dimensional procedure for the characterization of human faces. *Journal of the Optical Society of America A*, 4(3):519–524, 1987.

[7] D.W. Patterson. *Artificial neural networks: theory and applications*. Prentice Hall PTR Upper Saddle River, NJ, USA, 1998.

[8] Y. Takeuchi. *Global dynamical properties of Lotka-Volterra systems*. World Scientific Pub Co Inc, 1996.

[9] J.C. Bartlett, S. Hurry, and W. Thorley. Typicality and familiarity of faces. *Memory & Cognition*, 1984.

[10] L.L. Light, F. Kayra-Stuart, and S. Hollander. Recognition memory for typical and unusual faces. *Journal of Experimental Psychology: Human Learning and Memory*, 5(3):212–228, 1979.

[11] A.G. Goldstein. Recognition of inverted photographs of faces by children and adults. *The Journal of genetic psychology*, 127(1st Half):109, 1975.

[12] E. Sober. The principle of parsimony. *The British Journal for the Philosophy of Science*, 32(2):145–156, 1981.

[13] R.N. Tsujimoto, J. Wilde, and D.R. Robertson. Distorted memory for examplars of a social structure: Evidence for schematic memory processes. *Journal of Personality and Social Psychology*, 36(12):1402–1414, 1978.

[14] J.W. Alba and L. Hasher. Is memory schematic? *Psychological Bulletin*, 93(2):203–231, 1983.

[15] T. Valentine. A unified account of the effects of distinctiveness, inversion, and race in face recognition. *The Quarterly Journal of Experimental Psychology Section A*, 43(2):161–204, 1991.

[16] V. Bruce and A. Young. Understanding face recognition. *British journal of psychology*, 1986.

[17] Victoria Bourne Graham Hole. *Face Processing: Psychological, Neuropsychological, and Applied Perspectives*. OUP Oxford, 2010.

[18] A.M. Burton and V. Bruce. Naming faces and naming names: exploring an interactive activation model of person recognition. *Memory*, 1(4):457–480, 1993.

[19] I. Cohen. Immune system computation and the immunological homunculus. *Model Driven Engineering Languages and Systems*, pages 499–512, 2006.

[20] C. Cobelli, G. Nucci, and S. Del Prato. A physiological simulation model of the glucose-insulin system. In *BMES/EMBS Conference, 1999. Proceedings of the First Joint*, volume 2, page 999. IEEE, 2002.

[21] F. Wood, K. Esbensen, and P. Geladi. Principal component analysis. *Chemometr. Intel. Lab. Syst*, 2:37–52, 1987.

[22] S. Roweis. EM algorithms for PCA and SPCA. *Advances in neural information processing systems*, pages 626–632, 1998.

[23] A.P. Dempster, N.M. Laird, D.B. Rubin, et al. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977.

[24] L. De Lathauwer, B. De Moor, J. Vandewalle, and B.S.S. by Higher-Order. Singular Value Decomposition. In *Proc. EUSIPCO-94, Edinburgh, Scotland, UK*, volume 1, pages 175–178, 1994.

[25] R. Sprengelmeyer, AW Young, I. Pundt, A. Sprengelmeyer, AJ Calder, G. Berrios, R. Winkel, W. Vollm "oeller, W. Kuhn, G. Sartory, et al. Disgust implicated in obsessive–compulsive disorder. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 264(1389):1767, 1997.

[26] A.J. Calder, J. Keane, T. Manly, R. Sprengelmeyer, S. Scott, I. Nimmo-Smith, and A.W. Young. Facial expression recognition across the adult life span. *Neuropsychologia*, 41(2):195–202, 2003.

[27] R. Sprengelmeyer, A.W. Young, U. Schroeder, P.G. Grossenbacher, J. Federlein, T. Buttner, and H. Przuntek. Knowing no fear. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 266(1437):2451, 1999.

[28] R. Cabeza and A. Kingstone. *Handbook of functional neuroimaging of cognition*. The MIT Press, 2001.

[29] S.J. Smith. Circuit reconstruction tools today. *Current opinion in neurobiology*, 17(5):601–608, 2007.

[30] PW McOwan and A. Johnston. The algorithms of natural vision: The multi-channel gradient model. In *Genetic Algorithms in Engineering Systems: Innovations and Applications, 1995. GALESIA. First International Conference on (Conf. Publ. No. 414)*, pages 319–324, 1995.

A The UCL Vision Research Lab's expression space pipeline

This section describes how a single still face-image (and, by extension, a video formed of a sequence of images) can be represented as a series of coefficients which (together with a specially defined "face space") can reproduce a given face-image with a high degree of accuracy and a great deal of compression; a face-image can be specified by the coefficients alone (provided the face space is already given), information which is orders of magnitude smaller than the bitmapped version of the image.

Heavy use is made of the technique of *principal component analysis* or *PCA*[21][5], a technique which highlights the axes of largest variance in a set of multivariate point data. Consider an d-dimensional space containing n points. Each point is defined by d coordinates along the normal axes of the space. PCA imposes b *new axes* on the space (the choice of b is down to the operator, subject to $b < d$) and expresses each point by b coordinates along each of these new axes.

Each new axis is chosen so that it spans the maximum possible variance among the points, given the important constraint that the new axes (like the old ones) must be orthonormal. For example, imagine a data set which looks like a vaguely cylindrical point cloud in the original d-space. The axis of greatest variance will be along the longditudinal axis of the point cloud; this, therefore, will be the first axis of b-space. See figure 3 for an illustration. Mathematically, the transformation can be done by finding the eigenvectors and eigenvalues of the covariance matrix of the initial variables.

After PCA each point is described by a b-vector, along with the origins and directions of the b new axes (which are the same for every point and must therefore only be given once for the data set). If $b = d$, PCA merely rotates the data set and every point is still described fully; information is not lost. If $b < d$, some information about the precise location in d-space of each point is lost. The larger b, the smaller the information loss and the smaller the compression. PCA is therefore user-parametrisable in terms of b, allowing an adjustable tradeoff between compression and accuracy.

The first research on the application of PCA to faces was done by Sirovich *et al* in 1987[6], who evaluated the feasibility of the technique on a set of example faces. Taking each face (a 128×128 pixel greyscale picture) as a point in D-dimensional space, with $D = 128^2 = 2^{14}$, they performed PCA to extract a series of B basis vectors spanning the subspace in which the example faces were situated. They noted that each face can be expressed by adding together a series of "eigenfaces" (points in B-space, each one generated by extending a basis vector by a certain coefficient). To make their method work, Sirovich *et al* had to perform a substantial degree of cropping and normalisation before applying PCA; their test set was also composed exclusively of young Caucasian males.

A major problem with feeding a linearised image directly into the PCA procedure, then linearly combining eigenfaces into an output face, is that linear combinations produce inherent blur; combining two widely differing eigenfaces

[5] Also known as the discrete Karhunen-Loève transform, the Hotelling transform or proper orthogonal decomposition.

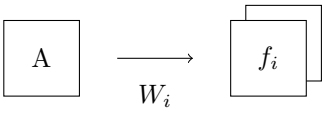
will not produce a realistic-looking face but a blurred combination of the two. This is a reflection of the fact that faces are not pure mathematical artefacts but physiological objects whose shape is constrained by their underlying musculoskeletal structure.

Work over the intervening years has extended this simplistic application of PCA into a more mature method capable of dealing with a wider variety of faces and expressions in full colour. The following procedure is typical of the processing pipeline used by the UCL Vision Research Lab.

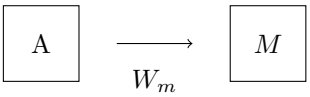
To take into account shape information as well as texture information, all frames are first conceptually shape-standardised using a warping operation[6] The input to the PCA process consists of *a)* RGB or greyscale pixel data specifying a base image, and *b* a vector flow field allowing this base image to be warped into a final output image. This separates texture data from shape data (even though this division is to some degree arbitrary, as discussed later).

The input to the procedure is a sequence of frames f_i and a reference frame A (selected so as to contain colour components allowing realistic warping to a large amount of expressions, such as an open mouth showing teeth and a dark area between the teeth. Warping can remove colours from an image by reducing their area to zero, but not add them).

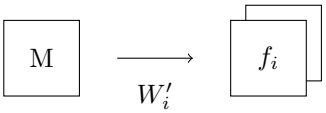
1. Using the multichannel gradient model (MCGM) algorithm[30], a neurally plausible motion-detection and warp-finding algorithm, generate warp fields W_i from each A to each f_i. Each warp field is simply a vector field showing how to deform A to obtain f_i.



2. Average these fields, generating a mean warp W_m.
3. Warp the reference image with the mean warp to generate the warp mean M.



4. Generate new warp fields W_i' which warp the warp mean to each frame f_i (not to the reference image). These warp fields are generated by taking the reverse of the mean warp, then doing the initial warps W_i. These new fields W_i' each symbolise the warp required to reshape the warp mean into the shape of the face in frame f_i.



[6] *Warping* simply spatially deforms an image; *morphing* combines spatial deformation with blurring between a start and finish image.

5. Apply the reverse warps of fields W_i' to each frame. This results in a set of frames f_i' which all have the *shape* of the warp mean, but different texture information.
6. For each frame, store f_i' and W_i'. Each original frame can be reconstructed by applying W_i' to f_i'.
7. Serialise this information (by concatenating the lists r_i, g_i, b_i, x_i, y_i containing red, green, blue and x, y-warp information for each pixel) for each frame. Each frame is now symbolised by a high-dimensional vector containing colour and warp information; these vectors exist in an f-dimensional *frame space*. Frame space, of course, contains all possible images, including those which do not represent faces.
8. Perform PCA on the resulting point cloud. The space spanned by the e principal components chosen is an e-dimensional *expression space*; each point therein represents an expression.
9. Reconstruct expressions by taking a point in e-space (either representing a real expression, or an artificial one), projecting it into frame space, warping its image data with its warp data, and displaying it.

This process is summarised in figure 3.

B Source code

The following pages show Matlab source code for the two main functions of the report: building a second-order PCA space and reclaiming a video therefrom. Much functionality is included in external functions.

C Video files

The included CD contains video files showing various comparative reconstructions generated during the experiment.

- One clip reconstructed with different values of s.
- Side-by-side comparisons of original and reconstructed clips, s=7.
- Some random dynamic expressions, s=7.
- Comparison of the clips produced in section 3.3 showing the effect of increasing or decreasing each principal component.

D Jokes used to generate the smiles during filming

- What did the ham say to the doctor?
  I'm cured!
- Why did Bambi start smoking?
  Deer pressure!
- Why did Little Jack Horner sit in the corner?
  Because his bum was square.
- Why was the electron depressed?
  It felt a bit negative.
- A horse walks into a bar. "Why the long face?" the barman says. "My wife has left me and I'm an alcoholic," says the horse.
- A hole has been found in a naturist camp wall. The police are looking into it.
- What do you call a man with seagulls nesting in his head?
  Cliff!
- What do you call a man with rabbits burrowing into his head?
  Warren.
- Some trees were stolen from my local forest. The police are stumped.
- Later the same day, all the toilet seats were stolen from my local police station. The police have nothing to go on.
- On the news today... police were called to my local nursery school where a two-year-old was resisting a rest.
- When Prince William joined the army he disliked the use of the phrase "Fire at Will!"
- We have a saying where I come from. Time flies like an arrow. Fruit flies like a banana.
- What's a prisoner's favourite punctuation mark?
  The full stop. It marks the end of his sentence.
- Did you hear about the soldier who survived attacks of mustard gas and pepper spray? He was a seasoned veteran.
- When is a door not a door?
  When it's a jar.
- When is a car not a car?
  When it turns into a side street.
- Did you hear about the short fortune-teller who escaped from prison? He was a small medium at large.
- The Energizer bunny was arrested for being violent. He was charged with battery.
- Have you heard about the new practice of coffee-stealing? They call it "mugging."
- A fire ripped through the campsite. The heat was intense.
- My friend went on holiday to Egypt, but couldn't accept that he was really there. I think he was in denial.
- Why do tennis players rarely marry?
  Because love means nothing to them.