Proceedings of the Institute of Acoustics

PCA IMAGE CODING SCHEMES AND VISUAL SPEECH INTELLIGIBILITY

N.M. Brooke and S.D. Scott

School of Mathematical Sciences, University of Bath, Bath

1. INTRODUCTION

There has recently been growing interest in the speech cues that are conveyed by the visible movements of speakers' faces. These can be exploited in at least two ways. Firstly, they can improve the performance of automatic speech recognizers whose conventional, acoustic inputs have been augmented with inputs that represent images of the speakers' visible facial gestures [e.g. 1-4]. Secondly, it is possible to generate speech syntheses which, in addition to an audio output, can display real-time, animated computer graphics of lifelike facial images that simulate the visible speech articulations [e.g. 5-7]. There are a number of potential applications in each of these areas [8,9]. Making use of the visible articulatory movements of speakers' faces in automatic visual or audio-visual speech processing involves handling time-varying data from sequences of images that must be captured or generated at 25-50 frames per second. This volume of data can only be handled within realistic timescales if some form of image data compression is employed which can encapsulate the essential visual speech cues. Some of the earlier approaches have been described elsewhere [8,10]. The simplest (and most extreme) form of compression is to extract and use the time-varying attributes, such as positions or areas, of a number of representative points or features on the face. The complete outlines of features such as the lip margins may also be tracked as they vary with time during speech production and there are sophisticated techniques for doing this [e.g. 11,12]. The main defects of these methods of compression are as follows. Firstly, some of the facial features, such as shadowing and skin texturing, are not amenable to simple parameterisation. Secondly, features such as the teeth and tongue are difficult to measure because they may be only intermittently or partially visible. Thirdly, it is not possible to identify *a priori* the set of facial features that conveys the important visual speech cues.

There is therefore a strong case for employing a data-driven compression scheme that uses facial images as the source data, but that needs to make no assumptions about their structure. For example, multi-layer perceptrons, or MLPs, can be used to code and map oral images into various output spaces [1,3,4,8]. However, there may be difficulties in using MLPs to code very large sets of oral images captured during speech production [8] and they may offer no significant advantages over coding using principal component analysis [13].

2. PRINCIPAL COMPONENT ANALYSIS FOR IMAGE COMPRESSION

Principal component analysis, or PCA [14], as applied to image compression, can be viewed in the following way. Each *(n* x *m)*-pixel image from a set of monochrome images can be represented by a point in *(n* x *m)*-dimensional space. The co-ordinate along each dimension in this space corresponds to the grey-level of a specific pixel. From the data for a set of training images, PCA can then compute a new set of axes, or principal components, defining a space into which each of the data points can be mapped. The principal components are ordered by the variance in the data accounted for. Thus, for example, the first component accounts for the biggest fraction of the total variance, the second for the biggest fraction of the remaining variance, and so on. Images can therefore be represented by a point in the transformed space whose co-ordinate values, or principal component coefficients, can then be applied to a basis function in order to reconstruct the image. If most of the data variance can be accounted for by a small number, *p*, of principal components, where $p < (n \times m)$, data compression can be obtained by encoding images as sets of *p* principal component coefficients.

PCA is now being widely used for image compression. Early applications included whole-face image coding [15,16], but there are many examples of its use specifically for coding oral images [3,11,17] and its potential virtues have been outlined elsewhere [8]. In more recent applications, PCA image coding has been incorporated into a prototypical, animated computer graphics display of talking faces [9], as well as into audio-visual speech recognizers [18,19].

This paper is concerned with an initial assessment of the usefulness of PCA coding schemes for the representation of oral images captured during speech production. Issues include, for example, the effect of varying a) the number of principal components, or channels, *p*, in the coder, b) the size of the training set and c) the spatial resolution of the images. The ability of PCA coding schemes to handle adequately data from outside the training set is also important for recognition tasks.

3. EXPERIMENTAL ASSESSMENT OF PCA IMAGE COMPRESSION

3.1. The database

A visual speech database was constructed from videorecordings of a single, native English speaker enunciating utterances from the NATO RSG-10 standard digit triple lists 3A, 3B and 3C. Each list consisted of 50 triples and was recorded twice. A simultaneous soundtrack was also recorded for identification purposes only.

3.2. Image capture

The speaker was seated in a chair with a head rest to which guide posts were attached, so that head movements were largely eliminated and the head could be repositioned in a standard orientation. A monochrome videocamera (Sony RSC1110 with rotary shutter) with a 75mm focal length, f/2.8 lens was positioned directly in front of the speaker's face level with the speaker's nostrils, at a distance of 1.2 m. The speaker's face was lit by two, 1 Kwatt quartz halogen floodlights at a distance of about 1 m and set 0.3 m to either side of the line from the camera to the speaker, level with the speaker's face. A microphone (AKG D202E1) was boom-mounted 0.2 m from the speaker's mouth, below the picture area, which was occupied by the lower part of the speaker's face from just above the nostrils to just below the chin. A TV monitor was placed in view of the speaker but below the picture area. Its screen was marked with the speaker's nostril positions so that a fixed head alignment could be maintained during recordings. The speaker's field of view was partially occluded to avoid direct exposure to the lights. The digit triples were announced by a prompter and repeated by the speaker at 5-second intervals. The monochrome recordings were captured on a U-matics video-cassette recorder (Sony VO5630) at 25 frames (50 fields) per second.

3.3. Image processing and PCA

The recorded images of the 200 digit triples from lists 3A and 3B were transferred to an Abekas A66 digital video disc store via a CEL-164 digitiser for analysis. They comprised a total of approximately 15000 fields, all of which were used to train the PCA coding scheme [14], after processing them as follows. A fixed oral area of the images was a) windowed out, b) reduced to one of three spatial resolutions: 32 x 24; 24 x 18; or 16 x 12 pixels, by averaging blocks of pixels, and c) contrast-enhanced by linearly extending the central 5/8 of the intensity range to span the full intensity range of 256 grey-levels. The spatial resolutions were chosen to lie at or above the limit at which significant visual cues may be lost [20]. All recorded image fields of the 100 digit triples from list 3C were processed in the same way as the training images and subsequently used to test the performance of the trained PCA coders on 'unseen' images, which were plotted in the principal component space identified by the training process.

3.4. Perceptual experiments

Two visual word identification experiments were carried out using videorecorded monochrome image sequences of digit triples from the database. Each experiment tested ten graduate students, all native English speakers with normal sight and hearing, who viewed on a TV monitor a randomised set of digit triples at 32 x 24 pixel spatial resolution. In experiment 1, the set comprised 50 triples from each of the following three conditions: a) original videorecorded images; b) images reconstructed from 15-channel PCA encodings of triples from the training set; and c) images reconstructed from 15-channel PCA encodings of triples from the test set. Experiment 2 used randomised stimuli consisting of 25 triples from each of the following six conditions: a) original videorecorded images; b)-d) images reconstructed from, respectively, 5-, 15- and 25-channel PCA encodings of triples from the training set; and e)-f) images reconstructed from, respectively, 5- and 25-channel PCA encodings of triples from the test set. The reconstructions from PCA codes were created and recorded using the Abekas A66 digital video disc store.

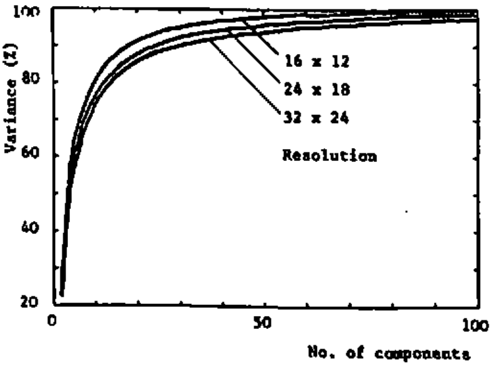

Figure 1: Variance accounted for by PCA as a function of the number of components, or coder channels, at different image resolutions.

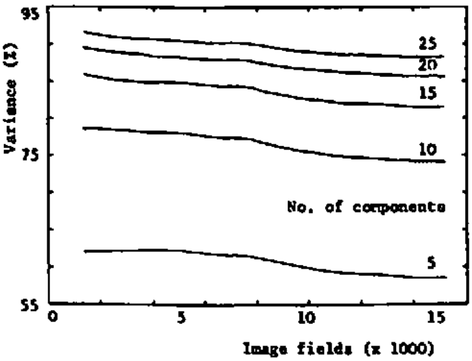

Figure 2: Variance accounted for by PCA as a function of size of the training set, for different numbers of components.

4. RESULTS AND DISCUSSION

Figure 1 shows the percentage variance accounted for as a function of the number of channels in the PCA coder, at the three spatial resolutions studied. The variance axis shows the variance remaining after removing the first component, since this essentially represents only an 'average oral image' for the training set. The variance accounted for increases very rapidly as the number of channels increases up to approximately 15 but much less quickly thereafter. Also, increasing the spatial resolution of the images requires only a small number of additional components to account for a fixed proportion of the data variance. Thus 10, 12 and 14 channels can account for approximately 80% of the variance of the image sets at 16 x 12, 24 x 18 and 32 x 24 pixel resolutions, respectively. This degree of data compression is perhaps not surprising since lip shapes and their movements are anatomically constrained. The oral images and their time variations therefore lie within a relatively small pattern space and are highly structured and correlated. At a image resolution of 32 x 24 pixels, a 15-channel PCA coder can account for about 83% of the variance of the training set and lies near the knee of the curve, above which further gains require many more channels in the coder.

The variance accounted for, given a coder with a fixed number of channels, declines only slightly as the size of the training set increases, as Figure 2 shows. Expressed slightly differently, for example, accounting for 90% of the variance of 32 x 24 pixel images requires a 21-channel PCA coder for a 1200-field set of training images, but only 29 channels for a training set of 12000 fields. The restricted, digit vocabulary of these experiments is such that a small training set is probably representative of the full range of visible gestures; thus, further training data adds little additional information.

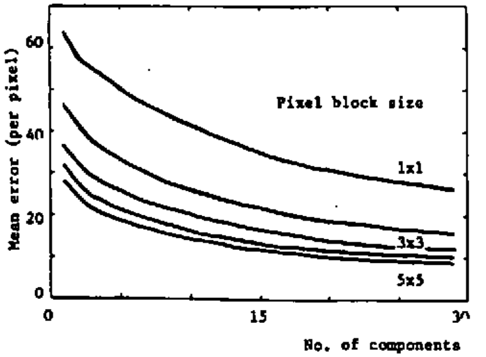

Figure 3: Image reconstruction from PCA codes showing how the error varies with the number of components for blocks of pixels of different sizes (see Section 4 of the text for a full description).

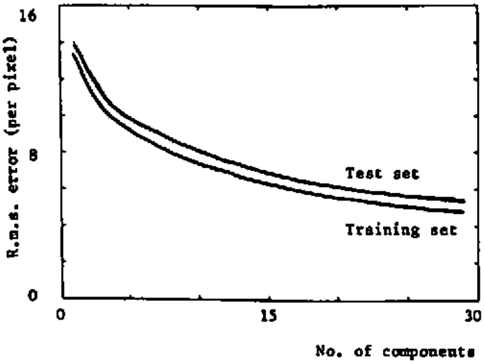

Figure 4: Image reconstruction from PCA codes showing how the root mean square error varies with the number of components for images a) inside, and b) outside, the training set (see Section 4 of the text for a full description).

In order to examine the nature of PCA encoding errors, the average absolute difference per pixel between the grey-levels of corresponding blocks of pixels in a) original images and b) their reconstructions from PCA-encodings, was computed for the worst-case block of each image and averaged over the training set. Square blocks between 1 x 1 and 5 x 5 pixels were examined. The results are shown in Figure 3, as a function of the number of channels in the PCA coder. The coding error decreases rapidly as the block size increases, suggesting that coding errors tend to be scattered throughout the whole image, rather than clustering in small areas. Visual inspection of specimen images confirms this impression; the largest errors tend to appear near high-contrast boundaries, such as tooth outlines, which move. Errors show only small reductions for coders with more than about 15-20 channels. The root mean square difference in grey-level (over all pixels) between the original and reconstructed images from the training set is only marginally smaller than the equivalent measure using images from the test set, as Figure 4 shows. To illustrate the general quality of images reconstructed from PCA codes. Figure 5 shows a specimen frame from a digit triple in the test set, together with its reconstructions from 3-, 6- and 15-channel PCA encodings. There is very little *visible* change in the images reconstructed from the 6- and 15-channel encodings.

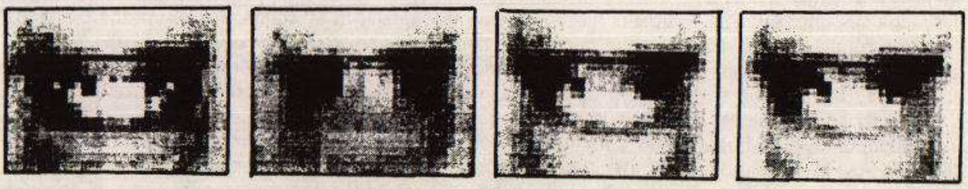

Figure 5: Image reconstruction from PCA encodings. The figure shows, from left to right, an image from the test set and its reconstructions from 3-, 6- and 15-channel PCA coders, respectively.

Objective measures of the type discussed above do not, by themselves, indicate the usefulness of the PCA coding scheme. For both speech synthesis and recognition applications, one appropriate measure of usefulness is the ability of a data compression scheme to retain the essential visual speech cues, so in order to examine this, a perceptual experiment was carried out. It consisted of the two visual word identification tasks described in Section 3.4. The results are shown in Figure 6, in which digit word recognition rates are plotted aginst the number of channels in the PCA encodings from which the images were reconstructed. The dotted horizontal line indicates the mean of the subjects' performances on the original images and forms an upper limit to visual speech recognition ability at the resolution used in the experiments.

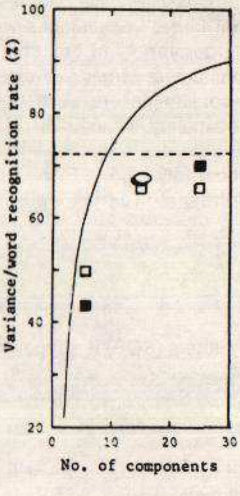

Figure 6: Results of the visual speech perception experiment on digit identification. The word recognition rates are plotted as a function of the number of channels used in the PCA coder from which the image sequences were reconstructed. The details are described in Section 3.4 of the text; ellipses are the results from experiment 1 and squares the results from experiment 2. Open symbols show the results using reconstructed images from the training set; filled symbols show the results using reconstructed images from the test set. The horizontal dotted line shows the recognition rate for the original images. The figure also shows the variance accounted for as a function of the number of channels in the PCA coder, at the same image resolution.

The figure also shows the variance accounted for by PCA encodings at the same spatial resolution. The performance of the two groups of subjects tested in the experiments was comparable, being not significantly different on either the original images or the reconstructions from 15-channel PCA encodings. There was no significant difference in performance on images reconstructed from 15- and 25-channel encodings of triples in the test or training sets, though there was a significant difference using the 5-channel encoder. The confusion matrices were also similar in all conditions except the 5-channel encoding conditions, the most frequent confusions being between 'zero' and 'seven' and between 'eight' and 'nine'. The recognition rates confirm the adequacy of the PCA coding scheme for images both within and outside the training set when 15 or 25 channels are used. The word recognition rates improve significantly as the number of channels in the coder increases from 5 to 15, but improve only slightly from about 64-66% to about 65-68% as the number of channels increases from 15 to 25. This is only slightly below the ceiling value of about 72% for the original images. The recognition results therefore broadly mirror the variation in performance of the PCA coding scheme as the number of channels increases, with approximately 15 channels marking a knee on the curve. These results suggest a useful correlation between the subjective and objective measures of performance of PCA image compression.

As indicated in Section 2, many workers are now quoting PCA coding as an appropriate form of data compression for oral images. We are not, however, aware of any published study of this kind on so large a set of images, or one which attempts to assess an encoding scheme in terms of its perceptual adequacy. The results of this study suggest that image data compression using PCA is an efficient and reliable technique. For example, oral images of spoken digit utterances at 32 x 24 pixels resolution can be represented with the essential visual cues largely intact using a 15-channel PCA encoder. The 15-channel PCA encoding has been used for video speech synthesis that creates 32 x 24 pixel images [9]. Different image resolutions may, as this study has suggested, employ a different number of channels. For example, a 6-channel PCA encoding scheme was used to handle 10 x 6 pixel images in a prototypical audio-visual speech recognizer [18]. PCA coding also has the virtue that the code values reflect the continuity of the time variations in the images of talking faces [8]. However, the digit vocabulary which has so far been examined is a small one; it fails to represent a large number of the phonemes of British English, still less a comprehensive range of phonetic contexts. It is probable, for example, that larger vocabularies would need more channels to encode images to a given accuracy, because a larger variety of oral shapes and movements would be encountered. It is also likely, for the same reason, that a larger set of training images would be needed. Nonetheless, the physical and anatomical constraints are such that the full range of mouth shapes and movements, even for an unlimited vocabulary, would still be comparatively small. As this paper has indicated, there are some encouraging indications that scaling up, for example, to large vocabularies or higher image resolutions, may not necessarily lead to PCA coders with either a very much larger number of channels, or a much poorer performance. Current work is examining the implementation of such larger-scale PCA coding schemes.

5. ACKNOWLEDGEMENT

One of the authors (SDS) is supported by a SERC studentship.

6. REFERENCES

[1] C BREGLER, H HILD, S MANKE & A WAIBEL, 'Improving connected letter recognition by lipreading', Proceedings of IEEE ICASSP, 1, 557-560 (1993)

[2] J ROBERT-RIBES, T. LALLOUACHE, P ESCUDIER & J-L SCHWARTZ, 'Integrating auditory and visual representations for audiovisual vowel recognition', Proceedings of Eurospeech '93, Berlin (European Speech Communication Association), 1753-1756 (1993)

[3] T J SEJNOWSKI, B P YUHAS, M H GOLDSTEIN & R E JENKINS, 'Combining visual and acoustic speech signals with a neural network imp'oves intelligibility', in 'Advances in Neural Information Processing Systems', edited by D S Touretzky (Morgan-Kaufman Publishers, San Mateo, California), 2, 232-239 (1990)

[4] D G STORK, G WOLFF & E LEVINE, 'Neural network lipreading system for improved speech recognition', Proceedings of the International Joint Conference on Neural Networks, Baltimore (IEEE), 2, 285-295 (1992)

[5] F LAVAGETTO, D ARZARELLO & M. CARANZANO, 'Lipreadable frame animation driven by speech parameters', Proceedings of the International Symposium on Speech, Image-processing and Neural Networks, Hong Kong (IEEE), 626-629 (1994)

[6] M M COHEN & D W MASSARO, 'Modeling coarticulation in synthetic visual speech', in 'Computer Animation '93', edited by N M Thalmann & D Thalmann (Springer-Verlag, Tokyo), 139-156 (1993)

[7] K WATERS & T M LEVERGOOD, 'DECface: an automatic lip-synchronization algorithm for synthetic faces', Digital Equipment Corporation Cambridge Research Lab. Report CRL 93/4, pp. 25 (1993)

[8] N M BROOKE & M J TOMLINSON, 'Processing facial images to enhance speech communications', to appear in 'Proceedings of the Second Venaco Workshop on the Structure of Multimodal Dialogue (Maratea, 1991)', edited by M M Taylor, F Neel & D G Bouwhuis; available as Bath University Mathematics and Computer Science Technical Report 94-71, pp. 18 (1994)

[9] N M BROOKE & S D SCOTT, 'Animated computer graphics of talking faces based on stochastic models', Proceedings of the International Symposium on Speech, Image-processing and Neural Networks, Hong Kong (IEEE), 73-76 (1994)

[10] N M BROOKE, 'Mouth shapes and speech', in 'Processing Images of Faces', edited by V Bruce & M Burton (Ablex Publishing, Norwood, N.J.), 20-40 (1992)

[11] C BREGLER & S OMOHUNDRO, 'Surface learning with applications to lip reading', International Computer Science Institute Report TR-94-001 (Berkeley, California), pp. 8 (1994)

[12] K WATERS & D TERZOPOULOS, 'The computer synthesis of expressive faces', Philosophical Transactions of the Royal Society of London (B), 335, 87-93 (1992)

[13] D ANTHONY, E HINES, J BARHAM & D TAYLOR, 'A comparison of image compression by neural networks and principal component analysis', Proceedings of the International Joint Conference on Neural Networks, San Diego (IEEE), 3, 339-344 (1990)

[14] B FLURRY, 'Common Principal Components and Related Multivariate Models' (Wiley, New York) (1988)

[15] M TURK & A PENTLAND, 'Eigenfaces for recognition', Journal of Cognitive Neuroscience, 3, 71-86 (1991)

[16] A PENTLAND, T STARNER, N ETCOFF, A MASOIU, O OLIYIDE & M TURK, 'Experiments with eigenfaces', M I T Media Lab. Perceptual Computing Technical Note 194, pp. 6 (1992)

[17] W J WELSH & D SHAH, 'Facial feature image coding using principal components', Electronics Letters, 28, 2066-2067 (1992)

[18] N M BROOKE, M J TOMLINSON & R K MOORE, 'Automatic speech recognition that includes visual speech cues', Proceedings of the Institute of Acoustics (Autumn Meeting, Windermere), this volume

[19] C BREGLER & Y KONIG, 'Eigenlips for robust speech recognition', International Computer Science Institute Report TR-94-002 (Berkeley, California), pp. 4 (1994)

[20] N M BROOKE & P D TEMPLETON, 'Visual speech intelligibility of digitally processed facial images', Proceedings of the Institute of Acoustics (Autumn Meeting, Windermere), 12(10), 483-490 (1990)