

# Towards a low bandwidth talking face using appearance models.

Barry Theobald, Gavin Cawley, Silko Kruse and J. Andrew Bangham  
School of Information Systems, University of East Anglia, Norwich, UK.  
Email: {b.theobald@uea.ac.uk, {gcc, smk, jab}@sys.uea.ac.uk}

## Abstract

The paper is motivated by the need to develop low bandwidth virtual humans capable of delivering audio-visual speech and sign language at a quality comparable to high bandwidth video. The number of bits required for animating a virtual human is significantly reduced by using an appearance model combined with parameter compression. A new perceptual method is introduced and used to evaluate the quality of the synthesised sequences. It appears that  $3.6 \text{ kbits.s}^{-1}$  can still yield acceptable quality.

## 1 Introduction

Many pre-lingually deaf people find closed caption subtitles in broadcast television of less help than might be expected. Sign language is their first acquired language and subsequently they have difficulties learning to read and write using the conventions of an oral language. The difficulty is similar to that experienced by hearing people when acquiring a second language [14]. Deaf people, therefore, value the presence of an on-screen signer [13] using, in the UK, British Sign Language (BSL). This has been recognised by UK legislation. It requires terrestrial digital television to provide on-screen signing. This paper is motivated by the need to develop virtual humans capable of delivering sign language at a quality comparable to high bandwidth video. An important feature of such an avatar will be the realistic reproduction of facial gestures. They should be clear enough for lipreading for which the face, particularly the tongue is extremely important, although the mouth shapes associated with signing are not those of spoken words. For television broadcast purposes an avatar [28] that can be driven at a bandwidth of less than  $32 \text{ kbits.s}^{-1}$  is desirable.

To broker the trade-off between perceived quality and bandwidth, practical methods for evaluating perceived quality are essential. A new variant of a method for evaluating perceived quality is proposed and illustrated by reporting progress towards a talking face that uses less than  $5 \text{ kbits.s}^{-1}$ .

### 1.1 Background

Research in computer facial animation [23] began in the early seventies with the pioneering work of Parke [22], where a set of parameters that account for the observable variation of the face were defined and used to animate a 3D mesh model. *Physically-based* models of the face have since been employed to animate the face in a more realistic manner. In particular the *Facial Action Coding System* (FACS), where the facial ‘system’ is broken

down into a set of about 60 *action units* that can be combined to reproduce any discernible facial expression, has become the basis of facial animation systems [27] and face trackers or recognisers [11, 26]. More recently, image processing techniques have been applied to modeling and animating the face [12, 25]. Ezzat represented each viseme (discernible mouth shape) with a static image and used an optical flow algorithm [15] to morph between the images to produce animated sequences. This technique has the disadvantage that an accurate flow field cannot be calculated if the distance between images is large. A further problem is that the approach does not make it easy for lip shapes to depend on context: those before and after the current lip shape at any given time. A different approach is to exploit statistical models of shape and appearance that have been used to track [10, 4, 7, 19] and recognise [21, 2, 3] faces in images and video sequences. Bregler combined the tracking of important points on the speaker with morphing of video frames [5] to generate new lip movements in existing video sequences. Rather differently, Brooke [6] used a principal component analysis (PCA) of the intensities of the general mouth region and used the principal modes to drive a statistical model. This can be improved by capturing and taking advantage of important shape changes.

Point distribution models (PDMs) [9] capture what we think (perceive) are the important shapes in images. A tracker is built by linking the shapes to the underlying image through grey level profile distribution models: the Active Shape Model (ASM). However, it appears that ASMs fitted to the inner and outer margins of the lips do not extract sufficient information from video of faces for lipreading [18]. On a database that can be read with a 60% recognition rate by humans and 45% by the best lipreading system, they deliver only about 25%. However, by using the shapes to delineate patches of the image to be included in a PCA (eigen-patch) then a combination of the coded shape and image information included in a third PCA can be used to track or identify faces [8]: the active appearance model (AAM). AAMs appear to be much better than ASMs for the lipreading problem [20, 19] as accuracy can be increased to 42% using, for example, models with 20 modes (of the mouth region). This suggests that, while lip shape is important, at least as much information is obtained from the intensity detail within the inner lip margin. Hence, AAMs should be useful for generating a lipreadable, and fully expressive, talking-head.

## 2 Design and methods

The main step for reducing the dimensionality of video is based on representing facial expressions using the principal components drawn from an appearance model. If 32 bit float values are used for encoding the parameters at a frame rate of 25Hz then  $32 \text{ kb.s}^{-1}$  would only allow 40 parameters to be encoded. Since the whole bandwidth cannot be used for the face alone (the rest of the avatar also has to be animated) the number of bits spent on the face must be as low as possible.

Following the notation of Cootes, a PDM is trained by placing landmarks on a set of images by hand and performing a PCA on the coordinates. Typically we used about 100 points for the whole face. Any training shape can be approximated using  $\mathbf{x} \approx \bar{\mathbf{x}} + \mathbf{P}_s \mathbf{b}_s$ . Where  $\mathbf{P}_s$  is the matrix of the first  $t_s$  eigenvectors, of the covariance matrix, chosen to describe some percentage of the total variation, and  $\mathbf{b}_s$  is a vector of  $t_s$  shape parameters.

An appearance model is computed by warping the training images to place each landmark in each image into the mean position derived from all training shapes. This nor-

malises the shape of each patch and allows the patch to be re-sampled with the same number of pixels in every example. Typically we use 18,000 red, blue, green pixels. By normalising out the contribution of shape to the statistical model of the patch the approach used here [10] differs from that in, for example, [6]. A PCA is performed on the re-sampled pixel values. With such a model any RGB appearance can be approximated using  $\mathbf{a} \approx \bar{\mathbf{a}} + \mathbf{P}_a \mathbf{b}_a$ . Where  $\bar{\mathbf{a}}$  is the mean shape-free image,  $\mathbf{P}_a$  is the matrix of the first  $t_a$  eigenvectors of the covariance matrix and  $\mathbf{b}_a$  a vector of appearance parameters.

Each image is, therefore, described by a set of shape parameters and a set of appearance parameters,  $\mathbf{b}_s$  and  $\mathbf{b}_a$  respectively. A combined model of shape and appearance is computed by concatenating the  $t_s$  shape and  $t_a$  appearance parameters for each image and performing a third PCA (in  $t_s + t_a$  dimensions). Where the number of modes,  $t_s$  and  $t_a$ , are chosen so that typically 95% of the variance of their respective models is captured. The combined shape and appearance model is given by Equation 1,

$$\mathbf{b} \approx \mathbf{Q}\mathbf{c}, \quad (1)$$

where  $\mathbf{Q}$  is the matrix of eigenvectors of the covariance matrix and  $\mathbf{c}$  a vector of parameters that reflect changes in shape and texture of the face. A large range of realistic images can be synthesised given a set of the first  $t$  parameters using Equation 2.

$$\mathbf{x} \approx \bar{\mathbf{x}} + \mathbf{P}_s \mathbf{W}_s \mathbf{Q}_s \mathbf{c}, \quad \mathbf{a} \approx \bar{\mathbf{a}} + \mathbf{P}_a \mathbf{Q}_a \mathbf{c}, \quad (2)$$

$$\mathbf{Q} = \begin{pmatrix} \mathbf{Q}_s \\ \mathbf{Q}_a \end{pmatrix},$$

Where the matrix  $\mathbf{W}_s$  takes into account the scaling mismatch between the weights  $\mathbf{b}_s$  (which model Euclidean distance) and  $\mathbf{b}_a$  (which model pixel RGB intensity). This is computed as shown in [7] and facial animations are generated by controlling the time trajectory of vector  $\mathbf{c}$ , Equation 1.

In practice two steps are particularly time consuming: warping the training images to map the shape-free patches onto their associated means and, during animation, re-warping the synthesised patch into the output shape. Typically it takes  $\approx 2$  s to fit a 6000 pixel AAM to a single image using the Mesa3D software library. However, the process becomes more practical if hardware OpenGL is exploited, e.g. a Diamond Viper 770 takes 0.5 s for the same model.

The second step for refining the information content is based on quantising the  $t$  shape and appearance parameters. As a first approach a simple lossy coder-decoder (codec) has been implemented. The encoder classifies incoming parameters according to their values. Each class is assigned to a specific word length, so that members of each class are represented by a relatively small number of integer values. The accuracy of this representation is controlled by means of the word length, which, in turn, is controlled by the user who can choose between different compression modes in this way. To enable the decoder to differentiate between different word lengths each parameter is accompanied by a leading token in the bit stream indicating the class to which the mode belongs. Currently only two classes are distinguished: members of the range  $[-1.0, 1.0]$  or not. Hence, the token is a single leading bit.

### 3 Perceptual Test using Equivalence

The purpose of perceptual testing is to rank the overall quality of different models. The models are ranked along a single axis: good to bad. Of course, it is easy to rank two scalar values, A and B, but the quality of an animated face model is to be measured in a multi-dimensional space. Different people take, quite literally, different views (projections) of the same sequence and rank the positions of the systems on their good-to-bad ranking axis. Moreover, this view can vary with time with developing theories about how things ‘should’ look. As result, the declared differences between models tend to vary between observers. As far as possible, this is controlled by fixing the viewing conditions and so forth. Standard testing practice used by, for example, the MPEG in assessing the quality of video coding algorithms, is defined in ITU Rec. BT.500 [1] and requires over 20 or so naive viewers to view the test sequences (double blind) and score quality in comparison with a good reference.

Here we try a modification that is modeled on a a method used in text-to-speech synthesis [16]. Rather than implicitly asking viewers to construct their own good-to-bad ranking axis, an axis is provided by degrading a good reference. Viewers are then asked to say when the test and degraded sequences are *equivalently* bad. This changes the viewer from being a ‘measuring instrument’ to that of a ‘detector’.

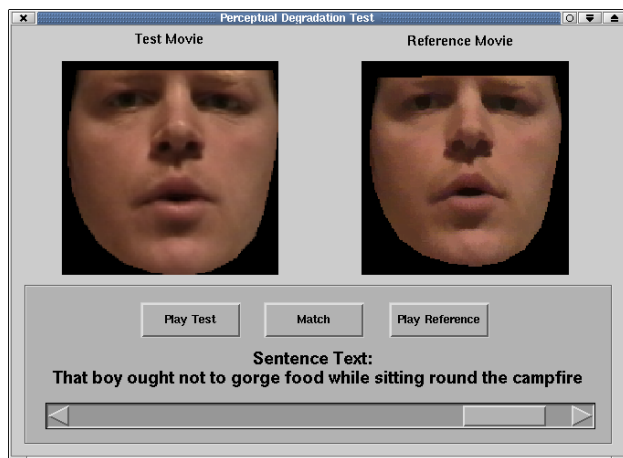


Figure 1: A GUI that plays either a test video (left panel) or a reference (right panel). The reference is a source video that is degraded to an extent controlled by the slider. When the viewer judges the two videos are equally good or bad the ‘match’ button is pressed and another test, chosen at random, appears.

To this end a Matlab GUI is used, see Figure (1). The reference movie on the right side is extracted directly from the original video and the slider allows the user to vary the degree of degradation applied to this movie. The idea is that as the degree of degradation increases monotonically from zero (slider value to the right) to maximum (left) so the quality decreases from good-to-bad. This is illustrated in Figure 2. The actual parameters of the *S* shaped curve are conceptually unimportant (e.g. curve 1 or 2 could be used). What is important is that at one end the reference is worse than, and at the other end

the reference is better than, the models under test. The viewer is then asked to select a position for the slider at which point the sequence under test and the reference video are equally good or bad. Figure 3 shows a frame from a test sequence and three frames from the reference video. The sensitivity of the test system depends on many things. For example, if the tests happen to fall at the positions  $A^*$  and  $B^*$  rather than  $A$  and  $B$  then the error bars in Figure 5 are likely to be large.

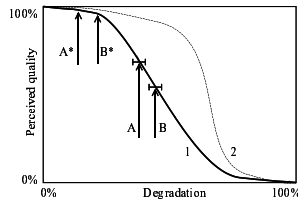


Figure 2: Projecting a high dimensional space onto pre-defined one dimensional good-to-bad axis.

In a typical experiment, 7 test models are compared by showing each 3 times at random: a procedure that takes about 10 minutes. Four different ways of degrading the video were considered: 1) Gaussian blurring, 2) temporal blurring, 3) morphological filtering, and 4) temporal warping, but many others could be devised. The first two are not pursued here as they produced a degradation that was too different from that introduced by the models. The sensitivity of the test is also dependent on the difficulty of the sentence (signing sequence) being communicated. Here we report results from the sentence that appeared the most appropriate challenge to articulate (the others seemed to easily reproduced by an appearance model).

There is plenty of room to improve on this test. For example, it would be appropriate to include the audio signal degraded sufficiently to force viewers to extract information from the synthesised face in order to understand what is being said.

## 4 Results

For this paper models were generated from a database of 9431 images of a single talker uttering the first 100 of a set of sentences constructed to be phonetically rich (BT ‘messiah’ sentences). The facial expression was held as neutral as possible and the pose of the head was roughly maintained throughout so that the main sources of variation were due to the speech. The data was collected in one sitting on a Panasonic DV99B digital camcorder and digitised at a frame rate of 25 frames per second using a Dazzle IEEE 1394 capture card with a frame size of 720x576 (colour). The audio was captured at 44.1 kHz, stereo and was used later to automatically segment the video (HTK matched phonemes to the source text).

The first few modes of the appearance model (Equation 1) capture most of the key variations of the facial features, but to design a talking face requires a judgement on both the number of modes,  $t$ , and how many training examples representative of the faces to be synthesised, are required to obtain a stable estimate of these parameters. Each curve reported in Figure 4 represents the mean of 20 models generated by sampling the image database 20 times at random. It shows how the number of modes required to explain

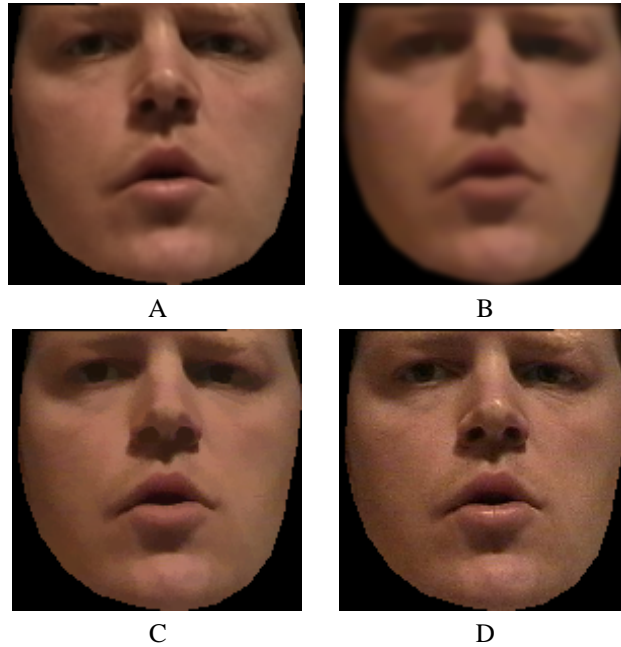


Figure 3: (A) Example of a test face synthesised from the model corresponding to position A in Figure 4. B,C,D show examples of the reference video: (B) maximally degraded, (C) 50% degradation and (D) no degradation. The quality of A lies somewhere between B and D.

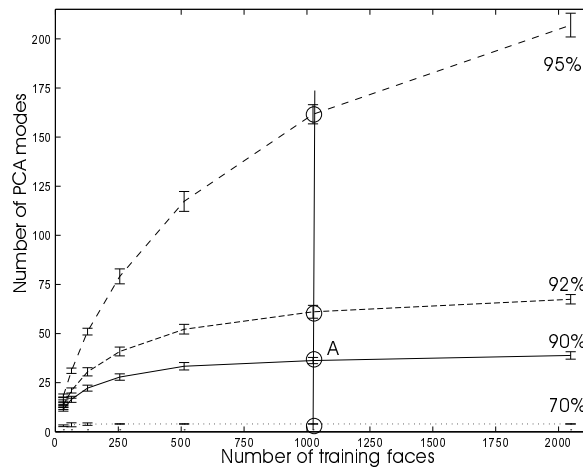


Figure 4: The relationship between number of modes required to capture a given proportion of the variation in the training set (ordinate) as the number of training images increases (abscissa). The error bars show  $\pm 2$  standard deviations.

70%, 90%, 92% and 95% of the total variation increases with the number of training examples. Clearly, 70% of the variation is captured very quickly in just a few modes, whereas to account for 95% of the variation would require in excess of 2000 training examples. In other words, if 1000 images are analysed to account for 90% of the variance (point A Figure 4) almost no extra information is then gained by incorporating another 1000 images. However, this does not necessarily mean that the 31 modes captured at, for example point A, are appropriate for a good reproduction and synthesis. To explore this requires perceptual testing.

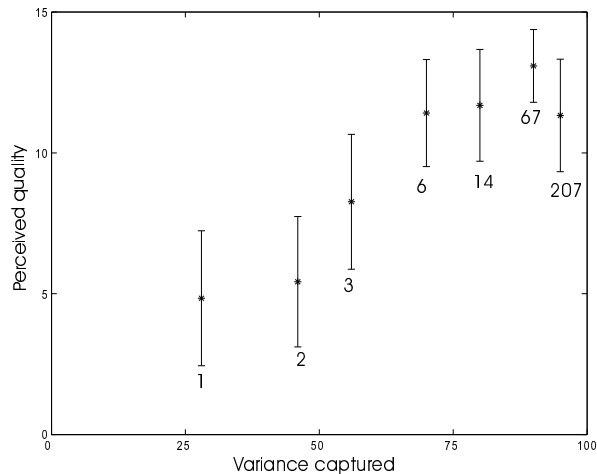


Figure 5: The relationship between perceived quality of a model and the proportion of the total variance captured by combined shape and appearance models. (Perceived quality is measured in arbitrary units drawn directly from the position of the slider shown in Figure 1. 0 is the worst quality and 15 the best: using degradation method 4.) The error bars show  $\pm 2$  times the standard error. The values indicate the number of modes used to synthesise each face.

To find the number of modes required for an adequate reproduction and synthesis, a combined model of shape and appearance was trained on 1000 images randomly sampled from the training set. The model was truncated by taking the first  $t$  modes, where  $t = \{1, 2, 3, 6, 14, 67, 207\}$  (these are points up the vertical line through point A in Figure 4). These models are then used to replay the sentence ‘That boy ought not to gorge food while sitting round the campfire’. (The sentence is chosen because it is particularly challenging.) It appears from the results in Figure 5, obtained from testing with 5 viewers, that perceptual quality does not increase markedly as the number of model parameters increase above about 14.

In order to further reduce the bitrate we apply the quantisation technique (Section 2) and carried out subjective tests for two compression modes. In the first, the parameters are output to a bitstream with only one digit accuracy, which resulted in a wordlength per parameter of 4 to 10 bits depending on the class to which the parameter is assigned. The average compression factor achieved is 3.12 which corresponds to a bitrate of  $3.6 \text{ kb.s}^{-1}$ . In the second method all parameters are mapped to a fixed number of integers so that the wordlength per parameter is between 2 and 4 bits. The average compression factor is 6.7

and the bitrate is  $1.7 \text{ kb}\cdot\text{s}^{-1}$ . In Figure 6 it can be seen that the higher compression factor leads to a more blurred face. This is consistent with the results of a perceptual test using five independent, naive, observers. On our scale of between 0 and 15 they scored a 14 mode model  $14.3 \pm 0.4$  which hardly changed after coding with method one,  $13.6 \pm 0.6$ , but which reduced significantly on using method two,  $11.0 \pm 1.2$ .

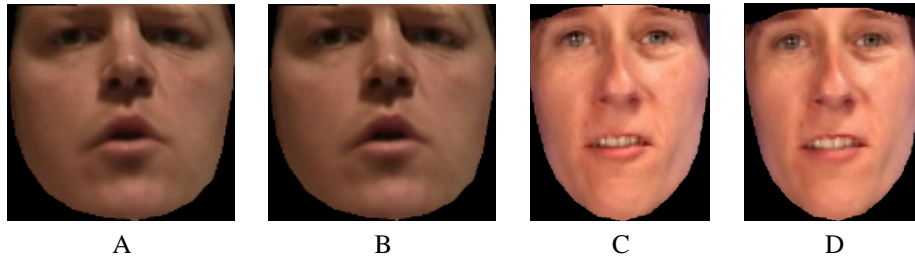


Figure 6: Samples drawn from two image sequences encoded with either  $1.7 \text{ kbits}\cdot\text{s}^{-1}$  (A and C) or  $3.6 \text{ kbits}\cdot\text{s}^{-1}$  (B and D).

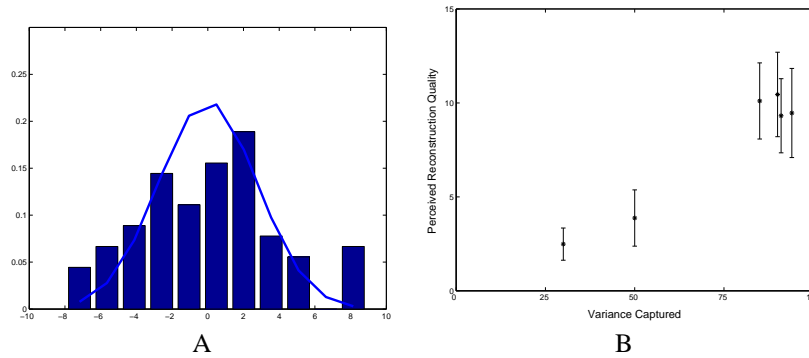


Figure 7: A) The distribution of data deviations about the means for the data shown in Figure 5 with overlaid Gaussian distribution. B) The same distinction between high and low quality models as that seen in Figure 5 but using degradation method 3.

## 5 Discussion

A low bandwidth talking face with fidelity that far exceeds those used in current commercial animations, e.g.[28] is possible using appearance models. The bit-rate reported here for producing a near video-realistic result ( $3.6 \text{ kbits}\cdot\text{s}^{-1}$  or 18 bytes per frame) is very low and should be treated only as a possibility until more extensive testing is complete. Bear in mind that it is not a general video codec, or even a general face codec, rather it has an extremely strong model and we expect this to explain why the low bandwidth can be achieved. To put it into the context of less constrained codecs, a  $64 \text{ kbits}\cdot\text{s}^{-1}$  MPEG-4 codec would handle a  $352 \times 288$  pixel video and a GSM half rate speech coder uses  $9.6 \text{ kbits}\cdot\text{s}^{-1}$ . The results presented here can be compared with research reviewed in [24],

however further work should involve evaluating this system against other model-based systems, such as [17] for example.

The perceptual ‘equivalence’ test procedure looks promising. It does not replace more detailed testing of error rates for viewers reading the information stream but looks to be a relatively quick way for different people to evaluate images and video in a consistent manner. The method has a place as a weekly or monthly check as opposed to a way of making a final evaluation of a system. The method of degrading the reference video does matter. A poor choice leads to larger errors. In these experiments neither spatial nor temporal blurring worked very well. Figure 7B shows the result of evaluating a selection of models using a low-pass morphological spatial filter to degrade each image. Although the characteristic of the curve is not resolved and although the (perceptual) values themselves differ between the two tests, the overall conclusions are the same. Figure 7A shows that the test results are approximately Gaussian distributed about the curve justifying the error bars set to  $\pm 2$  times the standard error.

## 6 Acknowledgements

The authors would like to thank all persons who took part in the perceptual tests and Dr. Matthews (CMU) for his collaboration.

## References

- [1] Methodology for the subjective assessment of the quality of television pictures. Technical report, RECOMMENDATION ITU-R BT.500-10, 1974-1978-1982-1986- 1990-1992-1994-1995-1998-1998-2000.
- [2] H. Abdi, D. Valentin, and B.G. Edelman. Eigenfeatures as intermediate level representations: The case for pca models. In *Brain and Behavioural Sciences*, volume 21, pages 175–177, 1997.
- [3] P.N. Belhumeur, J.P. Hespanha, and D.J. Kriegman. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):711–720, July 1997.
- [4] A. Blake, B. Bascle, M. Isard, and J. MacCormick. Statistical models of visual shape and motion. *Proceedings of the Royal Society of London*, A(356):1283–1302, 1998.
- [5] C. Bregler, M. Covell, and M. Slaney. Video rewrite: driving visual speech with audio. In *Proceedings of SIGGRAPH*, pages 353–360, 1997.
- [6] N.M. Brooke and S.D. Scott. Two- and three-dimensional audio-visual speech synthesis. In *Proc. Auditory-Visual Speech Processing*, pages 213–218, 1998.
- [7] T.F. Cootes, G.J. Edwards, and C.J. Taylor. Active appearance models. In H. Burkhardt and B. Neumann, editors, *Proceedings of the European Conference on Computer Vision*, volume 2, pages 484–498. Springer-Verlag, 1998.
- [8] T.F. Cootes, G.J. Edwards, and C.J. Taylor. Comparing active shape models with active appearance models. In *Proceedings of the British Machine Vision Conference*, pages 173–182, 1999.
- [9] T.F. Cootes, C.J. Taylor, D.H. Cooper, and J. Graham. Training models of shape from sets of examples. In *Proceedings of the British Machine Vision Conference*, pages 9–18, 1992.

- [10] G.J. Edwards, C.J. Taylor, and T.F. Cootes. Learning to identify and track faces in image sequences. In *Proc. British Machine Vision Conference*, 1997.
- [11] I. Essa and A. Pentland. Coding, analysis, interpretation and recognition of facial expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):757–763, July 1997.
- [12] T. Ezzat and T. Poggio. Visual speech synthesis by morphing visemes. Technical Report 1658/CBCL, MIT, 1999.
- [13] Bristol University Centre for Deaf Studies. Deaf people and television. *Research Notes*, 10, 1995.
- [14] S. Gregory, J. Bishop, and L. Sheldon. *Deaf young people and their families: developing understanding*. Cambridge, Univ. Press., 1995.
- [15] B.K.P. Horn and B.G. Schunck. Determining optical flow. *Artificial Intelligence*, 17:185–203, 1981.
- [16] R. D. Johnston. Beyond intelligibility - the performance of text to speech synthesizers. *BT Technical Journal*, 14(1):100–111, 1996.
- [17] I. Koufakis and B. Buxton. Very low bit rate face video compression using linear combination of 2d face views and principal components analysis. *Image and Vision Computing*, 17:1031–1051, 1999.
- [18] Iain Matthews, J. Andrew Bangham, Richard Harvey, and Stephen Cox. A comparison of active shape model and scale decomposition based features for visual speech recognition. In *European Conference on Computer Vision*, pages 514–528, June 1998.
- [19] Iain Matthews, Tim Cootes, Stephen Cox, Richard Harvey, and J. Andrew Bangham. Lipreading from shape, shading and scale. In D. Burnham, J. Robert-Ribes, and E. Vatikiotis-Bateson, editors, *Proc. Auditory-Visual Speech Processing*, pages 73–78, Sydney, Australia, December 1998.
- [20] Iain Matthews, Tim Cootes, Stephen Cox, Richard Harvey, and J. Andrew Bangham. Lipreading from shape, shading and scale. In *Proc. Institute of Acoustics*, 1998.
- [21] A. O’Toole, H. Adbi, K.A. Deffenbacher, and D. Valentin. Low-dimensional representation of faces in higher dimensions of the face space. *Journal of the Optical Society of America*, 10:405–410, 1993.
- [22] F.I. Parke. *A Parametric Model for Human Faces*. PhD thesis, University of Utah, Saltlake City, Utah, 1974.
- [23] F.I. Parke and K. Waters. *Computer Facial Animation*. A K Peters, 1996.
- [24] D.E. Pearson. Developments in model-based video coding. *Proceedings of the IEEE*, 83(6):892–906, June 1995.
- [25] F. Pighin, J. Hecker, D. Lischinski, R. Szeliski, and D. Salesin. Synthesizing realistic facial expressions from photographs. In *Proceedings of SIGGRAPH*, 1998.
- [26] Y. Tian, T. Kanade, and J.F. Cohn. Recognizing action units for facial expression analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(2):97–115, February 2001.
- [27] K. Waters. A muscle model for animating three-dimensional facial expressions. *Computer Graphics (SIGGRAPH ’87)*, 21(4):17–24, 1987.
- [28] M. Wells, F. Pezeshkpour, M. Tutt, J.A. Bangham, and I.A. Marshall. Simon - an innovative approach to deaf signing on television. In *Proceedings of the International Broadcasting Convention*, pages 477–482, 1999.