DiVA – Digitala Vetenskapliga Arkivet <http://umu.diva-portal.org>

This is a pre-print version of a paper published in **International Journal of Image and Graphics**.

Citation for the published paper:

Söderström, Ulrik; Li, Haibo

Representation bound for human facial mimic with the aid of principal component analysis

International Journal of Image and Graphics, 2010, Vol. 10, Issue 3: 343-363

URL: <http://dx.doi.org/10.1142/S0219467810003810>

Access to the published version may require subscription. Published with permission from:

World Scientific Publishing Company.

© [copyright World Scientific Publishing Company] [<http://www.worldscinet.com/ijig/ijig.shtml>]

**Representation bound for human facial mimic with the aid of principal
component analysis**

Ulrik Söderström

*Digital Media Lab, Dept. Applied Physics and Electronics
Umeå University, SE-90187, Umeå, Sweden
ulrik.soderstrom@tfe.umu.se*

Haibo Li

*Digital Media Lab, Dept. Applied Physics and Electronics
Umeå University, SE-90187, Umeå, Sweden
haibo.li@tfe.umu.se*

In this paper we examine how much information is needed to represent the facial mimic, based on Paul Ekman's assumption that the facial mimic can be represented with a few basic emotions. Principal component analysis is used to compact the important facial expressions. Theoretical bounds for facial mimic representation are presented both for using a certain number of principal components and a certain number of bits. When 10 principal components are used to reconstruct color image video at a resolution of 240x176 pixels the representation bound is on average 36.8 dB, measured in peak signal-to-noise ratio. Practical confirmation of the theoretical bounds is demonstrated. Quantization of projection coefficients affects the representation, but a quantization with approximately 7-8 bits is found to match an exact representation, measured in mean square error.

Keywords

Distortion bound; Rate-distortion bound; Facial mimic; Basic emotions; Principal component analysis; PCA.

1. Introduction

The human face is able to exhibit several facial expressions. All of the expressions are often referred to as the facial mimic and the facial expressions. The facial expressions can be represented in several ways, but how compact can this representation be? This question has been addressed within the psychological community. American psychologist Paul Ekman has shown that the visual expressions in a human face can be modeled in a compact way. According to Ekman all facial emotions consist of blended versions of only six basic emotions.^{1,2} The basic emotions are happiness, sadness, surprise, fear, anger and disgust (Fig. 1). By modeling these six emotions it should be possible to model all possible facial emotions. It is however not so straightforward as to say that six images of the basic expressions can be used to model all images of the facial expressions. There is a difference between the psychological representation of expressions and the representation in digital images. **We want to examine how much information is needed to represent Ekman's basic expressions from digital images, i.e. from video sequences.** Providing information about the boundaries for facial mimic representation is the main contribution of this work.

Fig. 1. The six basic emotions.

One way to achieve a compact model of human faces is to apply principal component analysis (PCA) to a series of facial images.³ Kirby and Sirovich were the first to introduce the use of PCA to create a compact model of human faces.^{4,5} They used normalized facial images as input to a PCA and used the compact description

Fig. 2. Neutral facial expressions.

to model the faces. The space that the input faces are distributed in can be seen as very high-dimensional, with one dimension for each pixel in the image. After PCA has been applied, the number of dimensions can be vastly reduced while the representation quality is almost retained. Kirby and Sirovich found that fewer than 100 principal faces are needed to model male Caucasian faces.⁵ This can be considered to be very compact; by using only 100 facial images you can model millions of faces. Kirby and Sirovich's idea of a compact representation for facial images has been used extensively for facial recognition.⁶ Pentland and Turk are pioneers in this area and have performed a substantial amount of research about using PCA for facial recognition.^{7,8} The model used for facial recognition does however only cover facial images in a neutral expression state (Fig. 2). All faces have the same expression so the model contains no information about the facial mimic. Ekman's claim that the facial mimic for one person can be modeled with six basic emotions is not considered for facial recognition of faces with neutral expressions but PCA can also be used to extract principal information from sequences of images. The extracted principal components become a model of facial mimic when the same person appears in all images, if the person has different facial expressions in different frames. Modeling of intra-personal facial expressions is discussed in detail in Section 3. Implementations that make use of the personal face space to encode facial video have been proposed by Torres *et al.* and Crowley *et al.*^{9,10,11,12} Both groups have aimed to encode facial parts of video frames through PCA. We have presented an approach which aims to encode entire video frames of facial video through the use of PCA.^{13,14}

Our goal with this work is to evaluate exactly how compact representation of facial mimics that can be created with the aid of PCA. Previous attempts of using PCA to encode images, both still images and video sequences, show that images can be encoded at very low bitrates with high reconstruction quality. This quality is measured in peak signal-to-noise ratio (PSNR) (Eq. (18)). No one has calculated a bound for the efficiency of this kind of coding. Crowley *et al.* show that they can produce a reconstruction quality of almost 35 dB (PSNR) when they use 15 basis images for reconstruction.¹¹ Torres *et al.* have an average reconstruction quality of 29.6 dB when they use 16 basis images.⁹ We have shown that when we use 10 basis images we have an average reconstruction quality above

31 dB.¹³ There are many results presented, but what is the performance limit?

Our belief is that the facial mimic modeled from a video sequence is in fact the facial mimic of the person in this video sequence. Hence, we want to calculate theoretical bounds for modeling facial mimic through PCA. The task is however not so straightforward. The representation bound is affected by multiple factors, e.g., the spatial resolution, the quality of the original video sequence and how expressive the person's facial mimic is. Therefore, we limit the task to the given circumstances:

- A spatial resolution of 240x176 pixels.
- A color depth of 24 bits for a video with RGB color space, 8 bits per pixel. (3x8)
- RGB color space in the original video.
- Theoretical bounds is calculated individually for each person.
- The video sequences are recorded with hands-free equipment where the global head motion, e.g., rotation and translation, is removed.
- The video is evenly illuminated.

The spatial resolution is equal to high resolution for mobile phones. The standard with the highest quality used today is called QVGA and has a total pixel number of approximately 77000.¹⁵ A resolution of 240x176 pixels is equal to 42420 pixels.

Two different bounds can be calculated; a rate-distortion bound and a distortion bound. Rate-distortion bound refers to the pair of minimum rate and minimum distortion. For a given distortion, the source cannot be modeled correctly by a rate lower than the rate described in the calculated function. The distortion bound describes the minimum distortion that can be achieved for a specific source. We will show how we calculate these bounds and present the resulting bounds for a number of different person's facial mimics.

The distortion that is measured in these bounds is measured in mean square error, where the pixel intensities are compared between the original and reconstructed images. This measurement is calculated on signal level. Lower mean square error does not guarantee that the reconstructed image has a higher visual quality. Subjective evaluation gives a better measurement of the actual visual quality. Since it is subjective it is impossible to calculate a theoretical bound for this kind of measurement. Therefore we focus on objective quality measurement based on mean square error.

The article is organized as follows. Section 2 reviews other low bitrate video coding techniques. In Section 3 the concept of modeling the facial mimic in a high-dimensional space is described. Section 4 and 5 calculates the theoretical bounds of rate and distortion for modeling the facial mimic. Section 6 discusses the difference between rate measured in bits and rate measured in number of principal components. Practical results for reconstruction of facial mimic video sequences are presented in Section 7. Section 8 examines the distribution of the projection

coefficients and the contribution of this paper is discussed in Section 9.

2. Related work: Low bitrate representation of faces

Several different methods and standards have been developed for very low bitrate encoding and representation of facial video sequences. Video coding standards like MPEG-4 and H.264 have been defined.^{16,17} Since both MPEG-4 and H.264 are based on block-wise encoding through Discrete Cosine Transform (DCT) they cannot reach reasonable quality at very low bitrates even if they work very well at higher bitrates. Other techniques than video compression have been proposed to reach very low bitrates. Model-based coding is a technique where a facial model is used with texture mapping to create video-like animations at very low bitrates.¹⁸ Eisert and Girod used this technique and represent faces at very low bitrates with high quality.¹⁹ Facial animation is similar to model based coding and enables visual communication at very low bitrates.²⁰ Facial animation is part of the MPEG-4 standard. An animation of a human face is driven according to parameters extracted from a real video sequence and very high compression is achieved by only sending the parameters. PCA can also be used in facial animation; it reduces the number of parameters needed to represent a face and confines the facial motion to a valid space; thereby preventing unnatural motions in facial animation.²¹ Matching pursuit is another technique that reaches very low bitrates for encoding.²² Matching Pursuit uses features from an alphabet to reconstruct video sequences. Only information about which features that are used in a video frame and what values, e.g. position, size and rotation, these features have has to be transmitted. Active Appearance Model makes use of a statistical model of the shapes and gray-levels of a face.²³ Little information is needed to be transmitted for this technique so the bitrate is very low.

All methods make use of different techniques to enable low bitrate representation of the facial mimic. Block-based techniques make use of block-matching of areas in the face. Changes in facial mimic are encoded through the changes in the movement and intra-block changes of blocks. Model based coding and facial animation uses facial changes to alter a wireframe model according to the changes in the face. Matching pursuit make use of the different features available in the different facial expressions to encode the facial mimic. PCA coding makes use of the principal information in the facial mimic to represent it with only a few principal components. Almost all methods can represent the facial mimic with little information but all techniques except PCA coding have limitations in visual quality and/or framerate. Pighin *et al.* provide a good explanation why high visual quality is important and why video is superior to animations.²⁴ The face simply exhibits so many tiny creases and wrinkles that it is impossible to model with animations or low spatial resolution. Therefore any technique based on animation or texture-mapping to a model is not sufficient. Some approaches have focused on retaining the spatial quality of the video frames at low bitrates at the expense of framerate. Wang and Cohen

presented a solution where high quality images are used for teleconferencing over low bandwidth networks with a framerate of one frame each 2-3 seconds.²⁵ The idea of using low framerate is however not acceptable since both high framerate and high spatial resolution is important for many visual tasks. According to Lee and Eleftheriadis different facial parts have different encoding needs.²⁶ The sense of having eye contact is sensitive to low resolution while lip synchronization is sensitive to low framerate. Therefore it is not sufficient to provide either high framerate or spatial resolution; both are important.

3. Personal mimic space

The idea of a face space has been widely used for facial recognition. An explanation of the face space can be found in.²⁷ Different faces can be distributed in a high-dimensional space and the distribution of the faces is often used to separate them in a recognition system. PCA can be used to reduce the dimensionality while still retaining the difference between the faces that is needed for separation and recognition.

PCA is a linear mapping from one high-dimensional space to another space with the same dimensionality.

$$f(z) : \mathbf{Z}^N \rightarrow \hat{\mathbf{Z}}^N \quad (1)$$

where N is the number of dimensions in the space.

Since the dimensions are highly correlated PCA can compact the data so that fewer dimensions ($M < N$) can be used to represent the source. This is equivalent to a mapping to a space with lower dimensionality:

$$f(z) : \mathbf{Z}^N \rightarrow \hat{\mathbf{Z}}^M \quad (2)$$

The distribution of faces in the new space with lower dimension is then used to separate them.

We want to find the minimum number of dimensions M that is needed to represent the facial mimic. The minimum number is however dependent on the distortion so we have to consider the distortion as well.

Each face has a personal face space that models the distribution of the facial mimic. This space is also high-dimensional and can be reduced in dimensionality through PCA. Ohba *et al.* present a good introduction to the personal face space.^{28,29} We call this space personal mimic space. The separation of the different faces in the personal mimic space corresponds to the separation of the facial expressions. The distributions in the first 3 dimensions for two facial mimic video sequences are shown in Fig. 3. All of the facial mimic will be distributed in the high-dimensional space but it is impossible to visualize more than three dimensions so we only show the first three dimensions. The values on the axes correspond to

8 *Ulrik Söderström and Haibo Li*

Fig. 3. The first three dimension of the personal mimic space for two facial mimics. Every position contained in this space and close to the points are faces.

the values of the projection coefficients α_i (described in Section 4). This is the space that contains the facial mimic; the space that we model in this work.

4. Distortion bound

The result of the representation of the facial mimic is affected by the number of dimensions that are used for the representation but it is also affected by the mean value for each pixel position. $\mathbf{X}_k$ represents frame k where the size of each $\mathbf{X}_k$ is K , the number of pixels in the frame. The mean value for each position is calculated according to Eq. (3) and stored as an image; the mean image, $\mathbf{X}_0$.

$$\mathbf{X}_0 = \frac{1}{K} \sum_{k=1}^K \mathbf{X}_k \quad (3)$$

where K is the number of frames in the video sequence.

The mean image is subtracted from all video frames before they are projected on the principal components. The video frame $\mathbf{X}$ can be modeled without any loss if all linearly independent basis functions ϕ_i and the mean values are used for representation (Eq. (4)).

$$\mathbf{X} = \mathbf{X}_0 + \sum_{i=1}^N \alpha_i \phi_i = \mathbf{X}_0 + \Phi \bar{\alpha} \quad (4)$$

where

$$\Phi = [\phi_1 \dots \phi_N] \quad (5)$$

and

$$\bar{\alpha} = [\alpha_1 \dots \alpha_N]^T \quad (6)$$

The values in the vector $\bar{\alpha}$ are called projection coefficients. They are calculated according to:

$$\alpha_i = \phi_i^T \bar{\mathbf{X}} \quad (7)$$

where the residual $\bar{\mathbf{X}}$ is defined as

$$\bar{\mathbf{X}} = \mathbf{X} - \mathbf{X}_0 \quad (8)$$

The columns of Φ span the N -dimensional space containing $\bar{\mathbf{X}}$ and are called basis vectors. The basis vectors form an orthonormal set so that,

$$\phi_i^T \phi_j = \begin{cases} 1 & \text{for } i = j \\ 0 & \text{for } i \neq j \end{cases} \quad (9)$$

If less features ($M < N$) are used, the signal can still be approximated but with an error. By replacing the components α_i that aren't calculated with preselected constants, β_i , the following approximation can be done:

$$\hat{\mathbf{X}}(M) = \mathbf{X}_0 + \sum_{i=1}^M \alpha_i \phi_i + \sum_{i=M+1}^N \beta_i \phi_i \quad (10)$$

No generality is lost by assuming that only the first M α_i 's are calculated. The resulting representation error is

$$\begin{aligned} \Delta \mathbf{X}(M) &= \mathbf{X} - \hat{\mathbf{X}}(M) \\ &= \mathbf{X} - (\mathbf{X}_0 + \sum_{i=1}^M \alpha_i \phi_i + \sum_{i=M+1}^N \beta_i \phi_i) \\ &= \mathbf{X} - \mathbf{X}_0 - \sum_{i=1}^M \alpha_i \phi_i - \sum_{i=M+1}^N \beta_i \phi_i \\ &= \sum_{i=M+1}^N \alpha_i \phi_i - \sum_{i=M+1}^N \beta_i \phi_i \\ &= \sum_{i=M+1}^N (\alpha_i - \beta_i) \phi_i \end{aligned} \quad (11)$$

The mean square error of $\hat{\mathbf{X}}(M)$ can be used as a measurement of the effectiveness of the subset of M features.

$$\begin{aligned} mse(\hat{\mathbf{X}}(M)) &= \frac{1}{N} E\{\|\Delta \mathbf{X}(M)\|^2\} \\ &= \frac{1}{N} E\left\{ \sum_{i=M+1}^N \sum_{j=M+1}^N (\alpha_i - \beta_i)(\alpha_j - \beta_j) \phi_i^T \phi_j \right\} \\ &= \frac{1}{N} \sum_{i=M+1}^N E\{(\alpha_i - \beta_i)^2\} \end{aligned} \quad (12)$$

For every choice of basis vectors and constant terms, we obtain a value for the mean square error. The optimum choice for β_i is obtained by minimizing $mse(\hat{\mathbf{X}}(M))$ with respect to β_i as follows:

$$\frac{\partial}{\partial \beta_i} E\{(\alpha_i - \beta_i)^2\} = -2[E\{\alpha_i\} - \beta_i] = 0 \quad (13)$$

Solving Eq. (13) for β_i is done according to:

$$\beta_i = E\{\alpha_i\} = \phi_i^T E\{\overline{\mathbf{X}}\} \quad (14)$$

The α_i 's that aren't used should be replaced by their expected values. The mean square error can then be written as

$$\begin{aligned}
 mse(\hat{\mathbf{X}}(M)) &= \frac{1}{N} \sum_{i=M+1}^N E[(\alpha_i - E\{\alpha_i\})^2] \\
 &= \frac{1}{N} \sum_{i=M+1}^N \phi_i^T E[(\bar{\mathbf{X}} - E\{\bar{\mathbf{X}}\})(\bar{\mathbf{X}} - E\{\bar{\mathbf{X}}\})^T] \phi_i \\
 &= \frac{1}{N} \sum_{i=M+1}^N \phi_i^T \xi_X \phi_i
 \end{aligned} \tag{15}$$

where ξ_X is the covariance matrix of $\bar{\mathbf{X}}$. The optimal choice for ϕ_i 's is those who satisfy

$$\xi_X \phi_i = \lambda_i \phi_i \tag{16}$$

that is, the Eigenvectors of ξ_X . After inserting Eq. (16) into Eq. (15) the mean square error becomes

$$mse(opt) = \frac{1}{N} \sum_{i=M+1}^N \lambda_i \tag{17}$$

The proof for Eq. (16) can be found in.³⁰ Fukunaga explains how the mean square error can be calculated for a high-dimensional source when the mean has been subtracted. We have extended this calculation to include the mean image.

A mean square error bound for the number of Eigenvectors ϕ_i that are used for reconstruction is calculated from Eq. (17) by varying the number M of projection coefficients α_i that are used for reconstruction. This is the distortion bound for representing the signal with the selected number of Eigenvectors.

The distortion measurement in Eq. (17) is measured as mean square error. From the mean square error we calculate the corresponding peak signal-to-noise ratio (PSNR) according to Eq. (18). In all figures and tables in this article the PSNR is an average of all video frames in a video sequence.

$$psnr = 20 \log_{10} \left(\frac{255}{\sqrt{mse}} \right) \tag{18}$$

where 255 is the maximum pixel intensity value when a color depth of 24 bits is used (8 bits per color channel).

Curves for two of the facial mimics are displayed in Fig. 4. Table 1 presents the average PSNR for all 10 facial mimic video sequences when different number of Eigenvectors ϕ_i are used for reconstruction.

When the visual results of the video sequences are examined it can be seen that as few as approximately 10 Eigenvectors can model the facial mimic in an efficient way. A detailed description of the video sequences is found in Section 7.1. The facial

Fig. 4. PSNR for two facial mimic video sequences compared to the number of Eigenvectors used for encoding and decoding.

Table 1. Average PSNR for the RD bound of facial mimic video sequences.

Number of Eigenvectors ϕ_i	5	10	15	20	25
PSNR [dB]	34.56	36.82	38.13	39.07	39.75

expressions are all modeled correctly and the visual quality of all video frames is very good. Example frames from one video sequences are shown in Fig. 5.

5. Rate-Distortion bound

In the previous section we calculated the distortion bound for using a restricted number of principal components to represent the facial mimic. The coefficients were then represented with floating point values, i.e., it could be considered to be an exact representation of the coefficients. In any real implementation, the coefficients will be compressed through quantization to enable a more efficient transmission. In this section we examine how optimal quantization of the coefficients can be used to calculate a rate-distortion bound.

In most low bandwidth applications we are limited by the number of bits that we can use for transmission, and not by the quality that can be reached. Often the quality has to be maximized depending on a given bitrate. The limiting factor is the available bits, not the quality. When quantization is used for representation of any source the question is how to assign the bits to model the source in the most efficient way. A theoretical bound for this assignment can be calculated with a Rate-Distortion function. Rate-Distortion functions refer to the minimum rate for representing a source at a given distortion, or the lowest possible distortion at

Fig. 5. Example frames for reconstructed videos using a certain number of Eigenimages. [a] 1 Eigenimage [b] 5 Eigenimages [c] 10 Eigenimages [d] 15 Eigenimages [e] 20 Eigenimages [f] 25 Eigenimages.

a given rate.

Like explained in Section 3 a source is represented through PCA by independent principal components. Such a source can be assumed to consist of several independent Gaussian variables. The principal components are in fact only independent if the distribution of them is Gaussian. For this kind of source the rate distortion function can be calculated through reverse "water-filling" where the rate and distortion is controlled by a variable γ .³¹ The variable γ controls how many principal components are used for representation of the source. An example of reverse "water-filling" when six principal components are used is shown in Fig. 6. Only the principal components that have a variance higher than γ are represented with bits.

The rate distortion function controlled by γ is given by

$$R(D) = \sum_{i=1}^M \frac{1}{2} \log_{10} \frac{\sigma_i^2}{D_i} \quad (19)$$

where

$$D_i = \begin{cases} \gamma, & \text{if } \gamma < \sigma_i^2, \\ \sigma_i^2, & \text{if } \gamma \geq \sigma_i^2, \end{cases} \quad (20)$$

Fig. 6. Reverse water-filling for independent Gaussian principal components. Only the components which have a variance larger than γ is allocated bits in the quantization process.

Fig. 7. Rate-distortion bound for two facial mimics.

where γ is chosen so that $\sum_{i=1}^M D_i = D$

This distortion D is measured through mean square error. According to Eq. (18) the PSNR can be calculated for each facial mimic video sequence.

In Fig. 7 the bounds for two facial mimic video sequences are shown. The mean results for all videos are noted in Table 2 and shown in Fig. 8.

Table 2. Average PSNR for the RD bound of facial mimic video sequences.

Rate [Bits]	1.84	4.22	6.46	8.86	10.90
PSNR [dB]	32.39	34.60	35.79	36.90	37.53

Fig. 8. Mean Rate-distortion bound for all 10 video sequences of facial mimic.

6. Distortion measurement comparison

In Section 4 the rate was measured in number of Eigenvectors used for reconstruction. This distortion calculation was the bound for using a selected number of Eigenvectors for encoding and decoding. One projection coefficient for each Eigenvector was used for reconstruction and the number of bits for representing these coefficients could be considered to be unlimited. In Section 5 the rate was measured in bits for a bound of a certain distortion. This is the bound for modeling the source with as few bits as possible, where the coefficients are subjected to an optimal assignment of bits.

The difference in distortion for using a selected number of Eigenvectors for the two formulas is depending on bit allocation. In Section 4 the bits can be considered to be unlimited and in Section 5 an optimal allocation of bits is used. The difference in distortion is decided by the value γ used for rate and distortion calculation in Eq. (20). The gap between the two distortion calculations is:

$$M \times \gamma \quad (21)$$

where M is the number of Eigenvectors used for reconstruction and γ is the level from Eq. (20). The comparison of the two bounds for four videos are shown in Fig. 9. The bounds for the other videos are similar and therefore not shown.

If the number of bits that are used for quantization of the coefficients is increased the gap between the two bounds is reduced. If almost unlimited number of bits are

Fig. 9. $R(D)$ from Eq. (19) compared to bound from Eq. (17) when 25 Eigenimages are used for reconstruction. – Bound for optimal bit assignment - - Bound for unlimited bit usage

used for quantization, i.e., an exact representation of the coefficients is used, the gap is eliminated. In Fig. 9 the bound from Eq. (17) (*dotted line*) is equal to using 25 Eigenimages so the gap is only eliminated if an exact representation of 25 Eigenimages is used in Eq. (19).

7. Practical experiments

In this section we present practical results and compare them with the theoretical bounds calculated in Section 4. In our previous articles we calculated the practical bounds by comparing the original RGB image with the reconstructed RGB image.^{13,14} Here we compare compressed YUV images with reconstructed compressed YUV images since this is what the Eigenspace models and encodes. An explanation of YUV and RGB frames can be found in the next section.

We also present distortion calculations for quantized projection coefficients. We compare reconstructed video quality when quantized and not quantized projection coefficients are used for reconstruction.

7.1. Practical implementation

We used 10 video sequences for the experiments. Each video sequence showed one person when he/she was displaying the six basic emotions proposed by Ekman.² Example frames are shown in Fig. 1 and 2. After each emotion the subject returned to a neutral state. The video sequences were approximately 30 seconds long, an emotion was displayed for 2-4 seconds and a new emotion was displayed approximately every 5 seconds. The framerate for all the video sequences was 15 fps and the resolution was 240x176 pixels. The color-depth of the original video was 24 bits; 8 bits per pixel.

YUV 4:1:1 compression of the original video was performed for all video sequences. Only every fourth color value is retained and the total size for a video frame is reduced to half the original size, enabling faster computation. Each pixel is stored with 1.5 Bytes instead of the 3 Bytes needed for storage with the RGB color space. Only every fourth value of U and V is stored together with all Y values. The pixel information that is retained is organized into a single frame. In this way we perform PCA for YUV information together. Eigenvectors were extracted through PCA for each video sequence. Each video sequence was then projected onto the Eigenvectors, yielding projection coefficients. The projection coefficients and the Eigenvectors were used to reconstruct the image.

The distortion measurement is done by comparing reconstructed YUV images to original YUV images.

The mean square error for the experimental result is identical (up to four decimals) with the theoretical bound calculated in Section 4. The result is the same regardless of video sequence and the number of Eigenvectors that are used for encoding and decoding.

7.2. Quantization of projection coefficients

Each projection coefficient is represented by a single-precision floating point value, i.e., it is represented by 32 bits. Compression is achieved by quantizing these values since they must be transmitted between an encoder and a decoder. We used both uniform quantization and probability distribution function (pdf) optimized quantization (Lloyd-Max quantization).³² Pdf-optimized quantization quantizes the coefficients according to the pdf of the function so that the quantization step size is smaller where there are more data points. The reconstruction levels and limits of the quantizer follow the distribution of the data to create a more efficient quantization. The drawback is that more information about the quantization limits and reconstruction levels needs to be transmitted.

Fig. 10. PSNR of reconstructed videos when quantized coefficients are used. – *Unquantized coefficients* - - *Uniformly quantized coefficients* --- *Pdf-optimized quantized coefficients*

We only use the most important Eigenvectors which limits the number of coefficients that should be quantized to $M \times K$. M is the number of Eigenvectors and K is the number of frames in a video sequence. For each Eigenvector we can extract 450 projection coefficients (one for each video frame). If 10 Eigenvectors are used the resulting number of coefficients to quantize is 4500. The levels and limits from uniform quantization are chosen as the initial estimate for the pdf-optimized quantizer and the parameters are iteratively calculated.³² The threshold when the iteration ends is set to 10^{-3} .

The quantized coefficients are used to reconstruct the video frames and the mean square error and PSNR for the reconstructed video frames are calculated.

PSNR for two reconstructed video frames corresponding to the number of bits used for quantization is presented in Fig. 10. The results for all the other videos are similar.

For all videos, the representation is equal to an exact representation when 7-8 bits are used for coefficient quantization. When approximately 6 bits are used for quantization there is almost no difference between using uniform and pdf-optimized quantization. Uniform quantization is used throughout this article since it is enough to transfer the starting value and step size to use uniform quantization; two floating-point values. The quality is equal to using pdf-optimized quantization so no improvement is gained from using pdf-optimized quantization. A fixed quantization for all video sequences can be used and then there is no need to transfer any information about quantization levels and limits for each individual video. The two floating-point values needed for uniform quantization is however so little amount of data that it does not validate the use of a fixed quantization. Fixed quantization will probably reduce the quality since the starting value and step size will differ slightly

for Eigenspaces from different persons. This will be addressed by transmitting only two floating-point values.

8. Distribution of projection coefficients

Kolmogorov-Smirnov (K-S) test is a test that compares a given set of data to the standard Gaussian distribution, having mean zero and variance one.³³ K-S test is a hypothesis test and examines if we can reject the hypothesis that the given data can have a Gaussian distribution, at a given significance level.

If Eq. (19) should provide correct rate, the projection coefficients should be Gaussian distributed. With K-S test we examine if that is the case. If the video sequence consists of 450 frames there are 450 principal components. We have previously shown that there is no need to use more than 25 of these to reconstruct the video sequence with satisfactory result.^{13,14} Therefore we only examine the distribution of the projection coefficients along the 25 first Eigenvectors.

To use the K-S test we normalize the projection coefficients so they have mean zero and variance one. We use a significance level of 5 % for the hypothesis test.

Most of the projection along the Eigenvectors is not Gaussian distributed, on average only 10 out of 25 distributions follows a Gaussian distribution. The video sequence with most Gaussian distributed data has 18 Gaussian distributions and the video with fewest Gaussian distributions has only 3.

A Quantile-Quantile (Q-Q) plot is a probability plot that plots the quantiles of two probability distributions against each other.³⁴ If the data sets agree the plot will be on a straight line. When they differ the line will diverge, mostly at the edges of the line. Fig. 11 shows a Q-Q plot of 4 coefficient distributions compared to a Gaussian distribution for one video sequence. This comparison implies that many of the distributions that are not considered as Gaussian in the K-S test are close to a Gaussian distribution.

9. Discussion

Principal component analysis enables facial mimic to be modeled with a few principal components. These principal components can be represented at very low bit-rates. We have presented two theoretical representation bounds for human facial mimic; one rate-distortion bound and one distortion bound. We have also presented practical implementations that confirm the distortion bound. These bounds can be useful in many fields that use facial modeling since they show how much information that is, at least, needed to represent facial mimic.

We have shown that uniform quantization is as efficient as pdf-optimized quantization when approximately 6 bits are used for quantization of the projection coefficients. We have also shown that the distortion bound is achievable when 7 to 8 bits are used for quantization of the projection coefficients.

The Kolmogorov-Smirnov (K-S) tests show that most of the projection coefficients are not Gaussian distributed. There are few coefficients and the result from

(a) Q-Q plot of coefficients along Eigenvector 25 (Gaussian distributed)

(b) Q-Q plot of coefficients along Eigenvector 7 (Gaussian distributed)

(c) Q-Q plot of coefficients along Eigenvector 24 (Not Gaussian distributed)

(d) Q-Q plot of coefficients along Eigenvector 6 (Not Gaussian distributed)

Fig. 11. Q-Q plots of coefficient distributions compared to a Gaussian distribution for one video sequence. + Coefficient values --- Least square fit of the coefficients

the K-S test is affected by this low number. With more coefficients the distribution may follow a Gaussian distribution and the model assumption is correct. The Q-Q plot indicates that non-Gaussian distributions are close to being Gaussian.

A Gaussian distribution is more random than any other kind of distribution. If a mixture of Gaussians, or any other distribution, models the data more accurately it is therefore probable that the rate-bound will be lowered compared to the one that is calculated assuming that the data follows a Gaussian distribution. The bound for a Gaussian distribution is higher than the bound for any other distribution; it may be possible that lower bitrates can be reached.

References

1. P. Ekman and W. Friesen. *Unmasking the face. A guide to recognizing emotions from facial clues*. (Prentice-Hall, New Jersey, 1975).

2. P. Ekman. *Emotion in the Human Face*. (Cambridge University Press, New York, 1982).
3. I. Jolliffe. *Principal Component Analysis*. (Springer-Verlag, New York, 1986).
4. M. Kirby and L. Sirovich, Application of the karhunen-loeve procedure for the characterization of human faces. *IEEE Transactions on pattern Analysis and Machine Intelligence*, **12**(1), (1990), p. 103-108.
5. L. Sirovich and M. Kirby. Low-dimensional procedure for the characterization of human faces. *Journal of the Optical Society of America*, (1987), p. 519-524.
6. W. Zhao, R. Chellappa, A. Rosenfeld, and P. Phillips. Face recognition: a literature survey. *ACM computing surveys*, **35**, (2003), 399-458.
7. M. Turk and A. Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, **3**, (1991), p. 71-86.
8. M. Turk and A. Pentland. Face recognition using eigen faces. In *Proceedings of IEEE Computer Vision and Pattern Recognition*, (Maui, Hawaii 1991), p. 586-591.
9. L. Torres and D. Prado. A proposal for high compression of faces in video sequences using adaptive eigenspaces. In *Proceedings of International Conference on Image Processing (ICIP)*, **1**, (2002), p. I-189-192.
10. L. Torres and D. Prado. High compression of faces in video sequences for multimedia applications. In *Proceedings of IEEE International Conference on Multimedia & Expo (ICME)*, **1**, (Lausanne, Switzerland 2002), p. 481-484.
11. W. E. Vieux, K. Schwerdt, and J. L. Crowley. Face-tracking and coding for video compression. In *Proceedings of International Conference on Computer Vision Systems (ICVS)*, (1999), p. 151-160.
12. K. Schwerdt and J. Crowley. Robust face tracking using color. In *Proceedings of the 4th International Conference on Automatic Face and Gesture Recognition*, (2000), p. 90-95.
13. U. Söderström and H. Li. Full-frame video coding for facial video sequences based on principal component analysis. In *Proceedings of Irish Machine Vision and Image Processing Conference (IMVIP)*, (Belfast, Northern Ireland, 2005), p. 25-32.
14. U. Söderström and H. Li. Very low bitrate full-frame facial video coding based on principal component analysis. In *Signal and Image Processing Conference (SIP)*, (Honolulu, Hawaii, 2005).
15. J. Rasmusson, F. Dahlgren, H. Gustafsson, and T. Nilsson. Multimedia in mobile phones - the ongoing revolution. *Ericsson review*, **02**, (2004).
16. Iso/iec 14496-2:2004 information technology coding of audio-visual objects part 2: Visual.
17. R. Schäfer, T. Wiegand, and H. Schwarz. The emerging h.264 avc standard. *EBU Technical Review*, **293**, (2003).
18. K. Aizawa and T. Huang. Model-based image coding: Advanced video coding techniques for very low bit-rate applications. *Proceedings of the IEEE*, **83**, (1995), p. 259-271.
19. P. Eisert and B. Girod. Analyzing facial expressions for virtual conferencing. *IEEE Computer Graphics and Applications*, **18**, (1998), 70-78.
20. J. Ostermann. Animation of synthetic faces in mpeg-4. In *Proceedings of Computer Animation*, *IEEE Computer Society*, (1998), p. 49-55.
21. K. Goudeaux, T. Chen, S.-W. Wang, and J.-D. Liu. Principal component analysis for facial animation. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP01)*, **3**, (2001), p. 1501-1504.
22. R. Neff and A. Zakhor, Very low bit-rate video coding based on matching pursuits, *IEEE Transactions on Circuits and Systems for Video Technology*, **7**(1), (1997), p.

- 158-171.
23. T. Cootes, G. Edwards and C. Taylor, Active appearance models. In *Proceedings of European Conference on Computer Vision. (ECCV)*, **2**, (1998), p. 484-498.
24. F. Pighin, J. Hecker, D. Lishchinski, R. Szeliski and D. H. Salesin, Synthesizing realistic facial expression from photographs. In *SIGGRAPH Proceedings*, (1998), p. 75-84.
25. J. Wang and M. F. Cohen, Very low frame-rate video streaming for face-to-face teleconference. In *Proceedings of the Data Compression Conference (DCC)*, (2005), p. 309-318.
26. J. Lee and A. Eleftheriadis, Spatio-temporal model-assisted compatible coding for low and very low bitrate video telephony. In *Proceedings of the 3rd IEEE International Conference on Image Processing (ICIP 96)*, (Lausanne, Switzerland, 1996), p. II-429-432.
27. M. A. Burton and J. R. Vokey, The face-space typicality paradox: Understanding the face-space metaphor, *The Quarterly Journal of Experimental Psychology*, **51A(3)**, (1998), p. 475-483.
28. K. Ohba, G. Clary, T. Tsukada, T. Kotoku and K. Tanie, Facial expression communication with fes. In *International conference on Pattern Recognition (ICPR)*, **2**, (1998), p. 1376-1378.
29. K. Ohba, T. Tsukada, T. Kotoku and K. Tanie, Facial expression space for smooth tele-communications. In *Proceedings of the 3rd IEEE International Conference on Automatic Face and Gesture Recognition*, (1998), p. 378-383.
30. K. Fukunaga, *Introduction to statistical pattern recognition (2nd ed.)*, (Academic Press Professional Inc., San Diego, 1990).
31. T. M. Cover and J. A. Thomas, *Elements of Information Theory, Wiley Series in Telecommunications*, (John Wiley & Sons, New York, 1991).
32. K. Sayood, *Introduction to data compression*, (Morgan Kaufmann Publishers Inc., San Francisco, 1996).
33. W. H. Press, S. A. Teukolsky, W. T. Vetterling and B. P. Flannery, *Numerical recipes in C, 2nd Edition*, (Cambridge university press, 1992).
34. M. B. Wilk and R. Gnanadesikan, Probability plotting methods for the analysis of data, *Biometrika*, **55(1)**, (1968), p. 1-17.

Photo and Bibliography

Ulrik Söderström received his M.S. degree in Electrical engineering and Ph.D. degree in Applied electronics from Umeå University, Umeå, Sweden, in 2003 and 2008, respectively.

From October 2008 to March 2009 he was at the Department of Applied Physics and Electronics (TFE), Umeå University, as lecturer. He is currently working as Assistant Professor at Digital Media Lab, Umeå University.

Haibo Li received his M.S. degree from Southeast University, Nanjing, China, in 1986 and his Ph.D. degree in Information Theory from Linköping University, Sweden, in 1993. After graduation Dr. Li joined the technical faculty of Linköping University first as Assistant Professor and then he was promoted to Associate Professor in 1997. During his period at Linköping University he developed advanced image and video compression algorithms, including extremely low bitrate video compression, 3D

video transmission, tele-operation and telepresence.

He is currently Professor in Signal Processing at the Department of Applied Physics and Electronics (TFE), Umeå University, Sweden. After joining Umeå University in 1999, Prof. Li is now directing the Digital Media Lab, Umeå Center for Interaction Technology (UCIT), Umeå University, and working on advanced Human, Thing and Information interaction techniques. Prof. Li has been chairing sections at relevant international conferences and was actively involving in MPEG activities in low bitrate video compression.

Prof Li has contributed to several EU projects, like VIDAS, SCALAR, INTERFACE and MUCHI. He has published more than 100 technical papers including chapters in books and holds six international patents as the first inventor in multimedia area. Prof. Li is also active in commercialization of his research works. Four companies have been spun-off from Digital Media Lab and one of them was listed in the London Stock Market, 2005.