6

Asymmetrical Principal Component Analysis Theory and its Applications to Facial Video Coding

Ulrik Söderström and Haibo Li
Digital Media Lab, Dept. of Applied Physics and Electronics, Umeå University
Sweden

1. Introduction

The use of video telephony has not become a big success but it still has potential to become widespread. Video communication is becoming a well-used application for both personal use and corporations. This kind of communication is used in conversations between people and is essential for saving travel bills for companies. Less traveling also saves the environment and is therefore expected to be an important factor in the future. Even if the available bandwidth will increase it is desirable to use as low bandwidth as possible for communication since less bandwidth means lower cost, more availability in networks and less sensitivity to network delays. As more video is transmitted over networks, lower bandwidth need for video transmission means that more users can make use of video at the same time. Low bandwidth means low power consumption for transmission while low encoding and decoding complexity means low power consumption when the video is encoded and decoded. The impact of power consumption is expected to become much more important in the future as the availability of power is decreased and pollution from energy production needs to be halted.

Every human face is contained within a space called the face space. Every face can be recognized, represented or synthesized with this space. Principal component analysis (PCA) [Jolliffe (1986)] can be used to create a compact representation of this space. This enables PCA to be used for highly efficient video coding and other image processing tasks. The faces in the face space all have the same facial expression but PCA can also be used to create a space with different facial expressions for a single person. This is referred to as the personal face space, facial mimic space or personal mimic space [Ohba et al. (1998)]. This space consists of faces for a single person but with several different facial expressions. According to the American psychologist Paul Ekman it is enough to model six basic emotions to actually model all facial expressions [Ekman & Friesen (1975); Ekman (1982)]. The six basic emotions; happiness, sadness, surprise, fear, anger and disgust (Fig. 1), are blended in different ways to create all other possible expressions.

The combination of basic emotions is not directly applicable for linear processing with images so more than six dimensions are needed. We have previously evaluated exactly how many dimensions that are needed to reach a certain representation quality [Söderström & Li (2010)]. Efficient use of PCA for modeling of any data requires that the global motion is removed from the data set. For facial video this motion corresponds to motion of the entire head, e.g., positional shift and facial rotation. The motion that is modeled with PCA is the local

(a) Happiness (b) Sadness (c) Surprise

(d) Fear (e) Anger (f) Disgust

Fig. 1. The six basic emotions.

motion, i.e., the changes in the face, the facial mimic. The global motion can be removed with hardware techniques, e.g., hands-free video equipment [Söderström & Li (2005a)] or software implementations such as facial detection and feature tracking.

PCA provides a natural way for scaling video regarding quality. For the same encoding the decoder can select how many dimensions of the space that are used for decoding and thus scale the quality of the reconstructed video. The built-in scalability of PCA is easily utilized in video compression.

The operations with PCA involves all pixels, K, in a frame. When PCA is used for video compression the complexity for encoding and decoding is linearly dependent on K. It is desirable to have a low complexity for encoding but it is also desirable to have a high spatial resolution on the decoded video. A technique that allows the use of different areas for encoding and decoding is needed.

PCA extracts the most important information in the data based on the variance of the data. When it comes to video frames PCA extracts the most important information based on the pixel variance. The pixel variance is examined in section 5.1. Some pixels may have a high variance but no semantic importance for the facial mimic. These pixels will degrade the model efficiency for the facial mimic. To prevent that these high variance semantically unimportant pixels have effect on the model a region of interest (ROI) can be cropped or extracted from the video frames.

In this article we will examine how part of the frames can be used for encoding while we decode the entire frames. The usage of only a part of the frame for encoding while using full frame decoding is called asymmetrical PCA (aPCA) and it has been introduced by Söderström and Li [Söderström & Li (2008)]. In this work we will focus on different extractions and different sizes of the ROI.

The user can determine the important area in a video sequence or it can automatically be extracted using low-level processing of the frames. We present five different methods for extracting a ROI from video sequences where the face of a person is the most prominent information. An example of how the variance of the individual pixels vary is presented. This example clearly shows the motivation behind using a ROI instead of the entire frame. An example that visualizes the quality of the individual pixels is presented. This example shows that the quality for the semantically important pixels actually is increased when less information is used for encoding; if the correct information is used. Previously we presented how aPCA is used to reduce the encoding complexity [Söderström & Li (2008)]. Here we describe how aPCA can be used to reduce the decoding complexity as well.
Research related to this work is discussed in the next section and video coding based on principal component analysis (PCA) is explained in section 3. Asymmetrical PCA (aPCA) and how it is used to reduce encoder and decoder complexity is explained in section 4. Practical experiments of ROI extraction and usage with aPCA are explained in section 5 and the article is concluded in section 6.

2. Related work

Facial mimic representation has previously been used to encode sequences of faces [Torres & Prado (2002); Torres & Delp (2000)] and head-and-shoulders [Söderström & Li (2005a; 2007)]. These attempts try to represent facial video with a high quality at a very low bitrate. General video coding do not use PCA; the reigning transform is Discrete Cosine Transform (DCT) [Schäfer et al. (2003); Wiegand et al. (2003)]. Representation of facial video through DCT does not provide sufficiently high compression by itself and is therefore combined with motion estimation (temporal compression). DCT and block-matching requires several DCT-coefficients to encode the frames and several possible movements of the blocks between the frames. Consequently, the best codec available today does not provide high quality video at very low bitrates even if the video is suitable for high compression.
Video frames can also be represented as a collection of features from an alphabet. This is how a language functions; a small amount of letters can be ordered in different ways to create all the words in a language. By building an alphabet for video features it should be possible to model all video frames as a combination of these features. The encoder calculates which features that a frame consists of and transmit this information to the decoder which reassembles the frame based on the features. Since only information about which features the frame consists of is transmitted such an approach reach very low bitrates. A technique that uses such an alphabet is Matching Pursuit (MP) [Neff & Zakhor (1997)].
Facial images can be represented by other techniques then with video. A wireframe that has the same shape as a human face is used by several techniques. To make the wireframe move as a face it is sufficient to transmit information about the changes in the wireframe. To give the wireframe a more natural look it is texture-mapped with a facial image. Techniques that make use of a wireframe to model facial images are for example MPEG4 facial animation [Ostermann (1998)] and model based coding [Aizawa & Huang (1995); Forchheimer et al. (1983)]. Both of these techniques reach very low bitrate and can maintain high spatial resolution and framerate. A statistical shape model of a face is used by Active Appearance Model (AAM) [Cootes et al. (1998)]. AAM also use statistics for the pixel intensity to improve the robustness of the method.
All these representation techniques have serious drawbacks for efficient usage in visual communication. Pighin *et al.* provides a good explanation why high visual quality is

important and why video is superior to animations [Pighin et al. (1998)]. The face simply exhibits so many tiny creases and wrinkles that it is impossible to model with animations or low spatial resolution. To resolve this issue the framerate can be sacrificed instead. 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 [Wang & Cohen (2005)]. But high framerate and high spatial resolution are important for several visual tasks; framerate for some, resolution for others and some tasks require both [Lee & Eleftheriadis (1996)]. Any technique that want to provide video at very low bitrates must be able to provide video with high spatial resolution, high framerate and have natural-looking appearance.
Methods that are presented in Video coding (Second generation approach) [Torres & Kunt (1996)] make use of certain features for encoding instead of the entire video frame. This idea is in line with aPCA since only part of the information is used for encoding in this technique.
Scalable video coding (SVC) has high usage for video content that is received by heterogenous devices. The ability to display a certain spatial resolution and/or visual quality might be completely different if the video is received by a cellular phone or a desktop computer. The available bandwidth can also limit the video quality for certain users. The encoder must encode the video into layers for the decoder to be able to decode the video in layered fashion. Layered encoding has therefore been given much attention in the research community. A review of the scalable extension for H.264 is provided by Schwarz *et.al.* [Schwarz et al. (2007)].

3. Principal component analysis video coding

First, we introduce video compression with regular principal component analysis (PCA) [Jolliffe (1986)]. Any object can be decomposed into principal components and represented as a linear mixture of these components. The space containing the facial images is called Eigenspace Φ and there as many dimensions of this space as there are frames in the original data set. When this space is extracted from a video sequence showing the basic emotions it is actually a personal mimic space. The Eigenspace Φ=$\{\phi_1\ \phi_2\ ...\ \phi_N\}$ is constructed as

$$\phi_j = \sum_i b_{ij}(\mathbf{I}_i - \mathbf{I}_{\underline{0}}) \tag{1}$$

where b_{ij} are values from the Eigenvectors of the covariance
matrix $\{(\mathbf{I}_i - \mathbf{I}_{\underline{0}})^T(\mathbf{I}_j - \mathbf{I}_{\underline{0}})\}$. $\mathbf{I}_{\underline{0}}$ is the mean of all video frames and is constructed as:

$$\mathbf{I}_{\underline{0}} = \tfrac{1}{N}\sum_{j=1}^{N}\mathbf{I}_j \tag{2}$$

Projection coefficients $\{\alpha_j\}$=$\{\alpha_1\ \alpha_2\ ...\ \alpha_N\}$ can be extracted for each video frame through projection:

$$\alpha_j = \phi_j(\mathbf{I} - \mathbf{I}_{\underline{0}})^T \tag{3}$$

Each of the video frames can then be represented as a sum of the mean of all pixels and the weighted principal components. This representation is error-free if all N principal components are used.

$$\mathbf{I} = \mathbf{I}_{\underline{0}} + \sum_{j=1}^{N}\alpha_j\phi_j \tag{4}$$

Since the model is very compact many principal components can be discarded with a negligible quality loss and a sum with fewer principal components M can represent the image.

$$\hat{\mathbf{I}} = \mathbf{I}_{\underline{0}} + \sum_{j=1}^{M} \alpha_j \phi_j \tag{5}$$

where M is a selected number of principal components used for reconstruction ($M < N$).
The extent of the error incurred by using fewer components (M) than (N) is examined in [Söderström & Li (2010)]. With the model it is possible to encode entire video frames to only a few coefficients $\{\alpha_j\}$ and reconstruct the frames with high quality. A detailed description and examples can be found in [Söderström & Li (2005a;b)].
PCA video coding provides natural scalable video since the quality is directly dependent on the number of coefficients M that are used for decoding. The decoder can scale the quality of the video frame by frame by selecting the amount of coefficients used for decoding. This gives the decoder large freedom to scale the video without the encoder having to encode the video into scalable layers. The scalability is built-in in the reconstruction process and the decoder can easily scale the quality for each individual frame.

4. Asymmetrical principal component analysis video coding

There are two major issues with the use of full frame encoding:

1. The information in the principal components are based on all pixels in the frame. Pixels that are part of the background or are unimportant for the facial mimic may have large importance on the model. The model is affected by semantically unimportant pixels.
2. The complexity of encoding, decoding and model extraction is directly dependent on the spatial resolution of the frames, i.e., the number of pixels, K, in the frames. Video frames with high spatial resolution will require more computations than frames with low resolution.

When the frame is decoded it is a benefit of having large spatial resolution (frame size) since this provides better visual quality. A small frame should be used for encoding and a large frame for decoding to optimize the complexity and quality of encoding and decoding. This is possible to achieve through the use of pseudo principal components; information where not all the data are principal components. Parts of the video frames are considered to be important; they are regarded as foreground $\mathbf{I}^f$.

$$\mathbf{I}^f = crop(\mathbf{I}) \tag{6}$$

The Eigenspace for the foreground Φ^f=$\{\phi_1^f \ \phi_2^f \ \ldots \ \phi_N^f\}$ is constructed according to the following formula:

$$\phi_j^f = \sum_i b_{ij}^f (\mathbf{I}_i^f - \mathbf{I}_{\underline{0}}^f) \tag{7}$$

where b_{ij}^f are values from the Eigenvectors of the covariance matrix $\{(\mathbf{I}_i^f - \mathbf{I}_{\underline{0}}^f)^T (\mathbf{I}_j^f - \mathbf{I}_{\underline{0}}^f)\}$ and $\mathbf{I}_{\underline{0}}^f$ is the mean of the foreground. Encoding and decoding is performed as:

$$\alpha_j^f = (\phi_j^f)(\mathbf{I}^f - \mathbf{I}_{\underline{0}}^f)^T \tag{8}$$

$$\hat{\mathbf{I}}^f = \mathbf{I}^f_{\underline{0}} + \sum_{j=1}^{M} \alpha^f_j \phi^f_j \tag{9}$$

where $\{\alpha^f_j\}$ are coefficients extracted using information from the foreground $\mathbf{I}^f$. The reconstructed frame $\hat{\mathbf{I}}^f$ has smaller size and contains less information than a full size frame. A space which is spanned by components where only the foreground is orthogonal can be created. The components spanning this space are called pseudo principal components and this space has the same size as a full frame:

$$\phi^p_j = \sum_i b^f_{ij}(\mathbf{I}_i - \mathbf{I}_{\underline{0}}) \tag{10}$$

From the coefficients $\{\alpha^f_j\}$ it is possible to reconstruct the entire frame:

$$\hat{\mathbf{I}} = \mathbf{I}_{\underline{0}} + \sum_{j=1}^{M} \alpha^f_j \phi^p_j \tag{11}$$

where M is the selected number of pseudo components used for reconstruction. A full frame video can be reconstructed (Eq. 11) using the projection coefficients from only the foreground of the video (Eq. 8) so the foreground is used for encoding and the entire frame is decoded. It is easy to prove that

$$\hat{\mathbf{I}}^f = crop(\hat{\mathbf{I}}) \tag{12}$$

since $\phi^f_j = crop(\phi^p_j)$ and $\mathbf{I}^f_{\underline{0}} = crop(\mathbf{I}_{\underline{0}})$.
aPCA provides the decoder with the freedom to decide the spatial size of the encoded area without the encoder having to do any special processing of the frames. Reduction in spatial resolution is not a size reduction of the entire frame; parts of the frame can be decoded with full spatial resolution. No quality is lost in the decoded parts; it is up to the decoder to choose how much and which parts of the frame it wants to decode. The bitstream is exactly the same regardless of what video size the decoder wants to decode. With aPCA the decoder can scale the reconstructed video regarding spatial resolution and area.

4.1 Reduction of complexity for the encoder

The complexity for encoding is directly dependent on the spatial resolution of the frame that should be encoded. The important factor for complexity is $K * M$, where K is the number of pixels and M is the chosen number of Eigenvectors. When aPCA is used the number of pixels k in the selected area gives a factor of $n = \frac{K}{k}$ in resolution reduction. The number of computations are at the same time decreased by $n * M$.

4.2 Reduction of complexity for the decoder

The complexity for decoding can be reduced when a part of the frame is used for both encoding and decoding. In the formulas above we only use the pseudo principal components for the full frame ϕ^p_j for decoding but if both Φ^p and Φ^f are used for decoding the complexity can be reduced. Only a few principal components of Φ^p are used to reconstruct the entire frame. More principal components from Φ^f are used to add details to the foreground.

	Reconstructed Quality (PSNR) [dB]		
ϕ	Y	U	V
1	28,0	33,6	40,2
5	32,7	35,8	42,2
10	35,2	36,2	42,7
15	36,6	36,4	43,0
20	37,7	36,5	43,1
25	38,4	36,5	43,2

Table 1. Reference results. Video encoded with PCA using the entire frame $\mathbf{I}$.

$$\hat{\mathbf{I}} = \mathbf{I}_{\underline{0}} + \sum_{j=1}^{L} \alpha_j^f \phi_j^p + \sum_{j=L+1}^{M} \alpha_j^f \phi_j^f \tag{13}$$

The result is reconstructed frames with slightly lower quality for the background but with the same quality for the foreground $\mathbf{I}^f$ as if only Φ_j^p was used for reconstruction. The quality of the background is decided by parameter L: a high L-value will increase the information used for background reconstruction and increase the decoder complexity. A low L-value has the opposite effect. The reduction in complexity (compression ratio CR) is calculated as:

$$CR = \frac{K(M+1)}{(1+L)K + (M-L)k} \tag{14}$$

When $k << K$ the compression ratio can be approximated to $CR \approx \frac{M+1}{L+1}$. A significant result is that spatial scalability is achieved naturally; the decoder can decide the size of the decoded frames without any intervention from the encoder.

5. Asymmetrical principal component analysis video coding: practical implementations

In this section we show several examples of using aPCA for compression of facial video sequences. We show five examples, all for facial video sequences. As a reference we present the quality for encoding the video sequences with regular PCA in Table 1. This is equal to using the Eigenspace Φ for both encoding and decoding of video sequences.

5.1 Experiment I: Encoding with the mouth as foreground, decoding with the entire frame

The mouth is the most prominent facial part regarding facial mimic. By representing the shape of the mouth the entire facial mimic can be represented quite well. In this experiment, the foreground $\mathbf{I}^f$ consists of the mouth area and $\mathbf{I}$ is the entire frame (Fig. 2). $\mathbf{I}^f$ has a spatial size of 80x64 and $\mathbf{I}$ is 240x176 pixels large.

PCA extracts the most important information in a video sequence based on the variance of the pixels. This means that a pixel with high variance is important and low variance means the opposite. When the entire image is used some pixels which belong to the background may have high variance and be considered important. But these pixels have no semantical importance for the facial mimic and only degrades the model for the facial mimic. Asymmetrical principal component analysis (aPCA) [Söderström & Li (2008)] allow the use of foreground, i.e., important area, for encoding and decoding of entire frames. Fig. 3 shows the variance of the individual pixels in one of the video sequences used in Experiment I. The

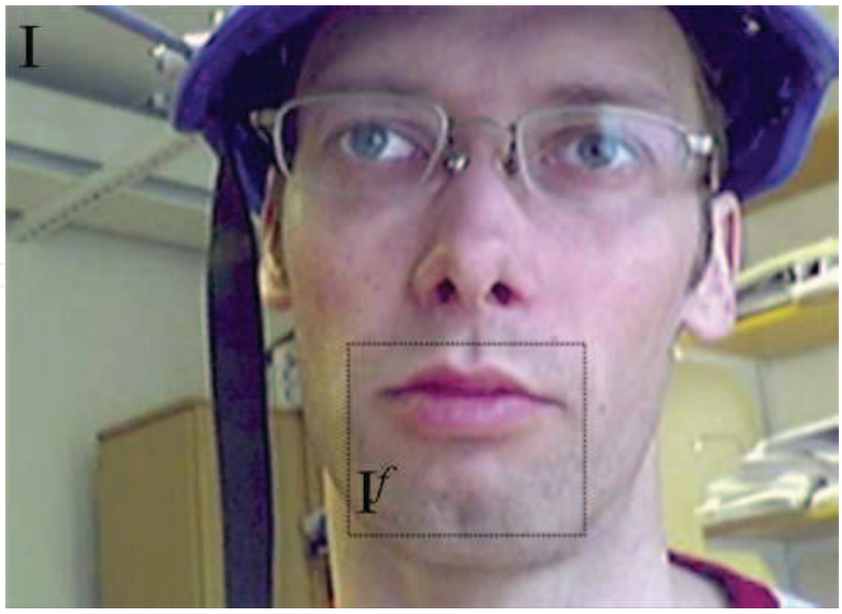


Fig. 2. The entire video frame and the foreground $\mathbf{I}^f$ used for Experiment I.

variance is noted in a heat-scale so white means low variance and yellow means high variance. It is clear that pixels from the background have high variance and are considered important by PCA. By only choosing the pixels that we know are semantically important we can increase the modeling efficiency of the facial mimic. With the use of aPCA it is still possible to decode the entire frames so no spatial resolution is lost.

Fig. 3. Variance for the individual pixels in a video sequence. *White = low variance Yellow = high variance*

	Lowered rec. qual. (PSNR) [dB]		
ϕ	Y	U	V
5	-1,4	-0,3	-0,2
10	-1,8	-0,3	-0,2
15	-2,0	-0,2	-0,2
20	-2,0	-0,1	-0,2
25	-2,1	-0,1	-0,2

Table 2. Average lowered reconstruction quality for 10 video sequences for a foreground $\mathbf{I}^f$ consisting of the mouth area (Experiment I).

The reconstruction quality is measured compared to the reconstruction quality when the entire frame $\mathbf{I}$ is used for encoding (Table 1). We also compare the complexity for encoding with $\mathbf{I}$ and $\mathbf{I}^f$. The reduction in complexity is calculated as the number of saved pixels. Since we use YUV subsampling 4:1:1 the number of pixels in $\mathbf{I}^f$ is 7680 and $\mathbf{I}$ consists of 63360 pixels. The reduction in complexity is then slightly more than 8 times. Table 2 shows the average reduction in reconstruction quality for 10 video sequences.

The quality for the pixels in the foreground is improved and the pixels which not are part of the foreground is reconstructed with lower quality when $\mathbf{I}^f$ is used for encoding. The reconstruction quality for each individual pixel in one video sequence is shown in Fig. 4. This figures clearly shows the advantage of aPCA compared to regular PCA. Semantically important pixels are reconstructed with higher quality while unimportant pixels have less reconstruction quality with aPCA compared to PCA.

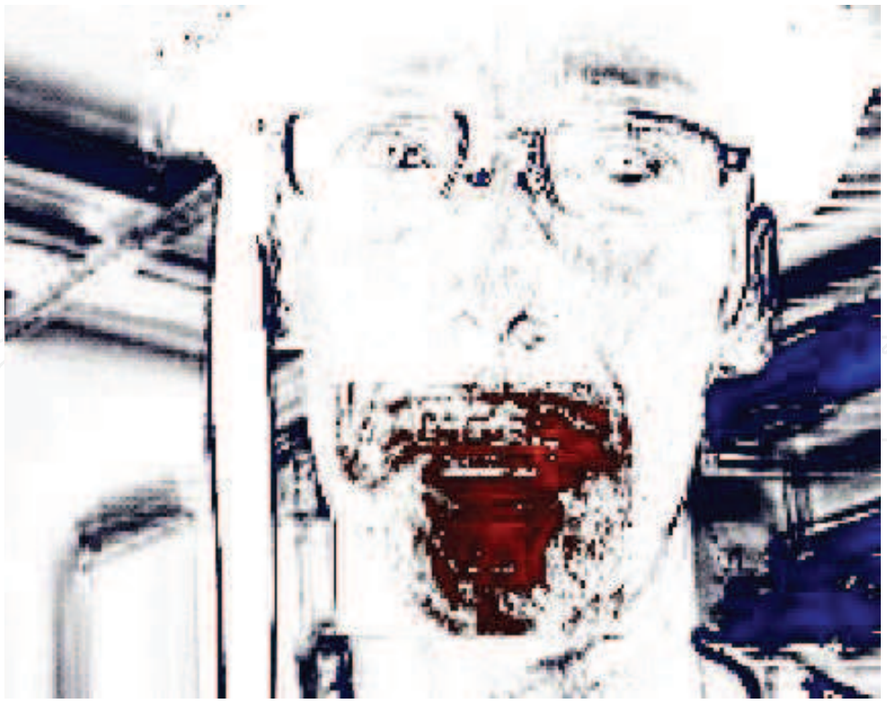
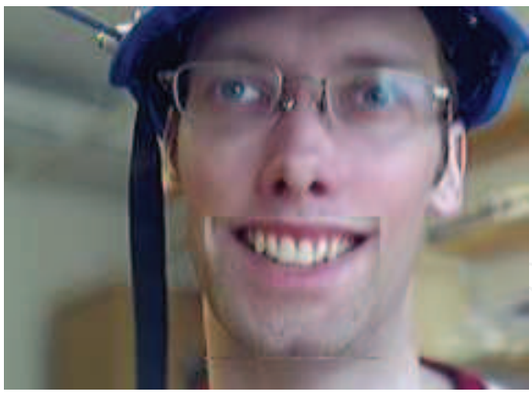
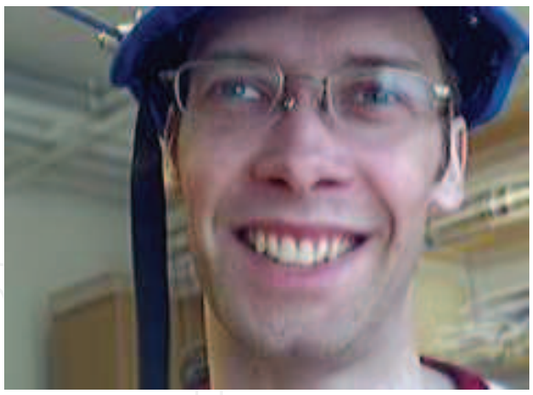


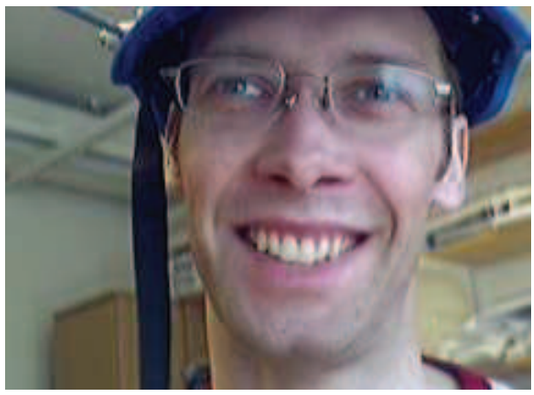
Fig. 4. Individual pixel PSNR for encoding with foreground $\mathbf{I}^f$ compared to encoding with the entire frame $\mathbf{I}$ *red*=improved PSNR *blue*=reduced PSNR.



(a) L=0,M=15



(b) L=10,M=15



(c) L=15,M=15

Fig. 5. Reconstructed frames with different setups for Experiment II.

5.2 Experiment II: Encoding with the mouth as foreground, decoding with the entire frame and the mouth area

In this section we show how the complexity for the decoder can be reduced by focusing on the reconstruction quality of the foreground. According to the equations in section 4.2 we reconstruct the frame with different background qualities. Fig. 5 show example frames for different L-values when M is 15.

Table 1 showed the average reconstruction result for encoding with the entire frame **I**. The reduction in reconstruction quality compared to those values are shown in Table 3. The PSNR for the foreground is the same regardless of the value of L since M always is 15 for the results in the table. The complexity reduction is also noted in Table 3. The complexity is calculated for a mouth area of 80x64 pixels and a total frame size of 240x176 pixels with M=15.

5.3 Experiment III: Encoding with the mouth and eyes as foreground, decoding with the entire frame

The facial mimic is dependent on more than the mouth area; the facial mimic of a person can be modelled accurately by modeling the mouth and the eyes. The mimic also contains small changes in the nose region but most of the information is conveyed in the mouth and eye regions. By building a model containing the changes of the mouth and the eyes we can model the facial mimic and we don't need to model any information without semantical importance.

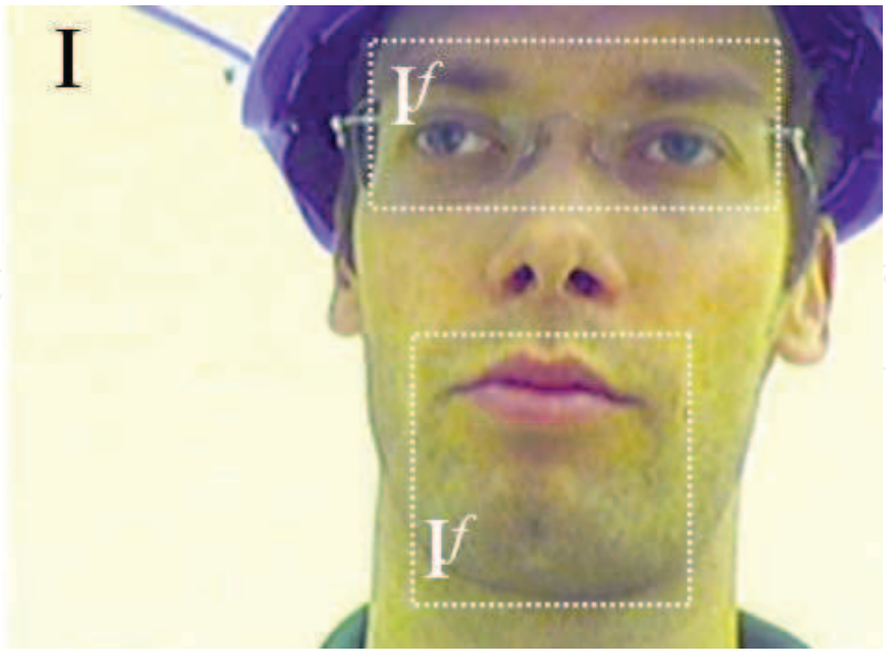


Fig. 6. Foreground $\mathbf{I}^f$ with the eyes and the mouth.

The areas which are chosen as foreground are shown in Fig. 6. The area around both eyes and the mouth are used as foreground $\mathbf{I}^f$ while the entire frame $\mathbf{I}$ is used for decoding. $\mathbf{I}^f$ has a spatial size of 176x64 and $\mathbf{I}$ still has a size of 240x176. The complexity for encoding is increased with an average of 55 % compared to only using the mouth area as foreground (Experiment I) but the complexity is still reduced $\approx$ 4 times compared to using the entire frame for encoding. The quality is also increased compared to experiment I (Table 4) and more accurate information about the eyes can be modeled.

5.4 Experiment IV: Encoding with features extracted through edge detection and dilation, decoding with the entire frame

In the previous three experiments we have chosen the foreground area based on prior knowledge about the facial mimic. In the following two experiments we will show how to select the region of interest automatically. In this experiment we choose one frame from a video sequence and use this to extract the foreground $\mathbf{I}^f$.

1. A representative frame is selected. This frame should contain a highly expressive appearance of the face.

	Lowered rec. qual. (PSNR) [dB]			
ϕ	Y	U	V	CR factor
1	-1,5	-0,7	-0,5	5,6
5	-1,4	-0,5	-0,4	2,4
10	-1,3	-0,3	-0,2	1,4
15	-1,2	-0,3	-0,2	1

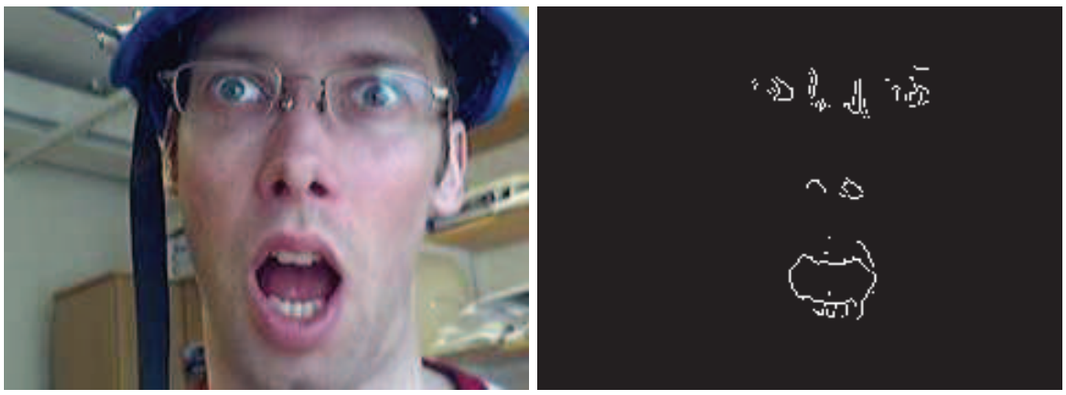
Table 3. Average lowered reconstruction quality for 10 video sequences using the mouth and eyes as foreground. (Experiment II) The complexity reduction (CR) factor is also shown and is calculated with M=15. In this calculation L is equal to Φ.

	Lowered rec. qual. (PSNR) [dB]		
ϕ	Y	U	V
5	-1,1	-0,2	-0,2
10	-1,5	-0,2	-0,1
15	-1,5	-0,2	-0,1
20	-1,6	-0,1	-0,1
25	-1,6	0	-0,2

Table 4. Average lowered reconstruction quality for 10 video sequences using an area around the mouth and the eyes for encoding (Experiment III).

2. Face detection detects the face in the selected frame.
3. Edge detection extracts all the edges in the face for the selected frame.
4. Dilation is used on the edge image to make the edges thicker. Every pixel in the dilated image which is 1 (white) is used for encoding.

The face detection method we use is described in [Le & Li (2004)]. The resulting area is similar to the area around the eyes and the mouth in the previous experiment. The result is shown in Table 5.



(a) Original (b) Detected edges

(c) Dilated edges (White pixels are foreground $\mathbf{I}^f$)

Fig. 7. Region of interest extraction with edge detection and dilation.

	Lowered rec. qual. (PSNR) [dB]		
ϕ	Y	U	V
5	-1,1	-0,2	-0,2
10	-1,4	-0,2	-0,2
15	-1,5	-0,1	-0,2
20	-1,6	-0,1	-0,1
25	-1,6	0	-0,1

Table 5. Average lowered reconstruction quality for 10 video sequences using a foreground $\mathbf{I}^f$ from Experiment IV.

The average size of this area for 10 video sequences is lower than the previous experiment. It is 11364 pixels large and this would correspond to an square area with the size of $\approx$ 118x64. This corresponds to a reduction in encoding complexity of $\approx$ 5,6 times compared to using the entire frame $\mathbf{I}$.

5.5 Experiment V: Encoding with edge detected features as foreground, decoding with the entire frame

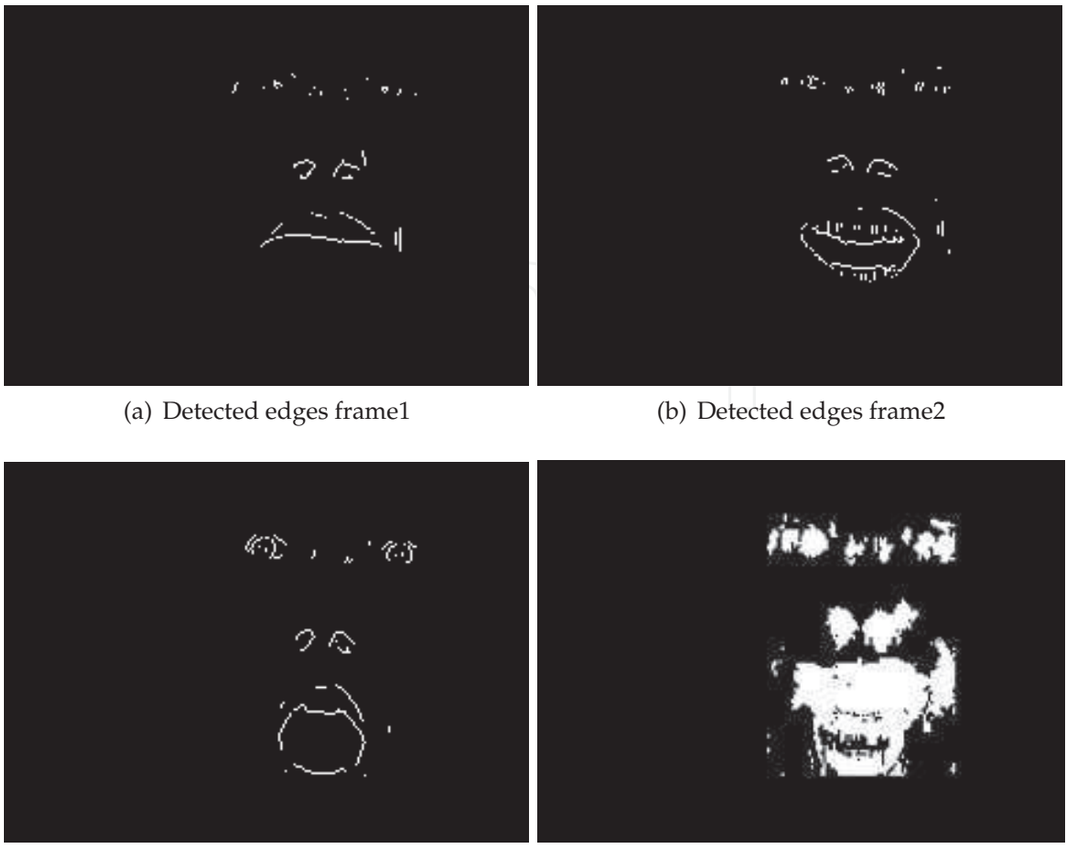
Another way to automatically extract the foreground is to use feature detection without dilation. This is a fully automatical procedure since no key frame is selected manually.

1. Face detection [Le & Li (2004)] detects the face in each frame.
2. Edge detection extracts all the edges in the face for each frame.
3. Every edge is gathered in one edge image. Where there is an edge in at least one of the frames there will be an edge in the total frame. Every pixel with an edge in any frame is used for encoding.

The complexity is on average reduced 11,6 times when the area extracted in this way is used for encoding compared to using the entire frame and $\approx$ 3 times compared to using the area around the mouth and the eyes from Experiment III. The reconstruction quality is shown in Table 6.
The reconstruction quality is almost the same for all cases when information from both the eyes and the mouth is used for encoding. When only the mouth is used for encoding the quality is lower. The complexity is at the same time reduced for all the different aPCA implementations. It is reduced heavily when the area is extracted from edges (Experiment V).

	Lowered rec. qual. (PSNR) [dB]		
ϕ	Y	U	V
5	-1,2	-0,3	-0,3
10	-1,6	-0,3	-0,1
15	-1,7	-0,2	-0,2
20	-1,7	-0,2	-0,2
25	-1,7	-0,3	-0,2

Table 6. Average lowered reconstruction quality for 10 video sequences using combined edges from all frames in a video for foreground extraction (Experiment V).



(a) Detected edges frame1 (b) Detected edges frame2

(c) Detected edges frame3 (d) Detected edges in all frames

Fig. 8. Region of interest extraction with edge detection on all frames.

6. Conclusion

We show some of the potential of asymmetrical principal component analysis (aPCA) for compression of facial video sequences. It can efficiently be used to reduce the complexity of encoding and decoding with only a slight decrease in reconstruction quality. The complexity for encoding can be reduced more than 10 times and the complexity for decoding is also reduced at the same time as the objective quality is lowered slightly, i.e., 1,5 dB (PSNR). aPCA is also very adaptive for heterogenous decoding since a decoder can select which size of video frames it wants to decode with the encoder using the same video for encoding. PCA provides natural scalability of the quality and aPCA also provides scalability in spatial resolution with the same encoding. The freedom of assembling the reconstructed frames differently also provides the decoder with the freedom to select different quality for different parts of the frame.

Low bitrate video is far from realized for arbitrary video. Regular video encoding has not reached these low bitrates and previous solutions to low bitrate facial video/representation do not have a natural-looking appearance. aPCA has a major role to play here since it can provide natural-looking video with very low bitrate and low encoding and decoding complexity.

7. References

Aizawa, K. & Huang, T. (1995). Model-based image coding: Advanced video coding techniques for very low bit-rate applications, *Proc. of the IEEE* 83(2): 259–271.

Cootes, T., Edwards, G. & Taylor, C. (1998). Active appearance models, *In Proc. European Conference on Computer Vision. (ECCV)*, Vol. 2, pp. 484–498.

Ekman, P. (1982). *Emotion in the Human Face*, Cambridge University Press, New York.

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

Forchheimer, R., Fahlander, O. & Kronander, T. (1983). Low bit-rate coding through animation, *In Proc. International Picture Coding Symposium PCS–83*, pp. 113–114.

Jolliffe, I. (1986). *Principal Component Analysis*, Springer-Verlag, New York.

Le, H.-S. & Li, H. (2004). Face identification from one single sample face image, *Proc. of the IEEE Int. Conf. on Image Processing (ICIP)*.

Lee, J. & Eleftheriadis, A. (1996). Spatio-temporal model-assisted compatible coding for low and very low bitrate video telephony, *Proceedings, 3rd IEEE International Conference on Image Processing (ICIP 96)*, Lausanne, Switzerland, pp. II.429–II.432.

Neff, R. & Zakhor, A. (1997). Very low bit-rate video coding based on matching pursuits, *IEEE Transactions on Circuits and Systems for Video Technology* 7(1): 158–171.

Ohba, K., Clary, G., Tsukada, T., Kotoku, T. & Tanie, K. (1998). Facial expression communication with fes, *International conference on Pattern Recognition*, pp. 1378–1381.

Ostermann, J. (1998). Animation of synthetic faces in mpeg-4, *Proc. of Computer Animation, IEEE Computer Society*, pp. 49–55.

Pighin, F., Hecker, J., Lishchinski, D., Szeliski, R. & Salesin, D. H. (1998). Synthesizing realistic facial expression from photographs, *SIGGRAPH Proceedings*, pp. 75–84.

Schäfer, R., Wiegand, T. & Schwarz, H. (2003). The emerging h.264 avc standard, *EBU Technical Review* 293.

Schwarz, H., Marpe, D. & Wiegand, T. (2007). Overview of the scalable video coding extension of the h.264/avc standard, *Circuits and Systems for Video Technology, IEEE Transactions on* 17(9): 1103–1120.

Söderström, U. & Li, H. (2005a). Full-frame video coding for facial video sequences based on principal component analysis, *Proceedings of Irish Machine Vision and Image Processing Conference (IMVIP)*, pp. 25–32. online: www.medialab.tfe.umu.se.

Söderström, U. & Li, H. (2005b). Very low bitrate full-frame facial video coding based on principal component analysis, Signal and Image Processing Conference (SIP'05). online: www.medialab.tfe.umu.se.

Söderström, U. & Li, H. (2007). Eigenspace compression for very low bitrate transmission of facial video, *IASTED International conference on Signal Processing, Pattern Recognition and Application (SPPRA)*.

Söderström, U. & Li, H. (2008). Asymmetrical principal component analysis for video coding, *Electronics letters* 44(4): 276–277.

Söderström, U. & Li, H. (2010). Representation bound for human facial mimic with the aid of principal component analysis, *International Journal of Image and Graphics (IJIG)* 10(3): 343–363.

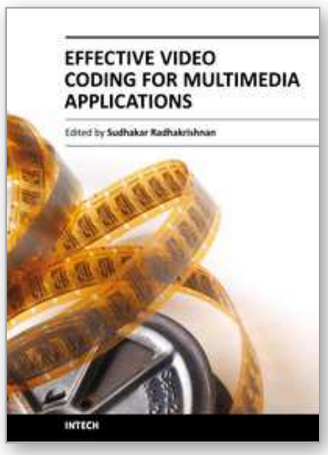
Torres, L. & Delp, E. (2000). New trends in image and video compression, *Proceedings of the European Signal Processing Conference (EUSIPCO)*, Tampere, Finland.

Torres, L. & Kunt, M. (1996). *Video Coding (The Second Generation Approach)*, Kluwer Academic Publishers.

Torres, L. & Prado, D. (2002). A proposal for high compression of faces in video sequences using adaptive eigenspaces, *Proceedings International Conference on Image Processing* 1: I–189– I–192.

Wang, J. & Cohen, M. F. (2005). Very low frame-rate video streaming for face-to-face teleconference, *DCC '05: Proceedings of the Data Compression Conference*, pp. 309–318.

Wiegand, T., Sullivan, G., Bjontegaard, G. & Luthra, A. (2003). Overview of the h.264/avc video coding standard, *IEEE Trans. Circuits Syst. Video Technol.,* 13(7): 560–576.



Effective Video Coding for Multimedia Applications

Edited by Dr Sudhakar Radhakrishnan

ISBN 978-953-307-177-0
Hard cover, 292 pages
Publisher InTech
Published online 26, April, 2011
Published in print edition April, 2011

Information has become one of the most valuable assets in the modern era. Within the last 5-10 years, the demand for multimedia applications has increased enormously. Like many other recent developments, the materialization of image and video encoding is due to the contribution from major areas like good network access, good amount of fast processors e.t.c. Many standardization procedures were carrried out for the development of image and video coding. The advancement of computer storage technology continues at a rapid pace as a means of reducing storage requirements of an image and video as most situation warrants. Thus, the science of digital video compression/coding has emerged. This storage capacity seems to be more impressive when it is realized that the intent is to deliver very high quality video to the end user with as few visible artifacts as possible. Current methods of video compression such as Moving Pictures Experts Group (MPEG) standard provide good performance in terms of retaining video quality while reducing the storage requirements. Many books are available for video coding fundamentals.This book is the research outcome of various Researchers and Professors who have contributed a might in this field. This book suits researchers doing their research in the area of video coding.The understanding of fundamentals of video coding is essential for the reader before reading this book. The book revolves around three different challenges namely (i) Coding strategies (coding efficiency and computational complexity), (ii) Video compression and (iii) Error resilience. The complete efficient video system depends upon source coding, proper inter and intra frame coding, emerging newer transform, quantization techniques and proper error concealment.The book gives the solution of all the challenges and is available in different sections.

How to reference

In order to correctly reference this scholarly work, feel free to copy and paste the following:

Ulrik Söderström and Haibo Li (2011). Asymmetrical Principal Component Analysis Theory and its Applications to Facial Video Coding, Effective Video Coding for Multimedia Applications, Dr Sudhakar Radhakrishnan (Ed.), ISBN: 978-953-307-177-0, InTech, Available from: http://www.intechopen.com/books/effective-video-coding-for-multimedia-applications/asymmetrical-principal-component-analysis-theory-and-its-applications-to-facial-video-coding

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai

Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821

IntechOpen

IntechOpen