M.I.T Media Laboratory Perceptual Computing Section Technical Report No. 317
Appears in: *IEEE Data Compression Conference, Snowbird, Utah, March 1995.*

An Automatic System for Model-Based Coding of Faces

Baback Moghaddam and Alex Pentland
Perceptual Computing Group, The Media Laboratory
Massachusetts Institute of Technology
20 Ames St., Cambridge, MA 02139

Abstract

We present a fully automatic system for 2D model-based image coding of human faces for potential applications such as video telephony, database image compression, and face recognition. The system operates by locating a face in the input image, normalizing its scale and geometry and representing it in terms of a parametric image model obtained with a Karhunen-Loeve basis. This leads to a compact representation of the face that can be used for both recognition as well as image compression. Good-quality facial images are automatically generated using approximately 100-bytes worth of encoded data. The system has been successfully tested on a database of nearly 2000 facial photographs.

1 Introduction

Model-based image coding of human faces has been proposed as a means of achieving quality at reduced bit-rates in applications such as video telephony. However, most existing systems rely on complex and brittle methods for the segmentation of the face from the background. It is clear that the success of any facial image coder depends on the ability to successfully segment the face from the scene and the ability to represent the facial appearance in a compact and parametric representation which is amenable to low-bit rate compression. The Karhunen-Loeve (KL) or principal component representation of 2D facial images is one parametric image model which is especially attractive from an image compression point-of-view: it yields statistically uncorrelated coefficients which are optimal in the mean-square-error sense for reconstruction.

Our system is computationally simple and is able to locate a face in an input scene despite changes in head scale, slight head tilts ($< 15^\circ$), moderate lighting variations and variable image contrast. This detection technique is based on *eigentemplates* [5] and is a generalization of the standard matched filter formulation and uses the KL expansion of a set of training faces. The use of an eigenspace formulation in the detection stage naturally suggests a similar approach for coding, but the detection system is general enough to be of utility to *any* model-based facial coding system.

Portions of the face processing system described in this paper were originally developed as part of a face recognition system [3]. The interface to the recognition system consists of a browsing tool called Photobook which allows the user to interactively search through the database based on facial similarity. The user begins by selecting the types of faces he/she wishes to examine; *e.g.,* adult Caucasian males. Photobook then presents the user with a screenful of the selected type of images. The user can select a face from among those presented and issue a search query. Photobook will then use the encoded parametric description of that face (derived by techniques described later in this paper) to search the entire database for possible matches. The user is then presented with the top candidates sorted by degree of similarity to the selected face. Figure 1 shows an example of a Photobook query. The face at the upper left selected by the user; the remainder of the faces are the most-similar faces amongst 575 images from the FERET database. Note that the first four images (in the top row) are all of the same person (taken a month apart and exhibiting different hairstyles, clothing, scale, etc.). Previously, the face recognition system has achieved a 95% recognition accuracy on the Media Lab database of 7,562 facial images [5].

2 KL Expansions for Facial Image Coding

The use of the KL expansion for characterization of human faces was first suggested by Sirovich & Kirby [7]. This scheme was later extended by Turk & Pentland [8] and others [4] to the problem of automatic face recognition. Welsh & Shah [10] also demonstrated a low-bit rate compression scheme for transmission of facial features such as lips, using a KL expansion.

Most of these systems (with the exception of [8] and later [5]) did not address the problem of detection, and used manual registration of images for demonstrating the concept. For a KL compression scheme to be useful in real-life coding applications, the face must first be detected and normalized prior to the computation of the expansion coefficients. This requires the ability to locate the face, estimate its scale and hence correct for positional and scale variations. In addition, individual facial features (such as eyes, nose and mouth) must be detected and used to geometrically normalize the shape of the face. Our systems accomplishes these pre-processing steps using a series of computationally simple image processing operations consisting of linear filtering, image warping, and point-wise transforms.

3 Face Detection

The standard detection paradigm in image processing is that of simple normalized correlation or template matching. This approach however is only optimal in the case of a *deterministic* signal embedded in white Gaussian noise.

Figure 1: Photobook: an interactive image database tool

When we begin to consider a target *class* detection problem — *e.g.,* finding a generic human face in a scene — we must incorporate the underlying probability distribution of the signal of interest. Subspace methods, such as Karhunen-Loeve expansions, allow for the compact representation of the statistical variability of the signal model and lead to much more robust signal detection schemes.

Indeed, the eigenspace formulation leads to a powerful alternative to simple template matching. The residual error in a KL expansion (referred to as the "distance-from-face-space" in the context of "eigenfaces" [8]) is a an effective indicator of a match. The residual error is easily computed using the projection coefficients and original signal energy. This detection strategy is equivalent to matching with *eigentemplates* and allows for a greater range of distortions in the input signal (including lighting, rotation and scale). In a statistical signal detection framework, the use of eigentemplates has been shown to yield superior performance in comparison with standard matched filtering [2].

Pentland *et al.* [5] used this same formulation for eigenspace representation of facial *features* (*e.g.,* eyes, noses and mouths). In this domain, the equivalent "distance-from-*feature*-space" (DFFS) can be effectively used for the detection of visual features. Given an input image, a distance map is constructed by computing the DFFS at each pixel. When using M eigenvectors, this requires M convolutions (which can be efficiently computed using an FFT) plus an additional local energy computation. The global minimum of this distance map is then selected as the best location of the target.

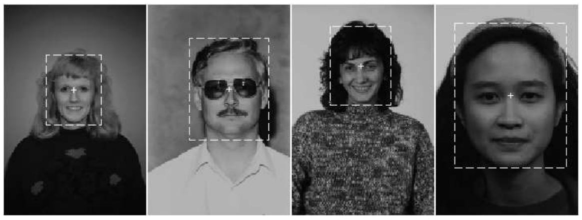
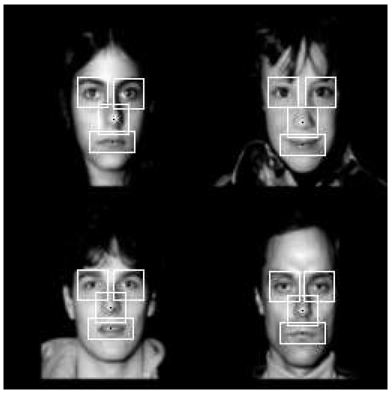


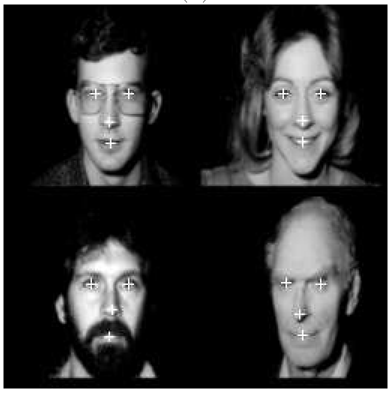
Figure 2: Multiscale Face Detection

3.1 Multiscale Face Detection

This detection technique can be easily extended to a *multiscale* search by using a single set of eigentemplates (at a fixed scale) and linearly remapping the input image through a given range of scales and computing a separate distance map at each scale. The estimate of the position *and* scale is obtained by identifying the best global minimum among all scale-indexed distance maps. Figure 2 shows several examples of the resulting detections. The estimated center of the face (midpoint of the eyes) is marked by crosshairs and the face scale is indicated by the dashed rectangle. This image region is then rescaled and repositioned in a fixed reference frame for subsequent processing.



(a)



(b)

Figure 3: (a) Examples of facial feature training templates used and (b) the resulting typical detections.

3.2 Facial Feature Detection

In addition to whole-face detection, automatic detection of facial features is also an important component for face processing. Over the years, various strategies for facial feature detection have been proposed, ranging from the early work of Kanade with edge-map projections [1], to more recent techniques using generalized symmetry operators [6] and multilayer perceptrons [9]. In our face processing system this task is critically important since after face detection, the scale-normalized face must then be geometrically normalized by aligning it with a canonical geometrical model.

The eigentemplate technique can be simply extended to the detection and coding of facial features, yielding eigeneyes, eigennoses and eigenmouths. In this eigenfeature representation the equivalent "distance-from-*feature*-space" (DFFS) can be effectively used for the detection of facial features. Given an input image, a feature distance-map is built by computing the DFFS at each pixel. When using n eigenvectors, this requires n convolutions (which can be efficiently computed using an FFT) plus an additional local energy computation. The global minimum of this distance map is then selected as the best feature match. The performance of the eigentemplate technique was recently tested on a database of approximately 8,000 "mugshot" photographs, where it achieved a 94% detection rate showing an order of magnitude improvement over standard template matching [5].

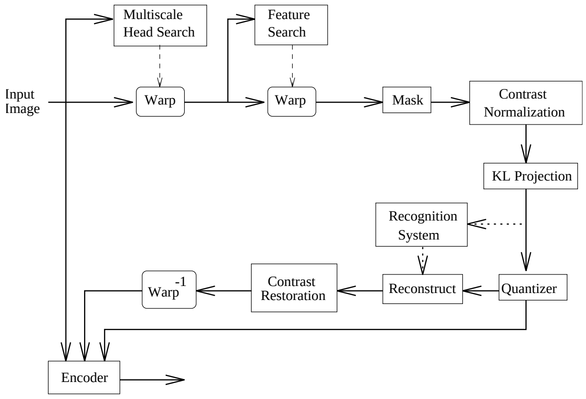


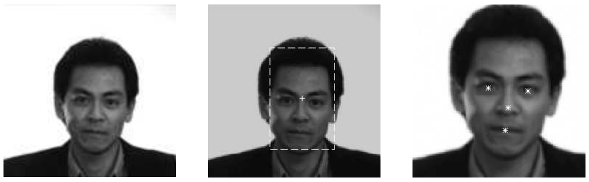
Figure 4: The face processing system.

Our system conducts a parallel search for the eyes, nose and mouth using separate sets of eigentemplates for each feature. However, in order to rule out single-point failures in detecting a feature's location, multiple local minima in each distance map are examined to determine the one combination of candidates which represent the most likely spatial configuration of features (consistent with *a priori* geometric constraints). The incorporation of geometrical constraints leads to a more robust method for detecting facial features. Figure 3 shows several examples of the training subimages used for computing the eigenfeatures and the resulting typical detections on novel images. The ability to automatically locate a face, normalize for scale and detect salient facial features leads naturally to a modular or layered representation, where a coarse (low-resolution) description of the whole head is augmented by additional (higher-resolution) details in terms of salient facial features. This modularity in face description has distinct advantages for face coding in low-bit rate video teleconferencing. This detection system was recently tested on the ARPA FERET face database where over two thousand facial photographs were processed with 97% reliability [3].

4 System Description

The block diagram of our face processing and coding system is shown in Figure 4. The first stage is a multiscale search for a face. After the position and scale have been identified the image is warped (scaled) to center the face at a standard scale. The feature detection stage then operates on this scale-normalized image. Figure 5 illustrates the results of the detection stages on an input image.

The coding power of the KL expansion is best exploited when the facial images are spatially registered and normalized with respect to lighting and contrast variations. Therefore, the face is geometrically warped by aligning the location of the detected facial features with those of a standard model. Next the image is masked so as to only include the interior of the face so as to concentrate the descriptive power of the KL expansion on the most salient parts of the face. In addition, after masking the image is contrast-normalized to compensate for changing global



(a) (b) (c)

Figure 5: (a) Input image, (b) face detection, (c) feature detections

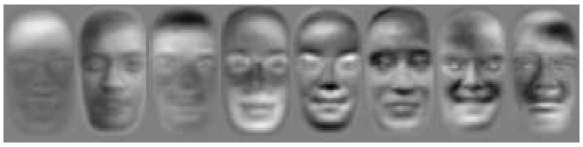


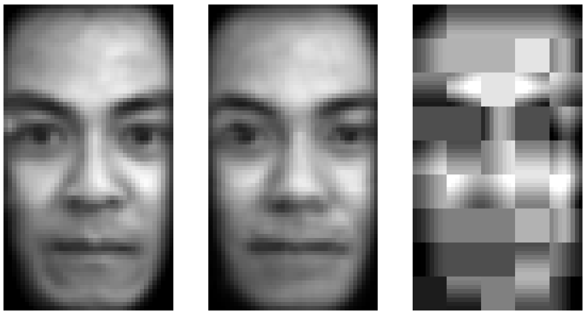
Figure 6: The KL basis functions for normalized faces.

illuminations as well as linear changes in the CCD camera's gain and offset.

Once the image is suitably normalized with respect to individual geometry and contrast, it is projected onto a set of eigenfaces. Figure 6 shows the first few eigenfaces obtained from a KL expansion on an ensemble of 500 normalized faces. In our system, the projection coefficients are used to index through a database to perform identity verification and recognition. In addition, For coding purposes, these coefficients are normalized by their standard deviations (the square-roots of the eigenvalues computed by the KL expansion) and quantized using a Lloyd-Max quantizer for a Gaussian source. Since the coefficients are ranked, we also use a variable number of quantization levels for different coefficients: the number of bits allocated to a given coefficient is proportional to its eigenvalue.

Figure 7 shows the normalized image, along with the reconstruction obtained using an 85-byte coding of the expansion coefficients. In order to compare our model-based compression scheme to a standard transform coder, we have also shown the lowest-quality acceptable JPEG compression of the normalized image (yielding a total of 540 bytes at a Q-factor of 2%).

Note that since all the transformations leading to the normalized image are reversible (with the exception of the masking), we can remap the reconstruction back into the original image, placing it at the correct location, with the correct scale and contrast. This leads to a partial face-only coded image as shown in Figure 8, where an additional 30-bytes are used to encode the 6 affine parameters of the inverse warp and the 2 parameters of the contrast normalization step. This partial coded image (at a cost of a mere 105-bytes) can then be used in conjunction with the original image for a variety of possible image compression schemes (*e.g.,* error-coding of the facial region). Note that this system leads very naturally to an *attention-based* coding scheme where, for example, only the salient region of the input image (*i.e.,* that containing the face) is coded with fidelity and the remainder is transmitted with a lossy compression scheme at low bit-rates. Such a scheme will preserve the quality of the facial image (nec-



(a) (b) (c)

Figure 7: (a) Normalized face, (b) 85-byte KL reconstruction, (c) 540-byte JPEG reconstruction (Q = 2%).

(a) (b)

Figure 8: (a) Original image, (b) 105-byte facial-coded partial image.

essary for recognition, etc.) while maintaining the reduced bandwidth needed in limited-capacity transmissions.

Finally, we note that our current system is implemented in general-purpose hardware (using standard UNIX workstations) and takes approximately 15 seconds to process an image. The majority of the processing load is in performing the convolutions required in computing the eigenspace projections. Therefore, with specialized image processing hardware, real-time operation at video frame rates should be feasible, especially in view of the fact that the temporal continuity of the video stream can be exploited to limit the search regions used by the detection subsystems from one frame to the next.

5 Conclusions

We have described a system that automatically detects faces and face features, and then maps them to a canonical view (*i.e.,* fixed position, scale, geometry and contrast) suitable for model-based compression. The system has been tested on more than 2,000 images, including wide variations in scale, contrast, etc., and has achieved a detection accuracy of 97%.

Our system is an unbalanced (asymmetric) coding scheme, requiring modest computational resources at the transmitter (mainly for the detection stages), but no more than standard decoding power at the receiver. Face detection and encoding requires approximately 15 seconds on a modern computer workstation; the reconstruction of a face image from its encoded representation requires less than one-hundredth of a second.

The ability to normalize extrinsic variations caused by scale and position mis-alignment, global illumination and contrast changes, has allowed us to fully exploit the power of the KL expansion for encoding differences in facial geometry between different individuals. This facial detection algorithm can also be used in conjunction with a conventional coding scheme, allowing preferential bit allocation to faces and facial features.

Acknowledgments

This research was funded by British Telecom and the US Army Research Laboratory. The FERET face database was provided by ARL.

References

[1] Kanade, T., "Picture processing by computer complex and recognition of human faces," Tech. Report, Kyoto University, Dept. of Information Science, 1973.

[2] Kumar, B., Casasent, D., and Murakami, H., "Principal Component Imagery for Statistical Pattern Recognition Correlators," *Optical Engineering,* vol. 21, no. 1, Jan/Feb 1982.

[3] Moghaddam, B. and Pentland, A., "Face recognition using view-based and modular eigenspaces," in *Automatic Systems for the Identification and Inspection of Humans*, SPIE vol. 2277. 1994.

[4] O'Toole, A., Abdi, H, Deffenbacher, K, and Valentin, D., "Low-dimensional representation of faces in higher dimensions of the face space," *Journal of the Optical Society A*, Vol. 10, pp. 405-410, 1993.

[5] Pentland, A., Moghaddam, B. and Starner, T., "View-based and modular eigenspaces for face recognition," *Proc. of IEEE Conf. on Computer Vision & Pattern Recognition*, June 1994.

[6] Reisfeld, D., Wolfson, H., and Yeshurun, Y., "Detection of Interest Points Using Symmetry,", *ICCV '90*, Osaka, Japan, Dec. 1990.

[7] Sirovich, L, and Kirby, M., "Low-dimensional procedure for the characterization of human faces," *J. Opt. Soc. Am. A*, Vol. 4, No. 3, March 1987, 519-524.

[8] Turk, M., and Pentland, A., "Eigenfaces for Recognition," *Journal of Cognitive Neuroscience,* Vol. 3, No. 1, 1991.

[9] Vincent, J. M., Waite, J. B., and Myers, D. J, "Automatic Location of Visual Features by a System of Multilayered Perceptrons," *IEE Proceedings,* vol. 139, no. 6, Dec. 1992.

[10] Welsh, J. W., and Shah, D., "Facial-Feature Image Coding Using Principal Components,", *Electronic Letters,* vol. 28, no. 22, October, 1992.