Short Papers

Face Recognition System Using Local Autocorrelations and Multiscale Integration

François Goudail, Eberhard Lange, Takashi Iwamoto, Kazuo Kyuma, and Nobuyuki Otsu

Abstract—In this paper we investigate the performance of a technique for face recognition based on the computation of 25 local autocorrelation coefficients. We use a large database of 11,600 frontal facial images of 116 persons, organized in training and test sets, for evaluation. Autocorrelation coefficients are computationally inexpensive, inherently shift-invariant and quite robust against changes in facial expression. We focus on the difficult problem of recognizing a large number of known human faces while rejecting other, unknown faces which lie quite close in pattern space. A multiresolution system achieves a recognition rate of 95%, while falsely accepting only 1.5% of unknown faces. It operates at a speed of about one face per second. Without rejection of unknown faces, we obtain a peak recognition rate of 99.9%. The good performance indicates that local autocorrelation coefficients have a surprisingly high information content.

Index Terms—Classification, face recognition, autocorrelation, object recognition, shift invariant feature extraction, multiresolution image analysis.

1 Introduction

AUTOMATIC face recognition is an important problem with many potential applications. Earlier works were mainly based on traditional methods such as extraction of facial features and template matching (see [1], [2] for recent reviews and [3] for a comparison of the approaches). More recently, Karhunen-Loeve expansion [4], deformable templates [5], [6], and feed-forward neural networks [7], [8], [9] were utilized, sometimes in connection with novel image processing methods [10].

In a previous article [11], we have applied a pattern recognition method proposed by Otsu and Kurita [12] and Kurita et al. [13] to the recognition of a large number of faces. The method is based on computation of local autocorrelation coefficients. A comparative study of three different classification methods and the introduction of a new rejection method increased the reliability of the system by more than a factor of 10 [14]. In this paper, we present the system configuration that resulted in the best overall performance—linear discriminant analysis in combination with contrast-based rejection criteria and multiresolution integration. The relatively low computational effort involved in both the learning and the recognition process of the system makes it very fast. In addition, it showed a good performance on a large database of 11,600 images of 116 different faces. These two characteristics are important in practical applications, which require short response time and high memorization capacity.

The methods primarily target the problem of face recognition, but should be general enough to be applied to other real world recognition tasks. In the next section, we describe the database used throughout this research, and explain the feature extraction, classification and rejection methods. In the third section, we continue with an analysis of the system performance. In the fourth section we introduce a computationally efficient modular architecture.

2 System Architecture

Fig. 1 illustrates the architecture of the face recognition system. In the first step, the input image is scaled down. In the second step, shift invariant features are extracted from the images, using a set of 25 local autocorrelation coefficients as feature vector. In the third step, classification of the feature vectors by mapping into a classification space and subsequent Euclidean distance classification takes place. Finally, the image is either recognized (classified as a known person) or rejected (classified as unknown person) according to certain rejection criteria. This section describes the methods employed in the individual steps, starting with a description of our database of images.

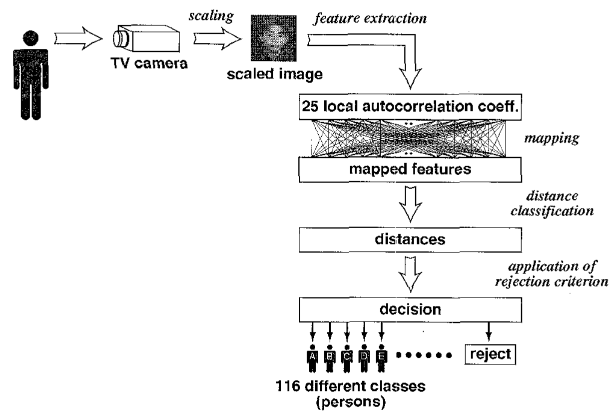

Fig. 1. System architecture.

2.1 Database of Facial Images

The database used throughout this research consists of 11,600 images of 116 persons, organized in training and test sets containing 50 images each. The recording of facial images is performed as follows: after asking the persons to sit down in front of a uniform, black background, two 30-second periods are recorded on videotape using a CCD camera. All persons wear dark company jackets. During the recording sessions, the persons are encouraged to move the head continuously, up to about 15 degrees up, down, to the left, and to the right. Later, a frame grabber operating at a speed of about three frames per second generates training and test sets from the 30-second video clips. The stored gray-scale images have a resolution of 180×120 pixels and a depth of 8 bits. Previous experiments with a database of 3,000 images showed that higher resolutions contribute very little to the capabilities of the recognition system.

The system is able to recognize unsegmented images—a feature rarely found in face recognition systems. Segmentation of a face from a fixed background and also from clothing is not subject of

• *F. Goudail is with the Laboratoire Signal et Image, Ecole Supérieure de Physique de Marseille, Domain Universitaire de Saint-Jérôme, 13397 Marseille Cedes 20, France.*
• *E. Lange, T. Iwamoto, and K. Kyuma are with the Neural and Parallel Processing Technology Department, Advanced Technology R&D Center, Mitsubishi Electric Corp., 1-1 Tsukaguchi-Honmachi 8-chome, Amagasaki, Hyogo, 661 Japan. E-mail: el@qua.crl.melco.co.jp.*
• *N. Otsu is with the Electrotechnical Laboratory, AIST MITI, 1-1-4 Umezono, Tsukuba-shi, Ibaraki, 305 Japan.*

Manuscript received Nov. 17, 1994; revised May 30, 1996. Recommended for acceptance by S.K. Nayar.
For information on obtaining reprints of this article, please send e-mail to: transpami@computer.org, and reference IEEECS Log Number P96066.

0162-8828/96$05.00 ©1996 IEEE

this investigation, but could be implemented without too much additional effort.

Fig. 2 shows images of the first nine people in our database. Most of them are male Japanese nationals. The large number of 11,600 images makes the collection to one of the largest databases ever used to test a face recognition system.

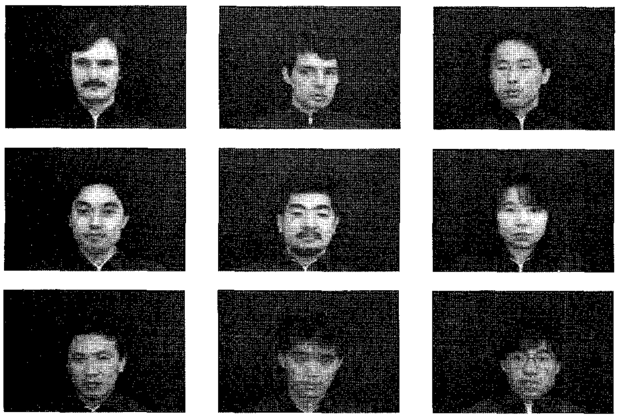

Fig. 2. Nine images from the 11,600-image database used throughout this research. The database contains a training set of 50 images and a test set of 50 images for each of the 116 people.

2.2 Feature Extraction Using Local Autocorrelation Coefficients

Local autocorrelation coefficients have the advantage of being shift invariant and computationally inexpensive. The feature extraction module of the face recognition system computes 25 local autocorrelation coefficients from a digital facial image, using the kernels represented in Fig. 3 [12], [13]. Each kernel is scanned over the entire image, and for each possible position, the product of the pixels marked in black is computed. All the products corresponding to a kernel are then summed so as to provide one coefficient. This operation is performed using the 25 different kernels, thus providing a 25-dimensional *feature vector*. By dividing the autocorrelation coefficient by powers of the zero-order coefficient, optional brightness invariance may be implemented. Such a low-dimensional face representation translates into small storage space requirements and fast classification times.

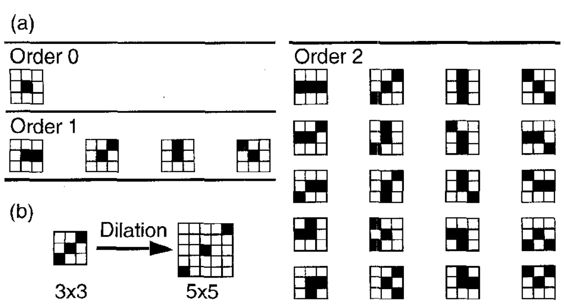

Fig. 3. Set of autocorrelation kernels used for feature extraction. (a) The kernels are defined on a neighborhood of 3×3 pixels. The *order* of a kernel is its number of active pixels (pixels marked in black) minus 1. (b) Example of scaling of an autocorrelation kernel with 3×3 pixels to a size of 5×5 pixels.

2.3 Classification Method

In order to classify the feature vectors extracted from the images, we employ linear discriminant analysis (LDA) [12], [13], [14].

The system is trained using feature vectors derived from the training set of images (see Section 2.1). The training set is comprised of K classes, each containing M images. In the following equations, $\mathbf{x}_m^k$ denotes the 25-dimensional feature vector representing the mth image of the kth class, $\overline{\mathbf{x}}^k$ denotes the mean vector of the kth class and $\overline{\mathbf{x}}$ denotes the mean vector of the whole dataset.

In LDA one determines the mapping

$$\mathbf{y}_m^k = \mathbf{A}^{\mathsf{T}}\mathbf{x}_m^k \tag{1}$$

which simultaneously maximizes the inter-class distance while minimizing the intraclass distance. Let us define the matrix sum

$$\Sigma_W = \sum_{k=1}^{K}\sum_{m=1}^{M}\left(\mathbf{x}_m^k - \overline{\mathbf{x}}^k\right)\left(\mathbf{x}_m^k - \overline{\mathbf{x}}^k\right)^{\mathsf{T}} \tag{2}$$

as the within-class covariance matrix and

$$\Sigma_B = \sum_{k=1}^{K}\left(\overline{\mathbf{x}}^k - \overline{\mathbf{x}}\right)\left(\overline{\mathbf{x}}^k - \overline{\mathbf{x}}\right)^{\mathsf{T}} \tag{3}$$

as the between-class covariance matrix. One way to find the desired mapping $\mathbf{A}$ is to maximize the quantity

$$tr\left(\hat{\Sigma}_W^{-1}\hat{\Sigma}_B\right), \tag{4}$$

where $\hat{\Sigma}_W$ and $\hat{\Sigma}_B$ are the within-class and between-class covariance matrices in the mapped space. The solution optimizing this criterion is the matrix $\mathbf{A}$ whose rows are the solutions $\mathbf{u}_n$ of the Eigen-equation

$$\Sigma_B\mathbf{u}_n = \lambda_n\Sigma_W\mathbf{u}_n. \tag{5}$$

In the mapped space, a Euclidean distance classifier with respect to the class means is used.

2.4 Rejection Criteria

The simplest way to add rejection ability is to set a threshold on the minimum Euclidean distance, which we denote by d_1, and to reject a vector if d_1 exceeds this threshold.

We found that setting a threshold on the confidence measure

$$c = \frac{d_2 - d_1}{d_2 + d_1} \tag{6}$$

reduces the false access rates by about one order of magnitude [15]. Here d_2 denotes the distance between the input vector and the second nearest class. In case of a perfect match, the input vector is identical to one of the class representatives, resulting in the distance $d_1 = 0$ and the maximum confidence value, unity. Vectors whose distances to the nearest and second nearest class are equivalent, $d_1 = d_2$, receive the minimum confidence value, zero (totally uncertain classification). The confidence values of other input vectors always lie between these two extrema. Rejection of input vectors takes place if the measure falls below a threshold. One of the benefits of using this rejection criterion is that the threshold value does not depend on the scaling of the input vectors. Other benefits will become apparent in Section 4, where the normalized confidence measure eases the fusion of data from several recognition modules.

3 System Performance

The experiments conducted in this section utilize the database described above. Training and testing of the system are always conducted on the separate, non-overlapping test and training sets.

3.1 Optimum Kernel Size

We conducted numerous recognition experiments using various image resolutions and kernel sizes [15]. Setting the kernel size to approximately half of the the face size consistently yielded the highest recognition rates. Thus quite coarse and low-frequent features seem to be of highest significance for the recognition algorithm. At a constant face to kernel size ratio, larger image resolutions result in slightly higher recognition rates. This might be attributable to the smaller effect of subpixel shifts in the case of the high-resolution images. For performance evaluation we use now fixed settings for image resolution (90×60 pixels) and kernel size (9×9 pixels, see also Fig. 3b).

3.2 Recognition Performance

Fig. 4 shows the recognition rates of the system in dependence of the number of persons that are stored. Linear discriminant analysis performs consistently at recognition rates around or above 98%. Earlier simulations using least square discriminant mapping resulted in a much lower recognition rate of 85% for 116 stored faces [15]. The recognition rate curve does not show a definite downward trend—a clear encouragement to conduct experiments with images of at least a few hundred persons. A recognition rate of 98% means that only one out the fifty faces stored for each person is not classified correctly—a surprisingly good result in view of the partly quite large head movements in the data base and the small number of features used for classification.

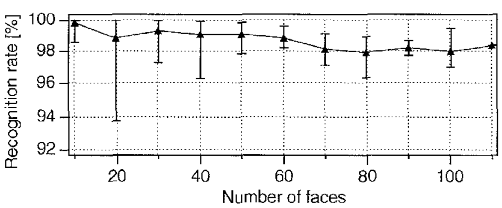

Fig. 4. Evolution of the recognition rate with the number of faces involved in the training process of linear discriminant analysis. For each number K of faces, the recognition rate was computed on 10 different subsets of dimension K from the database. The plotted points are the average values over these 10 results, and the upper and lower extrema of the error bars represent the maximum and minimum values, respectively.

Besides the recognition rate, the training and recognition times of an classifier are also important performance criteria, especially in real-time applications. Due the relatively low dimensionality of the feature vectors (25 elements), the resolution of the Eigen-equation (5) is quasiinstantaneous. The bulk of the training time is spent with the reading of files and the computation of autocorrelation matrices: around six seconds for 5,800 feature vectors of 116 faces on a 15 Mflops RISC workstation.

The recognition time is around seven milliseconds per feature vector. The factor limiting the speed of our current real time system is consequently not the classifier, but rather either the frame grabber, which delivers about three images per second, or the feature extraction process, whose speed depends on the image size. In real time experiments, the system proofed to be quite robust against changes in head position, removal of spectacles, reflexes on spectacles and small changes in scale.

3.3 Rejection Performance

In complete absence of a rejection mechanism, all images presented to the recognition system, including images of unknown persons and background, are mapped to the closest known face. Separating faces from background and other clearly different input is not too difficult to accomplish. Reliably recognizing known persons while rejecting unknown persons, however, we found to be a much more challenging task. Rejecting unknown faces means that the system has not only to accept wide variations in facial expression, head rotation, and so on, but also to reject patterns which lie quite close in the pattern space.

For the evaluation of the rejection performance of the system, we divided the training set into two parts containing 60 *known* faces and 56 *unknown* faces. Throughout this section, we use an image resolution of 60×40 pixels and a kernel size of 5×5 pixels, which results in close-to-optimal peak recognition rates. We train the classifier on 3,000 images of the 60 known faces from the training set. The recognition and rejection performance of the system is then tested on all the 5,600 images of the test set.

Fig. 5 plots the recognition rate versus the false access rate for two different rejection criteria. Here, recognition rate is defined as the proportion of known faces that are correctly recognized, and the false access rate is the proportion of unknown faces that are not rejected. An ideal face recognition system would have a recognition rate of 100% and a false access rate of 0%, as represented by the lower right corners of the graphs in Fig. 5. For direct comparison of the performances of the two rejection criteria, we measure the false access rate of the systems at a fixed recognition rate of 90%. Simple thresholding of the minimum distance d_1 results in a false access rate of 33%. Setting a threshold with regard to the confidence measure from (6) performs eight times better: The false access rate drops to only 4%.

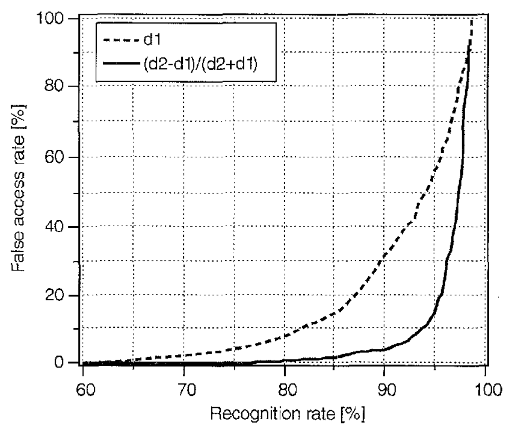

Fig. 5. Linear discriminant analysis: comparison of two rejection criteria.

4 MULTIPLE RESOLUTION ARCHITECTURE

The single-resolution system in the preceding section show a satisfactory performance in terms of peak recognition rate and rejection capability. In this section, we show that multiresolution integration improves the overall reliability of the system up to four times.

4.1 Modular Architecture

Integrating the feature vectors generated using different image resolutions can be done at several levels. The most straightforward method is to merge two or more feature vectors into a single vector with a larger number of elements. A classifier is then trained and tested on these larger vectors. This approach has the drawback of being computationally quite expensive, when the number of feature extraction modules n is large. The computation of the within-class covariance matrix Σ_W, for example, requires $O(n^2)$ operations. In this paper we employ a *modular architecture* which uses n classifiers similar to those studied in the previous section. Thus, the computation of the within-class covariance matrices requires only $O(n)$ operations. This reduction of the computational effort for training comes at the comparatively small cost of requiring an additional information fusion step. Fig. 6 illustrates the architecture of the complete system. Each classifier processes a singe feature vector, and it outputs a *decision* and a *matching score*, which is the value the confidence measures defined in (6). The classification decisions and matching scores of these modules are then combined so as to produce the final decision.

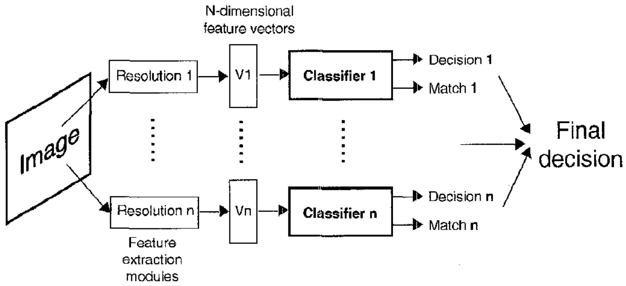

Fig. 6. Architecture of multiresolution system.

4.2 Data Integration Methods

We split the process of integrating the information from different classifying modules into a classification decision and a rejection decision. For the *classification decision* we use the decision of the module which gives the highest matching score. Counting the "votes" of the different modules gives similar recognition rates [15]. We perform the *rejection decision* according to the mean value of the matching scores provided by the different modules. The smoothing effect of averaging proved to be more beneficial to the accuracy of the rejection decision than the straightforward selection of the best match [15]. Recent simulations using Choquet integrals for information integration performed slightly better than simple averaging [16], [17].

4.3 Experimental Results

The purpose of using several modules is to fully utilize the information provided at different image resolutions. From here on we use a kernel size of 9×9 pixels, and images scaled down one to five times (denoted resolution 1–5). The rejection curves in Fig. 7 show that the system performance increases rapidly with the number of modules. At a recognition rate of 95%, five modules together achieve a false access rate of 1.5%, whereas the best module alone achieves only 15%.

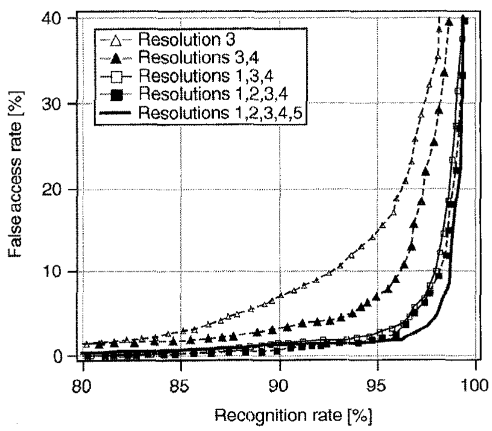

Fig. 7. Rejection curves of multiple resolution systems.

For comparing the overall reliability at simultaneously high recognition and low false access rates, let us define a *system error* ε_{sys}:

$$\varepsilon_{sys} = \frac{1}{\sqrt{2}} \min_{\lambda} \sqrt{\left[1 - rec(\lambda)\right]^2 + far(\lambda)^2}, \tag{7}$$

where λ denotes the rejection threshold. This system error is proportional to the minimum distance between a rejection curve and the lower right corner of the graph in Fig. 7. A system error of, for example, 5% roughly corresponds to a false access rate of 5% and a recognition rate of 100% – 5% = 95%.

Usage of the best module alone in a simple system operating at only one resolution results in a system error of 8.5%. A multiresolution system with five modules achieves a system error of 2.74%, a three-fold increase in performance. For larger numbers of modules, redundancies in the feature vectors become apparent, and the improvement levels off. A 15-module system achieves only a four-fold increase in performance. In addition, the training and recognition times increase linearly with the number of modules. Thus the system using five modules may be considered as a good tradeoff between rejection performance and computation time.

Considering the large increase in performance, the increase in computational expense is moderate. In a single-resolution system, feature extraction typically takes about 48 ms (using a 15 Mflops RISC workstation). Training on 50×116 feature vectors takes about 6 s, and the classification time per feature vector is 7 ms. In our multiresolution system with five modules, feature extraction takes about 630 ms. Dropping the module which processes the largest image resolution would reduce this time to 200 ms. Training and classification times scale linearly with the number of modules. Processing times may be reduced further by taking advantage of the parallelism of the method and implementing it using special hardware [18], [19], [20], [21], [22], [23].

Finally, it has to be noted that recognition in the five-resolution system is at least ten times more reliable than recognition in the single-resolution system, as indicated by peak recognition rates around 99.9% as opposed to 98% (cf. Fig. 4).

5 Discussion and Conclusion

We have investigated the performance of a technique for face recognition based on the computation of local autocorrelation coefficients. Using a database of 11,600 frontal facial images of 116 persons, we conducted a thorough statistical evaluation of the approach. A simple, yet effective rejection criterion decreased the false access rates by about one order of magnitude. The introduction of a computationally efficient modular multiresolution architecture allowed us to reduce the false access rate even further, from 15% to 1.5%, while achieving a recognition rate of 95%. The system error ε_{sys} dropped from 8.5% to 2.74%.

Large variations in the databases make the direct comparison of face recognition systems difficult. Databases might include variation of light direction, significantly smaller or larger numbers of persons, or smaller or larger numbers of samples per persons. Subject to these constraint, our system showed a comparatively good performance. Bichsel and Pentland report a peak recognition rate of 90% on a database with 8000 images of 3,000 people using a Eigen-face technique [22], [4]. Using a novel variant of template matching, Bichsel achieves a false access of 50% at a recognition rate of 90% using a database containing 400 images of 70 persons [10]. Brunelli and Poggio report a peak recognition rate of approximately 98% on a set of 188 images of 47 persons, using a interesting method based on template matching [3].

In general, we found rejection of unknown faces to be a much harder problem than sole recognition of faces known to the system. Future studies could use the system error ε_{sys}, which incorporates both recognition performance and rejection capability, and a standardized database for comparing different recognition methods.

The short classification times suggest to incorporate additional invariances by means of introducing a sufficient number of classes per person, each representing, for example, a typical lighting direction.

Information integration from multiple sources is essential for reliably performing difficult classification tasks. With our multiresolution system, we have taken one step into this direction. The next step will be the implementation of a segmentation module based on template matching, which will allow robust image segmentation and also give additional clues for classification.

Acknowledgments

This research was partly supported by a research grant from the Real World Computing Partnership (RWCP) by the Ministry of International Trade and Industry.

References

[1] A. Samal and P.A. Iyengar, "Automatic Recognition and Analysis of Human Faces and Facial Expressions: A Survey," *Pattern Recognition*, vol. 25, pp. 65–67, 1992.

[2] D. Valentin, H. Abdi, A.J. O'Toole and G.W. Cottrell, "Connectionist Models of Face Processing: A Survey," *Pattern Recognition*, vol. 27, pp. 1,209–1,230, 1994.

[3] R. Brunelli and T. Poggio, "Face Recognition: Features versus Templates," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 15, no. 10, pp. 1,042–1,052, Oct. 1993.

[4] M. Turk and A. Pentland, "Eigenfaces for Recognition," *J. Cognitive Neuroscience*, vol. 3, no. 1, pp. 71–86, 1991.

[5] A.L. Yuille, "Deformable Templates for Face Recognition," *J. Cognitive Neuroscience*, vol. 3, no. 1, pp. 59–70, 1991.

[6] M. Lades, C.C. Vorbruggen, J. Buhmann, J. Lange, C. v. d. Malsburg, R.P. Wurtz and W. Konen, "Distortion Invariant Object Recognition in the Dynamic Link Architecture," *IEEE Trans. Computers*, vol. 42, no. 3, pp. 300–311, Mar. 1993.

[7] G.W. Cottrel and M. Fleming, "Categorization of Faces Using Unsupervised Feature Extraction," *Proc. Int'l Neural Networks Conf.*, pp. 65–70, Paris, 1990.

[8] J.L. Perry and J.M. Carney, "Human Face Recognition Using a Multilayer Perceptron," *IJCNN Washington D.C.*, pp. 413-416, 1990.

[9] H. Bouattour, F. Fogelman-Soulié, and E. Viennet, "Solving the Human Face Recognition Task Using Neural Nets," *Proc. Artificial Neural Networks*, I. Aleksander and J. Taylor, eds., pp. 1,595–1,598. Amsterdam: North-Holland, 1992.

[10] M. Bichsel, "Strategies of Robust Object Recognition for Automatic Identification of Human Faces," PhD Thesis No. 9467, Eidgenössischen Technischen Hochschule Zürich, 1991.

[11] F. Goudail, E. Lange, T. Iwamoto, K. Kyuma, and N. Otsu, "Fast Face Recognition Method using High Order Autocorrelations," *Proc. IJCNN Int'l Conf. Neural Networks*, Nagoya, Japan, pp. 1,297–1,300, 1993.

[12] N. Otsu and T. Kurita, "A New Scheme for Practical, Flexible and Intelligent Vision Systems," *Proc. IAPR Workshop on Computer Vision—Special Hardware and Industrial Applications*, pp. 431–435, Tokyo, 1988.

[13] T. Kurita, N. Otsu, and T. Sato, "A Face Recognition Method Using Higher Order Local Autocorrelation and Multivariate Analysis," *Proc. 11th IAPR Int'l Conf. Pattern Recognition*, pp. 213–216, 1992.

[14] K. Fukunaga, *Introduction to Statistical Pattern Recognition*. Academic Press, 1990.

[15] F. Goudail, "Large Scale Face Recognition Using Local Autocorrelations," Rapport de stage de fin d'études, École Supérieure d'Optique, Amagasaki, Japan, 1993.

[16] G. Choquet, "Theory of Capacities," *Annales de l'Institute Fourier.*, vol. 5, pp. 131–295, 1953.

[17] T.D. Arbuckle, E. Lange, T. Iwamoto, N. Otsu, and K. Kyuma, "Fuzzy Information Fusion in a Face Recognition System," *Int'l J. Uncertainty, Fuzziness, and Knowledge-Based Systems*, submitted.

[18] D. Hammerstrom, "A VLSI Architecture for High-Performance, Low-Cost, On-Chip Learning," *Proc. Int'l Joint Conf. Neural Networks*, pp. 537-543, San Diego, 1990.

[19] U. Ramacher, J. Beichter, W. Raab, J. Anlauf, N. Bruls, M. Hachmann, and M. Wesseling, "Design of a First Generation Neurocomputer," *VLSI Design of Neural Networks*, Kluwer Academic, 1991.

[20] Y. Kondo, Y. Koshiba, Y. Arima, M. Murasaki, T. Yamada, H. Amishiro, H. Shinohara, and H. Mori, "1.2 GFLOPS Neural Network Chip Exhibiting Fast Convergence," *Proc. Int'l Solid-State Circuits Conf.*, pp. 218–219, San Francisco, 1994.

[21] E. Lange, Y. Nitta, and K. Kyuma, "Optical Neural Chips," *IEEE Micro*, vol. 14, no. 6, pp. 29–41, 1994.

[22] E. Lange, T. Arbuckle, F. Goudail, T. Iwamoto, K. Kyuma, and N. Otsu, "Towards a Vision Processor in Neural Architecture," *Proc. Int'l Conf. Artificial Neural Networks*, pp. 287–295, Paris, Oct. 1995.

[23] E. Lange, Y. Nitta, and K. Kyuma, "Direct Image Processing Using Artificial Retina Chips," *Trends in Optics: Research Developments and Applications*, A. Consortini, ed., pp. 63–82. Academic Press, 1996.

[24] M. Bichsel and A. Pentland, "Human Face Recognition and the Face Image Set's Topology," *CVGIP: Image Understanding*, vol. 59, no. 2, pp. 254–261, 1994.

[25] Y. Kaya and K. Kobayashi, "A Basic Study on Human Face Recognition," *Frontiers of Pattern Recognition*, S. Watanabe, ed., 1972.