MVA2005 IAPR Conference on Machine VIsion Applications, May 16-18, 2005 Tsukuba Science City, Japan

12-1

Towards Flexible and Intelligent Vision Systems – From Thresholding to CHLAC –

Nobuyuki Otsu
AIST / University of Tokyo
† National Institute of Advanced Industrial Science and Technology
Umezono 1-1-1, Tsukuba-shi, Ibaraki-ken, 305-8568 Japan
Email: otsu.n@aist.go.jp

Abstract

This paper presents a general approach and research results towards flexible and intelligent vision systems which the author has been interested in and devoted to for a quarter century. First, such a general approach is emphasised that is based on the general framework of pattern recognition consisting of two stages of feature extraction; invariant feature extraction as the first geometrical aspect, and discriminant feature extraction as the second statistical aspect. Along the line of the general approach, several methods the author has developed so far are introduced, such as multivariate analysis approach to automatic threshold selection, Higher-order Local Auto-Correlation (HLAC) feature extraction, and face recognition. Finally, some recent researches using expanded HLAC features for motion recognition are illustrated.

1 Introduction

"Vision" or visual information processing occupies up to 80 percent of the sensory information processing of human (and animals also), and it forms the base of their intelligence in the real world. Thus, vision has been an important topic in the research fields of pattern recognition, artificial intelligence, and also cognitive science. So many researches have been done in this nearly half century, however such a vision system that is flexible and intelligent enough like human and animals is still far on the way.

There are increasing needs for computer vision in various fields of industrial production, material and medical sciences, and in recent years particularly in the field of video surveillance for security purposes. In those application fields, such systems are expected that are as convenient (hopefully, personal computer based), practical (real-time high speed) and adaptive (trainable for various purposes) as possible.

The remarkable development of computer technology is providing more powerful computational environment. However, such requirements are still difficult to be satisfied, which seem to reveal a discrepancy of the conventional approach comprising the steps of image processing techniques.

In view of the situation, it seems that now is the time to reconsider the problem of computer vision and seek a new paradigm; beyond the traditional paradigm which is influenced by the serial and procedural processing by computers, and toward a more general and flexible vision, with also taking into account parallel distributed processing which is typical in the error back-propagation learning in multilayer neural networks.

The author has been engaged in the rather theoretical research of pattern recognition, viewing vision as a typical example of its application, and has been interested in and devoted to a new scheme of flexible and intelligent vision systems.

In this paper, firstly, we start with considering the general framework of visual information processing and review the traditional approach (serial and procedural processing) and also the neural network approach (parallel and distributed processing), pointing out the problems and drawbacks inherent in those approaches.

Secondly, as a theoretical foundation, the general framework of pattern recognition and feature extraction is reconsidered, with showing the importance of two stages of feature extraction: *invariant* feature extraction as a geometrical aspect and *discriminant* feature extraction as a statistical aspect. Theoretical analysis of the latter statistical feature extraction in general nonlinear case reveals the framework of Bayesian estimation that underlies the supervised learning in neural networks and also multivariate data analysis methods.

Thirdly, we present a scheme for flexible visual information processing and recognition which we have developed to put our theoretical standpoint into practice as the simplest model. The first geometrical feature extraction is based on higher order local autocorrelations so as to satisfy shift-invariance and additivity that are required and preferable as the fundamental conditions for vision systems. The second statistical feature extraction is based on multivariate data analysis, which linearly combines so obtained the primitive features in the first stage into effective new feature(s) for a given task. The system is very simple but can adaptively and quickly learn a given task from training examples and shows a considerably good performance.

Finally, some recent developments and several application examples of the scheme are shown.

2 Visual Information Processing

Let us start with considering visual information processing in a general framework, regarding it as a mapping “ : Y = “ (X) that converts input X to output Y. For example, when considering X as an input image, the output Y is an image in the cases of so-called image processing such as image enhancement or restoration. Y is a numerical value (or set of values) in the case of image measurement, and Y is a symbolic label (name) of category or its representation in the case of image recognition. For image understanding, the output Y will be more complicated description.

In this way, problems of visual information processing can be regarded as problems of how to construct the implicit mapping “ . In practice, it is not easy to construct the mapping directly as a whole. Therefore, it is usual to reduce the procedure into simpler procedures in sequence and/or in parallel (viz, shift-invariant filters). Thus the input X in the general framework is not necessarily restricted to a whole input image but can be a sub-image in parallel procedure, and both input and output can be intermediate information representation in sequential procedure. It should be remarked here that the mapping “ can treat various forms of input and output at each level in the hierarchy of processing or cognition, and the point is that it is generally abstracting redundant input information to more efficient and useful output information. This is also valid for other cases of pattern recognition and information processing in general as well.

The essential problems posed from this general framework are the following two. One is a problem of information representation. Basically the following three kinds of representation are considered.

R1: *continuous and distributed representation* of pattern information at signal level (function)

R2: *discrete and localized representation* of symbolic information at conceptual level (symbol)

R3: *compressed and summarized representation* of feature information at intermediate level which bridges the two levels in the above (vector)

In visual information processing and pattern recognition in general as well, it is important that these different types of information representation are smoothly integrated. In particular feature representation R3 is a central issue.

The other problem is posed on approach to information processing, that is how to construct the implicit mapping “ in an optimal way. Regarding this, the following three schemes (methods) are conceivable (Cf. Fig.1).

M1: *direct procedural method*

M2: *forward adaptation method*

M3: *backward adaptation method*

M1 is the most direct and ordinary method for the construction of processing (mapping Y = “ (X)), and supposed procedures are directly given in a sequence. On the other hand, M2 and M3 are indirect methods for adaptively

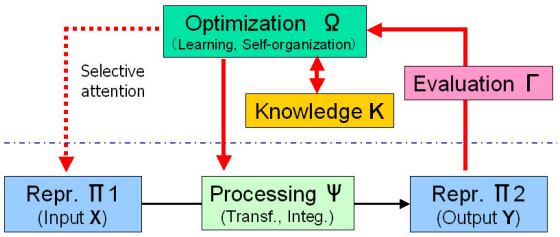


Figure 1: General scheme of information processing

obtaining the optimum processing in a framework of feedback by evaluation and optimization.

In M2 a family of mapping “ (parameterized model) is considered, and the optimum mapping is adaptively obtained (approximated) by optimizing the parameters by using training samples of input-output pairs. This is a supervised learning method, and comparison for evaluation is done between output Y and the ideal one that is given by a teacher. Back propagation learning in feed-forward neural networks and methods of multivariate analysis are the cases.

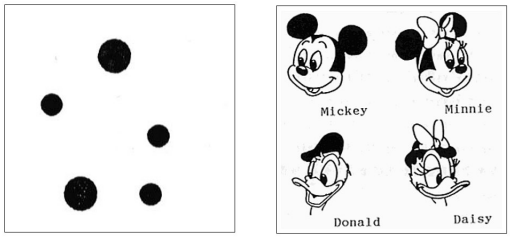
In M3 the mapping (processing) “ is not obtained explicitly, but equivalently the processed result (output) Y is obtained explicitly. Namely, a parameterized model of output Y is assumed and considered as an interpretation of input X, and the optimum model is adaptively obtained (approximated) by comparing and evaluating Y with X. For example, a model fitting such as line or surface fitting to image data is the case.

It is noted that M2 is regarded as forward inference of processing (“) while M3 as backward inference of interpretation (“ $^{-1}$).

3 Review of Ordinary Approach

Keeping this in mind, we shall review the ordinary approach in what follows.

To make the point clear, let us consider two simple examples of visual information processing for binary images in Fig. 2. One (Fig. 2-a) is a case of image measurement, where a number of round particles in two different sizes (large and small) are represented in silhouette without overlapping. The task requested here is to count the number of each kind of particles as quickly as possible. On the other hand, Fig. 2-b is a case of image recognition, where four kinds of cartoon faces are requested to be recognized.



a) Image measurement b) Image recognition

Figure 2: Examples of visual information processing

3.1 Serial and procedural approach

Our usual approach to these are the following serial and procedural approach. For the case of Fig. 2-a, it comprises to segment out each object one at a time by scanning, to classify it into two sizes by measuring, for example, the diameter, and to sum up the counts. Obviously, the computation time in this case grows proportionally to the numbers of objects (particles) in the image. For the case of Fig. 2-b, we usually consider what are the distinctive features for the categories (faces). Ear shape is different, and having a ribbon or not is also distinctive. Then the recognition procedure is reduced to how to recognize ears and ribbons. However, the subproblems are still pattern recognition problems in a hierarchy.

Such a serial and procedural approach corresponds to the direct construction method M1, where the given task is analyzed and decomposed into a sequence of subtasks (processing techniques) that are judged necessary to achieve the specified task. This seems quite natural at a glance, however deeper consideration will reveal the following problems inherent to such serial and procedural methods.

1. Applications are limited to such problems that have clear algorithms for what and how to compute.
2. The sequence of processing techniques is specified for the given task and therefore tends to lack adaptability for general-purpose application.
3. Minor errors at each step are accumulated in the end, resulting in fragile processing.
4. Even if each step of procedure (processing technique) is optimized, it does not necessarily guarantee the total optimality of the sequence.
5. Real-time high speed processing is difficult due to the successive complicated computations.

Originally, serial and procedural scheme is an abstraction of the aspect of our way of logical thinking. Actually, the scheme was adopted as a principle of information processing by modern computers and became a dominant paradigm of information processing in general. And, we are too much used to the scheme. The scheme basically assumes the complete and deterministic world of symbols and logic, excluding uncertainty and ambiguity. Therefore it is effective and efficient for dealing with well-defined logical thinking at a higher level but too strict and *hard* for dealing with intuitive cognition at a lower level.

3.2 Parallel and adaptive approach

There is a major trend to reconsider parallel information processing in accordance with the development of parallel computing facilities and the progress of brain science and cognitive science. Of course, such an idea that parallel processing is natural for images as distributed information is not new. There have been developed various methods for parallel processing, for example relaxation methods.

Neural computing, however, differs in that it tries to approach *flexible* information processing such as cognition at an intuitive or sub-symbolic level, referring to the parallel and distributed processing (PDP) in the brain. There are two major types of neural networks (NN). One is the type of feed-forward NN which is typical in multilayer analog perceptron with back propagation learning [23]. The other is the type of mutually connected NN which are typical in Hopfield networks [6] and Boltzmann machine [4].

In particular, the former type is widely applied as a new paradigm for flexible pattern recognition and control with *learning* capability. This type corresponds to the forward adaptation method M2, that is, the adaptive construction of the correspondence relation (mapping ") from learning samples of input and output pairs. Surely the processing speed after learning is fast because of parallel computation, however the learning process is slow in convergence and sometimes trapped into local minima.

Although such a neural computing is significant as a model of neurological information processing in the brain, the characteristic of processing elements, viz. sigmoidal nonlinearity and bounded values between 0 and 1 of input and output, makes the mathematical analysis difficult and is not necessarily important in practical application. It would be rather limitative and inefficient. For example, we could directly apply the type of NN to the image recognition problem shown in Fig. 2-b, but feeding an image itself directly to the input layer is computationally too heavy in learning mode. For the case of image measurement shown in Fig. 2-a, obviously we cannot directly employ the bounded output values of the output layer of the networks. In fact, the problem of information representation previously stated is not clear there.

On the other hand, the latter type of mutually connected NN is applied to combinatorial optimization problems as a new paradigm of parallel information processing. For visual information processing, those are applied as a method of stochastic relaxation for image restoration [2] or as a regularization (constrained optimization) method to solve various ill-posed problems [22], although they are still at the level of early vision. These are corresponding to the backward adaptation method M3.

In particular, the paper [2] points out that in some condition local relationship (MRF: conditional probabilistic structure) can be equivalently corresponded to global characteristic (potential) via the Gibbs transformation (distribution) and that a kind of optimization problem can be reduced to a problem of stochastic inference (Bayesian or maximum likelihood estimation) and solved by an iterative method in thermodynamical analogy (stochastic relaxation plus annealing).

4 Recognition of Pattern Recognition

In order to consider a new direction of computer vision, it seems necessary and important to look back again and reconsider the general and basic framework of pattern recognition. Pattern recognition is the front of intelligence in the real world, associating continuous distributed information representation (patterns $(\boldsymbol{r})$) with discrete localized information representation (symbols, or categories C_j), and vision is also included in the framework as a typical case.

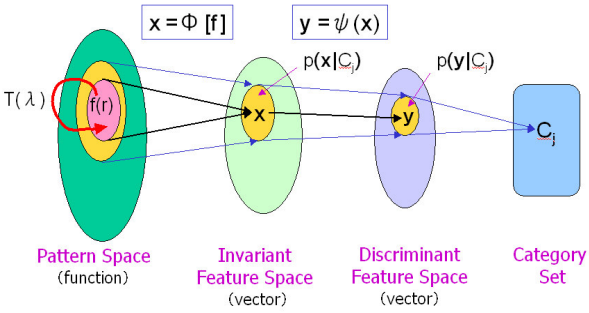


Figure 3: General framework of pattern recognition

4.1 General framework of pattern recognition

The essence of pattern recognition is to summarize the various objects (patterns) in the physical world into classes (categories) and to recognize each object in connection with the category name. The general framework of pattern recognition is schematically illustrated in Fig. 3.

Each physical pattern distributes spatially (viz. characters) or temporally (viz. speech sounds) and is generally represented by a function . In practice, it is approximately expressed as a high but finite dimensional vector in the observation, e.g. by sampling. Such patterns constitute a high dimensional continuous topological space called a *pattern space* (R1), which is a faithful representation space and not necessarily convenient for recognition or cognition. Actually it is quite redundant.

Therefore, such redundant patterns are usually subjected to extraction of features (information) which are essentially efficient for recognition. This process, called *feature extraction*, provides an efficient representation space called a *feature space* (R3) with less dimensionality, where each pattern is represented by a *feature vector* $\boldsymbol{y}$. In the feature extraction as information compression and dimensionality reduction, an important point is how to evaluate and select such features that enhance the class separability and clustering.

On the other hand, categories are represented by symbols $C_j(\ = 1 \quad K)$ and form a discrete and finite set (R2). Thus, recognition can be regarded as a discontinuous mapping from the feature space to the category set, resulting in partitioning and quantizing the feature space into finite number of subregions each of which consists of a cluster and is associated with an identical category name.

4.2 Bayesian decision

Patterns in the real world inevitably contain ambiguity and uncertainty. Therefore probability and statistical theory has been playing a central role from the early stage in the theory of pattern recognition [1].

When the feature vector of a pattern is given by $\boldsymbol{y}$ and its statistical structure is known, the problem of classification is completely formulated in the framework of statistical decision theory. Namely, Bayesian decision which decides an input pattern as belonging to the class C_j that has maximum *a posteriori* probability

$$P(C_j\ \boldsymbol{y}) = P(C_j)p(\boldsymbol{y}\ C_j)/p(\boldsymbol{y}) \qquad (1)$$

yields the optimum classification in the sense of minimum error rate, where $P(C_j)$, $p(\boldsymbol{y}\ C_j)$, $p(\boldsymbol{y}) = \sum_{j=1}^{K} P(C_j)p(\boldsymbol{y}\ C_j)$ are *a priori* probabilities, conditional probability density functions, and total probability density function, respectively.

Then the minimum error rate attained, of course, depends on the feature vector employed. Therefore the essential problem of pattern recognition is reduced to the problem of feature extraction, viz. how to extract $\boldsymbol{y}$ that is efficient for recognition.

4.3 Feature extraction theory

Feature extraction must satisfy opposing requirements: to extract essential information for recognition while to discard irrelevant information, and to identify (unify) the patterns within each class while to distinguish (separate) the patterns between different classes. In order to satisfy those requirements, the process of feature extraction is divided into two stages: *invariant* feature extraction as a *geometrical* aspect and *discriminant* feature extraction as a *statistical* aspect. Those are illustrated in Fig. 3, and each aspect has been theoretically studied in a general framework [16] as shown briefly in the followings.

4.3.1 Invariant feature extraction

Observed images are generally subjected to various geometrical transformations such as translation and dilatation due to the relative position and movement of observer to the objects. Nevertheless, we can recognize what the objects are and how those are transformed. Such a transformation is called the *invariant transformation* in the sense that it does not change and keeps *invariant* the correspondence to categories. Therefore, for the base of visual recognition it is important to investigate theoretically what feature values should be extracted from the observed image in order to make such recognition possible.

In general, let a pattern be denoted by a function , a feature value by , feature extraction by a functional $= [\ ]$, and an invariant transformation by an operator $T(\lambda)$. Then the feature that indicates the category should be invariant to the invariant transformation $T(\lambda)$. That is stated as

$$[T(\lambda)\ ] \qquad [\ ] = 0 \qquad (2)$$

In order to extract invariant features, preprocessing called normalization is usually employed for the transformed pattern $T(\lambda)$. However a more systematic theory has been developed for invariant feature extraction [16],[19]. The theory is based on Lie group theory and operator analysis. By considering infinitesimal operator, a necessary and sufficient condition for invariant feature $[\ ]$ in (2) is given by a partial differential equation. As its elementary solutions, a set of linear or nonlinear invariant features $_i$ is obtained and it forms a feature vector $\boldsymbol{x}$.

It is noticed that extraction of invariant features leaves variant features which contain the information on the geometrical transformation. Thus the theory of invariant feature extraction can also provide the theory of extracting a

feature $'[\ ]$ which describes the transformation (λ) of the pattern, that is

$$\lambda = \ '[T(\lambda)\ ] \qquad '[\ ] \tag{3}$$

Thereby a theoretical foundation for the recognition of shape and transformation is given [19]. It is interesting to note that invariant features and variant features form mutually orthogonal manifolds.

This aspect of invariant/variant feature extraction is important as a principle of geometrical aspect of pattern recognition and cognition as well, in particular in vision research. It also provides a guide to the direct construction method M1. Seemingly, this aspect receives less attention in the research on PDP and NN.

4.3.2 Discriminant feature extraction

Through the invariant feature extraction the patterns belonging to an identical class (category), or patterns mutually related by invariant transformation $T(\lambda)$, are in principle captured as an identical point $\boldsymbol{x} = \ [\ ]$ in the invariant feature space X. However actual patterns are subjected to irregular variation and contaminated with noise, and therefore the patterns belonging to a class are captured as a statistical distribution around the ideal point in X.

Hence, the general framework of statistical feature extraction as the second stage is given by a mapping $\text{“}\ ; \boldsymbol{y} = \text{“}\ (\boldsymbol{x})$, from the invariant feature space $X \subseteq R^m$ to a discriminant feature space $Y \subseteq R^n$ ( ), assuming multi-class structure of probabilities and statistical distributions $= P(C_j)\ p(\boldsymbol{x}\ C_j)\ _{j=1}^{K}$ in X, and a criterion J for evaluating the goodness of “ in Y. And the problem is to obtain the optimum mapping “ under the criterion J. This is corresponding to the forward adaptation method M2 in Section 2.

As for the simple and practical methods of statistical feature extraction, there are several nonparametric methods of multivariate data analysis, such as RA (regression analysis) and DA (discriminant analysis). However, those are usually formulated as linear mappings, and extracted statistical knowledge are only up to the second order statistics; viz. means and covariances. Therefore it is not clear how those reflect the underlying probabilistic structure and relate to classification. In order to clarify the underlying essential structure and meanings, it is necessary to remove the restriction of linear mapping and release it to a general nonlinear mapping.

Nonlinear discriminant analysis: Let W_Y and B_Y be the within-class and the between-class covariance matrices of $\boldsymbol{y}$, respectively. Then the discriminant feature extraction is formulated as an optimum mapping “ that maximizes the following discriminant criterion.

$$J[\text{“}\ ] = \mathrm{tr}(W_Y^{-1} B_Y) \tag{4}$$

When “ is linear, the coefficient matrix A of the optimum linear mapping $\text{“}\ _L(\boldsymbol{x}) = A'\boldsymbol{x}$ (where symbol ' denotes the transpose) is given by the eigenvectors of the matrix $W_X^{-1} B_X$ of $\boldsymbol{x}$. This is the result of the ordinary discriminant analysis.

On the other hand, the optimum nonlinear discriminant feature extraction $\text{“}\ _N$ is obtained by the variational calculus as the following simple form [14]:

$$\boldsymbol{y} = \text{“}\ _N(\boldsymbol{x}) = \sum_{j=1}^{K} P(C_j\ \boldsymbol{x})\boldsymbol{c}_j \tag{5}$$

and closely relates to the Bayes *a posteriori* probabilities and therefore to the Bayesian decision. The optimum discriminant feature space Y is then a $K \quad 1$ dimensional simplex the vertexes of which are given by $\boldsymbol{c}_j$, and obviously the Bayes decision boundaries are given by the barycentric subdivision. The vectors $\boldsymbol{c}_j$, which are called the class-representative vectors, are obtained from the eigenvectors of the following K by K stochastic matrix S which summarizes the between-class probabilistic relations.

$$S = [s_{ij}] \qquad s_{ij} = \int P(C_j\ \boldsymbol{x}) p(\boldsymbol{x}\ C_i)\ \ \boldsymbol{x} \tag{6}$$

It is noted that s_{ij} can be rewritten as follows by using the Bayes formula in (1).

$$s_{ij} = \ _{ij}/P(C_i) \tag{7}$$

where

$$= [\ _{ij}] \qquad _{ij} = \int P(C_i\ \boldsymbol{x}) P(C_j\ \boldsymbol{x}) p(\boldsymbol{x})\ \ \boldsymbol{x} \tag{8}$$

Least-mean-square nonlinear discriminant mapping: Suppose that $\boldsymbol{e}_j$ as the representative vectors of each class C_j are given and fixed in Y and vectors $\boldsymbol{x}$ belonging to C_j are mapped so as to concentrate around $\boldsymbol{e}_j$, respectively. Then the class separability (discrimination) can be evaluated by the following mean square error as a functional of mapping “.

$$\varepsilon^2[\text{“}\ ] = \sum_{j=1}^{K} P(C_j) \int \| \text{“}\ (\boldsymbol{x}) \quad \boldsymbol{e}_j \|^2\ p(\boldsymbol{x}\ C_j)\ \ \boldsymbol{x} \tag{9}$$

Then the least-mean-square discriminant mapping is formulated so as to minimize the criterion. From the standpoint of multivariate data analysis, this can be viewed as regressing the class representative vectors $\boldsymbol{e}_j$ by the feature vector $\boldsymbol{x}$.

It should be also noticed that this is just the same criterion as is used in the back propagation learning of multilayer feed-forward neural networks, where $\boldsymbol{e}_j$ are given as desired ideal output for class C_j respectively and usually taken to be orthonormal base vectors:

$$\boldsymbol{e}_I' \boldsymbol{e}_j = \delta_{ij} \tag{10}$$

The optimum nonlinear mapping $\text{“}\ _N$ is obtained by the variational calculus as follows [16],[17].

$$\boldsymbol{y} = \text{“}\ _N(\boldsymbol{x}) = \sum_{j=1}^{K} P(C_j\ \boldsymbol{x})\boldsymbol{e}_j \tag{11}$$

The least-mean-square error attained is given by

$$\varepsilon^2[\text{“}\ _N] = 1 \quad \mathrm{tr} \tag{12}$$

The solution turns out to be the same form as the solution of the nonlinear discriminant analysis in Eq.(5). It is remarked that $\boldsymbol{e}_j$ are given and fixed in this case and farther optimization with respect to the configuration $\boldsymbol{e}_j$ results in the nonlinear discriminant analysis.

On the other hand, the optimum linear solution (multiple regression analysis) is given by the following form

$$\boldsymbol{y} = \text{“}_L(\boldsymbol{x}) = \sum_{j=1}^{K} L(C_j\ \boldsymbol{x})\boldsymbol{e}_j \tag{13}$$

and, to be interesting, takes a similar form to the nonlinear solution in (11). Actually, the term $L(C_j\ \boldsymbol{x})$ is the linear approximation of the Bayes *a posteriori* probability $P(C_j\ \boldsymbol{x})$ and explicitly given by

$$L(C_j\ \boldsymbol{x}) = P(C_j)\ (\ _j\ \ \ )'\ \ ^{-1}_X(\boldsymbol{x}\ \ \ \ ) + 1 \tag{14}$$

where $_j$, , and $_X$ are the class mean, the total mean, and the total covariance matrix of $\boldsymbol{x}$, respectively.

4.4 Some commentary discussion

As has been shown in the above, the intrinsic structure of Bayesian estimation that underlies the statistical discriminant feature extraction is clarified by expanding the mapping “ to general nonlinear one and solving directly the ultimate optimum nonlinear mapping by using the variational calculus. It is important to remark here that discrimination (separation) of pattern classes results in Bayesian estimation in its ultimate nonlinear case, which also reveals the close relationship between statistical feature extraction and Bayesian decision.

From this standpoint we have developed a unified study which provides a theoretical foundation for various methods of multivariate data analysis (MDA) and quantification [21]. It has been shown that each MDA method is intrinsically based on the identical Bayesian structure shown in the previous subsection and closely related to each other, and that the stochastic matrix S or always plays an important role as the multi-class probabilistic and statistical knowledge obtained from data. It should be noticed that the usual linear methods of MDA can be regarded as linear approximations of the unified Bayesian structure and NN, and recent kernel methods [26] also, can be regarded as its approximations in some extents of nonlinearity.

The linear methods of MDA can be viewed as a linear feed-forward NN's. The ultimate nonlinear discriminant mapping “ $_N$ can be viewed as the opposite extreme, namely the ultimately optimum nonlinear feed-forward NN. Actually, it is seen that the back-propagation learning of ordinary NN tries to approximate the optimum nonlinear mapping “ $_N$ in Eq.(11), and Eq.(12) yields the theoretically lowest bound of the mean square error.

These theoretical results provides a foundation for flexible recognition systems. In practical applications, however, it is also important to reduce the theoretical framework to tractable new computation models, by taking into account the specific structures and knowledge and proper approximation according to respective actual application.

5 MDA Approach to Image Processing

MDA (Multivariate Data Analysis) simply utilizes linear models for integrating multivariate information and has some limitation. However MDA has been widely used as a practical and nonparametric method, in particular in the research fields such as psychology and social science which deal with ambiguous information related to human behavior. A merit of using MDA is that it simply results in a closed form solution with using up to the second order statistics and well-developed linear algebraic calculation.

The author pointed out in his early research that MDA would provide useful methods for image processing, and has developed several methods for practical applications.

5.1 Automatic threshold selection

The first application was the automatic threshold selection method [15], which is still widely used as a standard method.

Threshold selection is a typical example of an unsupervised method to convert/classify an input gray-level image to categorized subregions (objects and background), where intermediate representation (feature) is used as the histogram of the input image. Discriminant analysis, more strictly Fisher's discriminant criterion, was used to evaluate the class separability. The optimum threshold is selected so as to maximize the criterion, and it is also optimum in the sense of the least square error to approximate the input image with the resulting binary image. The method was naturally extended to the case of multi-thresholding, and an efficient method using DP was proposed.

5.2 Adaptive image processing

Many image processing problems can be formulated as linear filtering. Therefore, if the ideal resulting processed image to an input image is provided (by a teacher, for example), the problem to seek the optimum filter is reduced to the problem of Multiple Regression Analysis (MRA). Namely, by regarding the filter as a linear mapping $_j = \boldsymbol{a}'\boldsymbol{x}_j$ from the neighboring pixel values (vector $\boldsymbol{x}_j$) around each reference pixel to the corresponding ideal pixel value $_j$, the least squares optimum filter is obtained by applying MRA [18]. This method provides an approach to adaptive and trainable image processing systems.

6 General Approach to Flexible Vision

Finally, we introduce a general approach to flexible vision systems for adaptively trainable image measurement and recognition [16], [20], which has been developed as a simple realization of our theoretical standpoint.

There, we selected the following three conditions as fundamental and important requirements in practical image measurement and recognition:

C1: Shift-invariance

C2: Frame-additivity

C3: Adaptive trainability.

The first condition C1 means that no matter where an object exists within the image frame, its measured value (say, an area) or recognition result (say, a character " A") should be the same (invariant). The second condition C2 means that as is obvious from an example of counting the number of particles in the image frame, the total count over the whole image frame is equal to the sum of the counts on partitioned regions. The third condition C3 is important for a general-purpose system.

Here, it is interesting to note that the MDA methods in the previous section also satisfy these conditions eventually. In fact, histogram is a simple case of feature vector satisfying C1 and C2.

6.1 Scheme of feature extraction

In order to satisfy those conditions, we considered the following scheme of feature extraction consisting of two stages that is suggested as a basic framework of feature extraction in pattern recognition.

F1: *Geometrical Feature Extraction*; A large number of general and primitive features which satisfy C1 and C2 are extracted from the whole image frame as initial features.

F2: *Statistical Feature Extraction*; By means of linearly combining the initial primitive features, new features which are optimal to respective application are adaptively extracted through the learning from examples in order to satisfy C3.

It is noted that the linearity in F2 is important not only to simplify the learning process but also to conserve the condition C2 in F1.

As the simplest realization of the above scheme, we adopted Higher-order Local Auto-Correlation (HLAC) features for F1 and Multivariate Data Analysis (MDA) methods for F2.

6.2 HLAC features as F1

The Higher-order Local Auto-Correlation (hereafter denoted by HLAC) features are defined by

$$f(\boldsymbol{a}_1 \quad \boldsymbol{a}_N) = \sum_{\boldsymbol{r}\in P_2} (\boldsymbol{r})\ (\boldsymbol{r}+\boldsymbol{a}_1) \quad (\boldsymbol{r}+\boldsymbol{a}_N) \tag{15}$$

where N is the order of HLAC, $\boldsymbol{r}$ is the ( ) coordinate vector on the image plane P_2, $(\boldsymbol{r})$ the gray-level at position $\boldsymbol{r}$, and $\boldsymbol{a}_i$ the displacement vectors.

The number of HLAC features obtained by combining the displacements over P_2 is enormous, thus we reduce this number to enable practical application. We restrict the order N up to the second order ($N = 0\ 1\ 2$). Also, we restrict the range of displacements to within a local 3×3 region, whose center is the reference point $\boldsymbol{r}$, because the correlation within a local region is generally much higher than the correlation between distant points.

By eliminating displacements that are equivalent because of an even shift, we reduce the number of the displacement patterns to 25. Fig. 4 shows the 25 types of local displacement patterns, where "black" represents pixels to be examined while "white" represents "don't care".

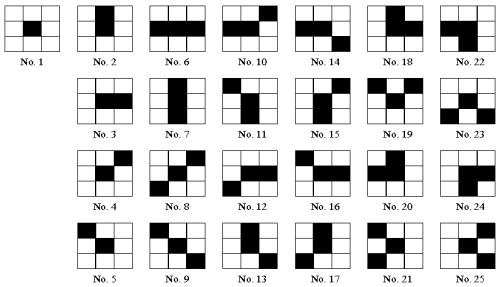


Figure 4: Local 3×3 masks up to the second order

Hence, HLAC features are obtained by once scanning the whole image over P_2 with the 25 local 3×3 masks and by computing the sums of the products of the gray values of the pixels corresponding to "black". HLAC features form an initial geometrical feature vector $\boldsymbol{x}$, and the dimension is 35 for a gray-level image and is reduced to 25 for a binary image due to the idempotents of product.

HLAC features are obviously additive for isolated objects on P_2, satisfying C2. Also, they are shift-invariant, satisfying C1, which makes the system robust to changes in the position of objects within an image and therefore segmentation-free.

6.3 MDA methods as F2

The initial HLAC features $_j$ are general and primitive and not necessarily adapted to a specified problem, however those in total contain enough information for given tasks. Thus we reorganize those by linear combinations so as to provide effective new features $_i$ which are estimates of measurements themselves or discriminant features for recognition purposes.

$$_i = \sum_{j=1}^{m} a_{ij}\ _j \quad (\ = 1 \qquad) \tag{16}$$

In this second stage of statistical feature extraction, the optimum coefficients $A = [a_{ij}]$ for a task is determined by using the MDA methods such as MRA or DA. The learning process is fast, because it consists of calculating statistics (means, covariances) of training examples $\boldsymbol{x}$ and solving a matrix equation which explicitly gives the optimal coefficients in a closed form.

6.4 Some practical applications

Although the system is designed as the simplest version of realizing the scheme, it has potential abilities as a flexible vision system, some of which will be shown in the following examples of experimental results [16], [20].

The first example in Fig. 5 illustrates the effectiveness of additive property C2 of the initial features. After each of six patterns (in the left) is taught once, the system can answer the numbers of each pattern in the image (in the right) simultaneously by simply decomposing the feature vector $\boldsymbol{x}$ of the image into those of learnt patterns as base vectors

(Factor Analysis). Then the coefficients are the numbers by virtue of additivity. It should be noticed here that the computation is very fast in real time by a simple matrix multiplication shown in Eq. (16) (in this case $_1$ through $_6$ are the numbers and A is determined from the six patterns), and it is a constant time no matter how many objects the image contains.

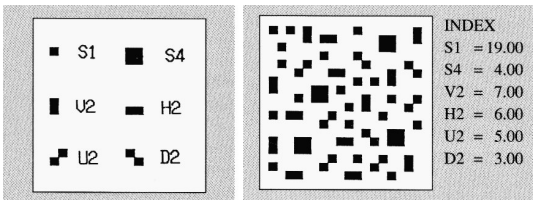
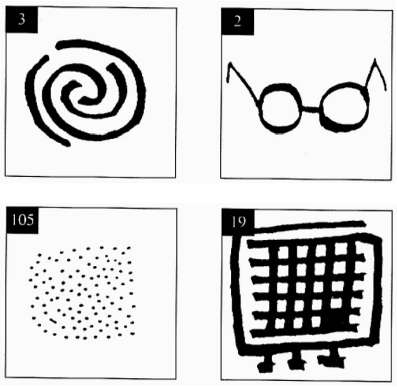


Figure 5: Simultaneous enumeration of multiple patterns

In Fig. 6 more interesting examples of application are shown, which are the measurements of topological characteristics (invariants) of binary images. The task is to count the number of isolated objects or the number of holes in an object. In this case single output $_1$ is the answer and the optimal coefficients are determined by MRA such that $_1$ approximates in least squares the ideal answers which are given in learning mode. After learning with 48 generated sample images for each task, the system could correctly answer to the new testing images which are shown in Fig. 6-a and in 6-b. It is interesting to note that the system learnt the Euler formula from the training sample images and utilized it to answer quite correctly.



a) Number of objects b) Number of holes

Figure 6: Measurements of topological characteristics

It should be noticed that we never teach the system what features should be taken for a given task and how to process, but the system learns it adaptively and automatically from training examples.

Recently, this system is easily implemented on an ordinary PC and straightforwardly extended to be applied to gray-level real images taken though a video camera in real time.

7 Generalized HLAC and its Applications

The formulation of HLAC feature extraction is so simple that it has recently been generalized in several ways to be applicable to more practical, complicated, and difficult tasks including color images and also motion images.

7.1 Recognition of Face and facial expression

We applied HLAC method to human face recognition [10], [3], where HLAC features were extracted from the multi-layered pyramid of input image. The recognition rate was quite high, 99% for 116 persons [3].

The method is also applied to more difficult task of facial expression [25], [13], where we used JAFFE facial image database [11] which consists of 7 facial expressions of 9 females (See Fig. 7) and obtained high recognition rate, about 80%. In [13] HLAC and the FA described in Fig. 5 are used to identify simultaneously person and facial expression.

On the other hand, in [25] HLAC and DA are used, and HLAC is expanded to weighted sum in order to adapt HLAC to local importance of face image over P_2. The optimal configuration of the weights, called "Fisher weight map", is adaptively obtained in DA. It is seen that the areas of eyebrows, cheeks, and lips have greater importance for recognition of facial expressions (See Fig. 8).

It is noted that the formulation is so general to include the Eigen-face and Fisher-face methods as special cases.

Figure 7: Some examples in JAFFE Database

Figure 8: Obtained Fisher weight maps (eigenvectors) @

7.2 Color HALC and robust tracking

Tracking is one of the most fundamental methods for motion image processing. Most of usual methods are based on segmentation of moving object and template matching, and identification is measured at the pattern (image) level. In the real-world environment, however, objects are so often occluded by obstacles, and which makes the methods difficult to be robust and reliable.

So as to realize robust and reliable tracking, it is better to measure similarity (matching) not at the pattern level but at the feature level. From this viewpoint, we applied Color HLAC features to tracking [7]. Because, color is also important feature to identify objects as well as object shape.

For color images, HLAC is extended by replacing the pixel gray value (scalar) with the pixel color values (RGB/HSV 3D vector). The dimension becomes combinatorially very high, thus the order of Color HLAC is trade off with the dimensionality.

Our method utilizes the background subtraction and the additivity of Color HLAC (dividing image to subregions and merging), and adopts k-NN decision rule. Experimental results show that the method quite robustly track moving

objects even when those are occluded by obstacles or crossing each other.

7.3 Cubic HLAC

For a motion image, HLAC feature vector forms a trajectory $\boldsymbol{x}(\ )$ in the vector space. Therefore motion features will farther be extracted by characterizing the trajectory in a more global term T. One intuitive way is to apply ARM to the trajectory, and which was applied to gesture recognition [5]. Another more direct and intensive way is to expand HLAC itself straightforward to 3D case [8]. The latter is called Cubic HLAC (hereafter denoted by CHLAC).

CHLAC is applicable to any form of three-way data, i.e. 3D ( ) data. The features are extracted by scanning the whole data $X \times Y \times Z$ (P_3) with a $3 \times 3 \times 3$ local cubic mask patterns (Fig.9). The dimension of CHLAC up to the second order is 279 for scalar data and degenerated to 251 for binary data. For motion image, we consider $=$ (time) and $Z = T$ is the width of a time-window.

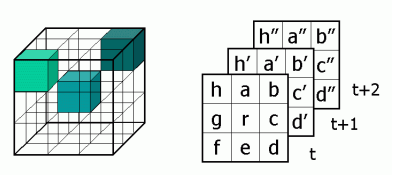


Figure 9: Example of mask pattern of CHLAC (h r' b")

7.3.1 Recognition of motion and gait

CHLAC is applied to motion and gait recognition in [8], showing effective performance (99.9% for 4 types of moves of 5 persons). There, input images are converted to binary ones by frame difference and thresholding, and CHLAC features are extracted and DA follows. Classification is based on the simple minimum distance decision rule.

CHLAC features also inherit the important properties, *shift-invariance* (rendering the method segmentation-free) and *additivity* as well as HLAC, and are robust to noise in data. Moreover, the method utilizes no *a priori* knowledge nor heuristics about objects such as human shape and angles of legs, etc.

Recently, we applied the method to the NIST gait dataset [24] and compared the identification result to those of the other methods. The dataset consists of 456 video sequences of 71 individuals, walking around the elliptical course. The result shows that our method is superior to other methods in spite of simple feature extraction and classification [9].

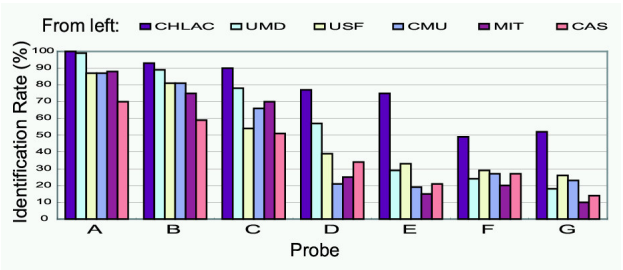


Figure 10: Comparison of gait recognition methods

7.3.2 Detection of abnormal movement

The detection of abnormal (unusual) movements in scenes is a crucial and urgent issue in video surveillance applications for security. For example, if abnormal movements of persons are automatically detected and screened, it saves a lot of labor for human to monitor all the time.

For detecting abnormal movements, systems need to recognize those. However it is almost impossible to learn all the examples of rare abnormal (unusual) movements in advance. Nevertheless, we can define abnormal movements as *not* being normal (usual) movements that are frequently happening in front of a camera. Therefore, abnormal movements can be detected only by learning normal movements statistically.

Based on this idea, we proposed an unsupervised method for abnormality detection in scenes containing multiple persons (See details [12] in this conference). Our method uses CHLAC (Cubic HLAC) to extract features of movements and utilizes a subspace method to detect abnormality. One particular advantage of this method is that it does not necessitate the object segmentation and tracking and also any prior knowledge about objects.

Actually, CHLAC can extract features of multiple persons' motions *without* segmenting and tracking each person (due to shift-invariance), and the computational cost is constant regardless of the number of persons. Furthermore, the additive property of CHLAC in combination with a linear subspace method is well suited to simplify the learning of normal movements and the detection of abnormal movements even in scenes containing multiple persons. Namely, all the normal movements are included in the subspace of normal movements even for scenes containing multiple persons' moves, and only abnormal movements depart from the subspace. Thus, once the subspace of normal movements S_N is constructed by using PCA for example, abnormality is easily detected by measuring the distance between an input feature vector $\boldsymbol{x}$ and S_N.

The distance $_\perp$, which is used as index of abnormality, is easily calculated as

$$_\perp = \|P_\perp \boldsymbol{x}\| \tag{17}$$

where $P_\perp$ is the projector onto ortho-complement subspace of S_N and easily calculated by using the eigenvectors.

An experimental result is shown in Fig. 11, where it is seen a tumbling movement (abnormal) of a person in the scene containing two other persons' walking (normal) is successfully detected as an abnormal movement.

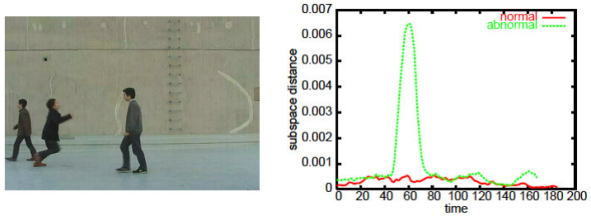


Figure 11: Example of abnormality detection.

For more practical uses, it is noted that the method is also implemented for on line and real-time learning and detec-

tion by incrementally solving the updated eigenvalue problem or by incrementally approximating the eigenvectors.

8 Concluding Remarks

An approach to flexible and intelligent vision systems has been discussed, showing a general framework of visual information processing and also of pattern recognition.

It was emphasized that intermediate representation, or feature vector representation, is crucial to smoothly bridge patterns (at signal level) to numerals (at measurement level) and also to symbols (at recognition level).

Another important point emphasized is the two stages of feature extraction in the general framework of pattern recognition; namely geometrical (invariant) feature extraction as the first stage, and statistical (adaptive and discriminant) feature extraction as the second stage. Theoretical foundations for the two stages have been shown. The former geometrical stage is formulated by Lie group theory and partial differential equations, while the latter is formulated by Bayesian estimation in the ultimate, and by MDA methods, NN, or kernel MDA in the approximations.

As a simplest implementation, we showed a scheme of vision system comprising HLAC and MDA, demonstrating its flexible and intelligent performance. HLAC has been straightforwardly expanded to CHLAC (Color or Cubic HLAC) so as to farther cope with color and/or motion images. The effective performance has been shown for actual applications such as motion or gate recognition, tracking, and abnormality detection in video surveillance.

These results present the potential prospect of the scheme towards flexible and intelligent vision systems.

References

[1] K. Fukunaga: *Introduction to Statistical Pattern Recognition*, Academic Press, 1972.

[2] S. Geman and D. Geman: "Stochastic Relaxation, Gibbs Distributions and the Bayesian Restoration of Images," IEEE Trans. PAMI, **6**, 721–741, 1984.

[3] F. Goudail, E. Lange, T. Iwamoto, K. Kyuma and N. Otsu: "Face Recognition System Using Local Autocorrelations and Multi-Scale Integration," IEEE Trans. PAMI, **18**, 1024–1028, 1996.

[4] G. Hinton, T. Sejnowski and D. Ackley: "Boltzmann Machines — Constraint Satisfaction Networks That Learn," Tech. Rep. CMU-CS-84-119, 1984.

[5] T. Ishihara and N. Otsu: "Gesture Recognition Using Auto-Regressive Coefficients of Higher-Order Local Auto-Correlation Features," Proc. 6th IEEE Int. Conf. on Automatic Face and Gesture Recognition, 583–588, 2004.

[6] J. Hopfield and D. Tank: "Neural Computation of Decisions in Optimization Problems," *Biol. Cybern.*, **52**, 141, 1985.

[7] M. Kawai: "Robust Tracking of Moving Objects in Motion Images in the Real Environment," (in Japanese) Graduation thesis, Univ. of Tokyo, 2004.

[8] T. Kobayashi, and N. Otsu: "Action and Simultaneous Multiple-Person Identification Using Cubic Higher-Order Local Auto-Correlation," Proc. 17th ICPR, 741–744, 2004.

[9] T. Kobayashi and N. Otsu: "A Three-way Correlation-based Method for Human Identification by Gait," (submitted to ICCV2005).

[10] T. Kurita and N. Otsu: "Face Recognition Method using Higher Order Local Autocorrelation and Multivariate Analysis," Proc. 11th ICPR, 213–216, 1992.

[11] M. Lyons and S. Akamatsu: "Coding Facial Expressions with Gabor Wavelets," Proc. 3rd IEEE Int. Conf. on Automatic Face and Gesture Recognition, 200–205, 1998.

[12] T. Nanri and N. Otsu: "Unsupervised Abnormality Detection in Video Surveillance," (to be presented at MVA2005).

[13] N. Nomoto and N. Otsu: "A New Scheme for Image Recognition using Higher-order Local Autocorrelation and Factor Analysis," (to be presented at MVA2005), 2005.

[14] N. Otsu: "Nonlinear Discriminant Analysis as a Natural Extension of the Linear Case," *Behaviormetrika*, **2**, 45–59, 1975.

[15] N. Otsu: "Discriminant and Least Squares Threshold Selection," Proc. 4th ICPR, 592–596, 1978.

[16] N. Otsu: *Mathematical Studies on Feature Extraction in Pattern Recognition*, (in Japanese) *Researches of ETL*, No. 818, 210 pages, 1981.

[17] N. Otsu: "Optimal Linear and Nonlinear Solutions for Least-square Discriminant Feature Extraction," Proc. 6th ICPR, 557–560, 1982.

[18] N. Otsu: "Multiple Regression Analysis Approach to the Automatic Design of Adaptive Image Processing Systems," Proc. SPIE, **435**-9, 70–75, 1983.

[19] N. Otsu: "Recognition of Shape and Transformation – An Invariant-theoretical Foundation," in *Science on Form*, ed. S. Ishizaka, KTK Reidel Pub. 1986.

[20] N. Otsu and T. Kurita: "A new Scheme for Practical Flexible and Intelligent Vision Systems," Proc. IAPR Workshop on Computer Vision (MVA1988), 431–435, 1988.

[21] N. Otsu, T. Kurita and H. Asoh: "A Unified Study of Multivariate Analysis Methods by Nonlinear Extensions and Underlying Probabilistic Structures," in *Recent Developments of Clustering and Data Analysis*, E. Diday *et al.* eds., Academic Press, 1988.

[22] T. Poggio and K. Koch: "Ill-posed Problems in Early Vision – from Computational Theory to Analogue Networks," Proc. Roy. Soc. London, **B226**, 303–323, 1985.

[23] D. Rumelhart, G. Hinton and R. Williams: "Learning Representations by Back-propagating Errors," *Nature* **323**-9, 533-536, 1986.

[24] S. Sarkar *et al.* : "The HumanID Gait Challenge Problem: Data Sets, Performance, and Analysis," IEEE Trans. PAMI, **27**, 2, 162–177, 2005.

[25] Y. Shinohara and N. Otsu: "Facial Expression Recognition using Fisher Weight Maps," Proc. 6th IEEE Int. Conf. on Automatic Face and Gesture Recognition, 499–504, 2004.

[26] J. Shawe-Taylor and N. Cristianini: *Kernel Methods for Pattern Analysis*, Cambridge Univ Press, 2004.