ARGUS: Adaptive Recognition for General Use System

— Its theoretical construction and applications —

Nobuyuki OTSU

[Translation from *Synthesiology*, Vol.4, No.2, p.70-79 (2011)]

In recent years, the need for computer vision systems is increasing in various fields, such as security and visual inspection. It is crucial there to realize simple and high-speed practical vision systems. The present paper addresses the author's theoretical research and its applications developed thus far in working toward this goal. First, the problem of the conventional approach is pointed out, and the general framework of pattern recognition, in particular the feature extraction theory, is referred to as the theoretical foundation. Next, a scheme of adaptive vision system with learning capability is presented, which comprises two stages of feature extraction, namely, Higher-order Local Auto-Correlation and multivariate data analysis. Several applications are demonstrated, showing the flexible and effective performance of the proposed scheme.

Keywords **:** Vision system, pattern recognition, feature extraction, adaptive learning

1 Introduction

In recent years, there have been great expectations for vision systems (computer vision). They are useful in various fields including surveillance cameras for crime prevention, appearance inspection of manufactured goods, CT scans and tissue examination in medicine, robot vision, and analysis and evaluation of movement in sports studies as well as image searching on the Internet. Furthermore, as an important aspect of vision systems, it should be noted that collection and processing of various images has become easier, owing to the technological development of CCD cameras, various sensors, computers, and visualization techniques.

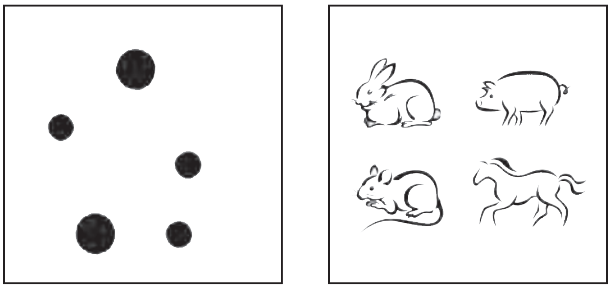
With developments in the field of vision systems, image recognition research has been pursued vigorously on an international level, but automation and implementation has been difficult. In addition, only distinct ad hoc methods and expensive specialized systems have been developed, and there is still a reliance on human abilities under actual settings. As a result, the implementation and distribution of a cheap, PC-based vision system that is versatile and delivers high speed is highly desirable.

With the above objective, this paper discusses the pattern recognition theory developed by the author thus far[1], focusing on feature extraction theory[2] and the Adaptive Recognition for General-Use System based on it that was proposed as a practical system construction method[3][4], as well as various practical developments[5][6]. Moreover, the effectiveness and importance of the theoretical approach in particular is demonstrated when considering a construction method for recognition (generally, information) systems.

2 Ordinary approaches and pattern recognition

First, let us consider the pattern recognition problem of image measurement and recognition. Fig.1-a) shows the image measurement (enumeration) task where there are round particles of two different sizes and the total number of each is enumerated. The method usually considered is similar to the following sequential method. First, the screen is scanned to segment individual particles, and then the radius is measured for the approximating circle of each particle; in this manner, the size of the particle can be determined from the radius, and the particles can thus be counted. However, this method will clearly result in an increase in the calculation time, proportional to the number of objects.

On the other hand, Fig.1-b) is the image recognition problem that identifies what each object (animal) is. It is usual to consider the characterizing features (parts) that distinguish these four objects. In the context of each model, ears, tails,



a) Image measurement b) Image recognition

Fig. 1 Examples of vision tasks[6]

Fellow, AIST Tsukuba Central 2, 1-1-1 Umezono, Tsukuba 305-8568, Japan E-mail : otsu.n@aist.go.jp

Original manuscript received April 14, 2010, Revisions received April 12, 2011, Accepted April 13, 2011

and body type among other features are compared (or quantified) as partial images in order to reach a final decision. However, overall recognition is dependent on the recognition of parts, which implies that the overall recognition will be incorrect if the part recognition is erroneous.

Thus, most ordinary approaches are "serial and procedural types" that first segment each individual object in this manner from the image, and then perform recognition according to a pre-prepared model. However, because the pattern generally has different variations, the model must also be made proportionately more complex in various ways. Moreover, accumulation of errors at each stage of processing in serial procedures results in overall vulnerability; a large amount of calculation is involved and it is difficult in practice to obtain the required recognition performance. The problem lies in the tendency to consider this as a logical procedure in an ad hoc manner at the image level. In a way, it is an approach dominated by the Neumann-type computer programming paradigm.

As the antithesis of this method, from the late 1980s, the "parallel and adaptive (learning) type" method was proposed using neural networks[7]; in addition to the study of the theoretical aspects, various applications have been attempted, especially in pattern recognition and control. However, because of the constraints that the elements are nonlinear and have bounded values [0, 1], information representation and feature extraction usually tend to be ambiguous. In recent years, additional problems such as the arbitrary nature of the model and the learning speed and convergence have been indicating the need for a change toward the nonlinear multivariate analysis, such as the kernel method[8].

To examine a new methodology for visual systems, recognition systems in general, it is necessary to theoretically reconsider the general framework of the underlying pattern recognition mechanism, especially information representation and feature extraction.

2.1 General framework for pattern recognition

In pattern recognition, recognition is accomplished by multiple extraction (thus represented by the vector $\boldsymbol{x}$) of some feature values effective for recognition (generally functionals $x_i = \phi_i[f]$ defined as functions of the function f) from the pattern, a signal expressed by a function f localized in space-time. Typically, as shown in Fig.2, the framework comprises a two-stage process of "feature extraction" and "recognition". Recognition can be divided into classification and clustering. Classification is the determination of whether the input pattern corresponds to one of the several known categories, and is called supervised learning because the answer is given in the learning stage. Clustering is called unsupervised learning, which discriminates the input pattern into several clusters (categories). Many techniques have been proposed regarding classification, and it is already theoretically known that the minimum error rate classification method is the Bayesian decision rule, which decides on the category C_j with a maximum posterior probability $P(C_j|\boldsymbol{x})$. This implies that the feature extraction at the first stage is important as the requirement which dominates the recognition system efficiency, however various ad hoc or heuristic techniques have been suggested until date.

2.2 Feature extraction theory

The author has conducted a theoretical study of these feature extractions[2]. General framework for feature extraction comprises "invariant feature extraction" as the geometrical aspect and "discriminant feature extraction" as the statistical aspect. In principle, it is important that feature extraction comprises these two stages in this order. Fig.3 thus demonstrates the general framework for pattern recognition, as a natural consequence of this theory.

2.2.1 Invariant feature extraction (geometrical aspect)

The observed image f as a pattern is subject to various continuous geometrical transformations (generally, projective transformations) such as translation, scaling, and rotation due to the relative position and movement of the observer and the object. However, recognition results are independent of these and remain invariant. In

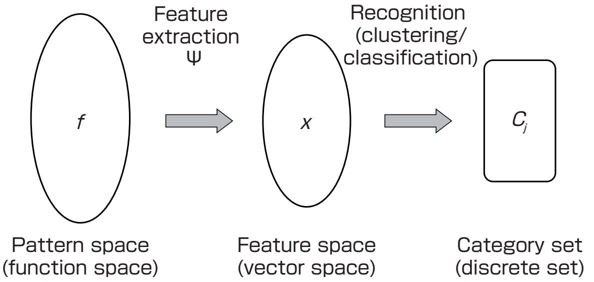


Fig. 2 General framework for pattern recognition (typical)

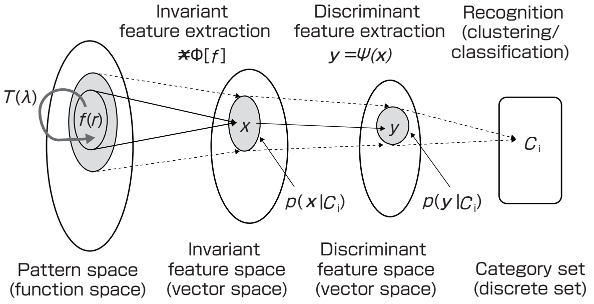


Fig. 3 General framework for pattern recognition (detail)

invariant feature extraction theory, such a geometrical transformation that operates on the pattern function f and keeps category correspondence invariant (called an invariant transformation) is represented by an operator $T(\lambda)$, under which the invariant feature values, and thus the corresponding invariant functionals $x = \phi[f]$ are pursued.

$$\phi[T(\lambda)f] - \phi[f] = 0 \tag{1}$$

Using operator analysis based on Lie group theory, the invariant features, corresponding to the given invariant transformation, are found as elementary solutions of the partial differential equation, derived as a necessary and sufficient condition[1][2]. In this manner, the pattern, as the fundamental features for recognition through abstraction of extraneous information, can ideally be treated in unity as a single point x in the invariant feature vector space.

2.2.2 Discriminant feature extraction (statistical aspect)

However, actual patterns are subject to variations and noise and are distributed according to a probability distribution $p(\boldsymbol{x}|C_j)$ for each category class C_j. The next stage of the discriminant feature extraction theory considers a mapping $\boldsymbol{y} = \Psi(\boldsymbol{x})$ of the invariant feature vector $\boldsymbol{x}$ to a new feature vector $\boldsymbol{y}$ with reduced dimensions and derives an optimum mapping that optimizes an evaluation criterion for $\boldsymbol{y}$, such as discrimination of the category classes. So-called multivariate analysis methods (such as discriminant analysis) are usually formulated as linear mappings, whereas a neural network or a kernel method is used for some type of nonlinear mapping.

In fact, the ultimate optimum nonlinear discriminant mapping is easily obtained in the following formula using variational calculus[1][2].

$$\boldsymbol{y} = \Psi_N(\boldsymbol{x}) = \sum_{j=1}^{K} P(C_j|\boldsymbol{x})\boldsymbol{c}_j \tag{2}$$

This result shows that pattern discrimination is closely related to Bayes posterior probability $P(C_j|\boldsymbol{x})$, and suggests the essential framework of Bayesian inference behind pattern recognition. Here, $\boldsymbol{c}_j$ are the vectors that represent each category in the target mapping Y, and in the case of discriminant analysis, they are derived as eigenvectors of the between-class stochastic matrix in the original space X. It is understood that the dimensions of the optimum discriminant space obtained are essentially determined by the number of classes, therefore coming to K-1 dimensions.

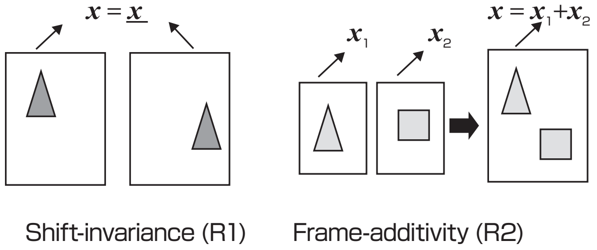

$\boldsymbol{x} = \underline{\boldsymbol{x}}$
$\boldsymbol{x}_1$
$\boldsymbol{x}_2$
$\boldsymbol{x} = \boldsymbol{x}_1 + \boldsymbol{x}_2$
Shift-invariance (R1)
Frame-additivity (R2)

Fig. 4 Shift-invariance and frame-additivity[6]

In real-life applications, it is necessary to make appropriate simplifications according to practical requirements considering these theoretical frameworks.

3 Approach and conditions for a constructing method

While considering the approach toward a flexible vision system and a constructing method, the following three points can be mentioned as basic conditions that are required for the vision system (Fig.4).

R1: Shift-invariance,
R2: Frame-additivity,
R3: Adaptive trainability.

The results of such recognition or measurement should be the same regardless of where the recognition or measured object is in the image frame. Thus, the first condition R1 demands that a feature $\boldsymbol{x}$ extracted from the pattern does not depend on the position of the object (it is invariant under a parallel shift). Size scaling, rotation, and other transformations can also be considered as invariant transformations. However, since a parallel shift is the most fundamental, it was made a required condition.

The next condition R2 requires that features for the entire screen are the sum of local features for individual objects. This is also a consequence of R1, and is a required condition where feature representation is a convenient representation (linear) for recognition (especially counting), and the processing afterward becomes simple and high speed.

Unlike the ordinary method where feature extraction is given as a heuristic procedure and the construction method changes in accordance with the change in recognition tasks, the last condition R3 requires that a new feature $\boldsymbol{y}$ suited to the task is automatically constructed (synthesized) from the initial feature $\boldsymbol{x}$ in an optimal manner using the learning acquired from the example; in addition, the condition requires that the method is a general-purpose formulation that is adaptively optimized with a structure indifferent to changes in the task.

In addition, for such a feature extraction method constructed to meet these required conditions, it is desirable for the computation amount to be low and that real-time processing is possible.

4 Adaptive Recognition for General-Use System(ARGUS)

An “adaptive image recognition system for general use”[Note 1] has been devised that meets these basic demands and is implemented using the simplest form for the aforementioned pattern recognition, especially for the framework of feature extraction theory[3][4]. This system comprises a two-stage feature extraction following the theoretical framework of feature extraction (Fig.5).

4.1 Invariant feature extraction (HLAC/CHLAC)

The most basic features were considered to be parallel shift-invariant (position-invariant), for an initial feature in the first stage, namely, feature extraction that is invariant from a geometrical aspect. This is because recognition is essentially independent of the position $\boldsymbol{r}$ of the spatio-temporal pattern $f(\boldsymbol{r})$.

As position-invariant features, the auto-correlation function $r(\tau) = \int f(t) f(t+\tau)\, dt$ in the field of time series analysis of audio, has been known for a long time. This extracts the relative relationship (correlation) in a wave profile pattern that does not depend on time position. The higher-order expansion of this, Nth order auto-correlation function is known mathematically,

$$x(\boldsymbol{a}_1, , , \boldsymbol{a}_N) = \int f(\boldsymbol{r}) f(\boldsymbol{r}+\boldsymbol{a}_1) \cdot\cdot\cdot f(\boldsymbol{r}+\boldsymbol{a}_N)\, d\boldsymbol{r} \qquad (3)$$

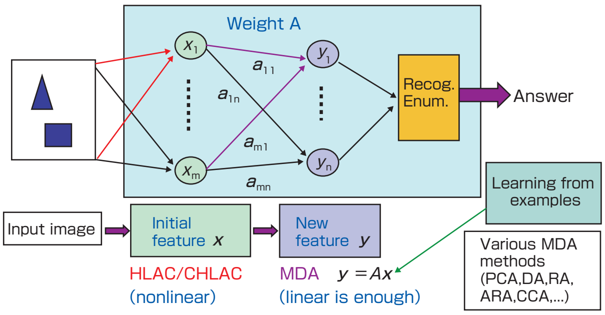


Fig. 5 Adaptive Recognition for General-Use System (ARGUS)

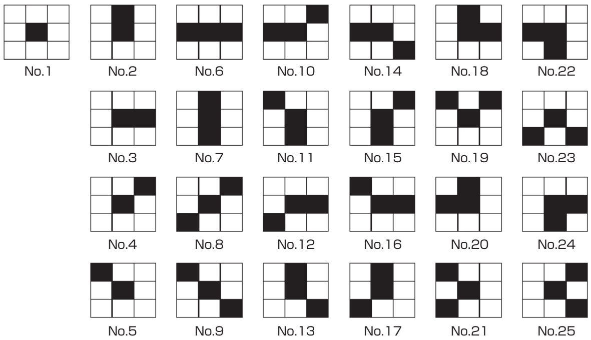


Fig. 6 Local 3 × 3 masks up to the 2nd order [3][4]

and the 2nd order (generally even-order) autocorrelation functions form a complete system. As such, several interesting properties pertaining to pattern recognition applications, have been discussed[9]. For an image, $f(\boldsymbol{r})$ is the gray-scale value at the reference point (image pixel) $\boldsymbol{r}$, and $\boldsymbol{a}_i$ is the relative displacement around the reference point $\boldsymbol{r}$. However, the number of feature values becomes exponentially large according to the combination of N displacements, and their computation is almost impossible. Thus, combinations of limited orders and displacements are used in a practical application.

In fact, patterns in the real world are spatio-temporally localized and the local relative relationships are essencial. Moreover, this localization also satisfies frame-additivity (R2). Therefore, as nonlinear features that satisfy both R1 and R2, Higher-order (Nth order) Local Auto-Correlation (HLAC) features, obtained from the higher-order autocorrelation function in Eq. (3) limited to a localized displacement, were devised and adopted[3][4].

HLAC: For an actual 2D image (a still image) $f(x, y)$, restricting the order to 2nd order and displacements to a local 3×3 region, there are 25 patterns of local masks for taking inequivalent and independent sum of products when considering the shift variance (Fig.6). For the full screen (or a sub-region) XY, scanning each of the local masks shown in Fig.6 and finding the sum of products of the pixel values corresponding to black dots gives the HLAC feature vector $\boldsymbol{x}$. Its dimension is 35 for a gray-scale image (e.g., for mask No. 1, distinguishes f , f^2, f^3), and for a binary image (0/1, white/black), it degenerates to 25 dimensions (e.g., for mask No. 1 yields $f = f^2 = f^3$ as idempotent)[Note2].

CHLAC: In the case of a moving image (3D) $f(x, y, t)$, since there are three-dimensional (solid) numerical data over XYT formed from the two-dimensional still images lined up along the time axis, features are extracted for CHLAC (Cubic HLAC), which naturally expanded HLAC by including the time axis[11]. Fig.7 shows an example of a local 3×3×3 mask for CHLAC. There are 251 independent local mask patterns. As with HLAC, the CHLAC feature vector $\boldsymbol{x}(t)$ is obained by finding the sum of products using

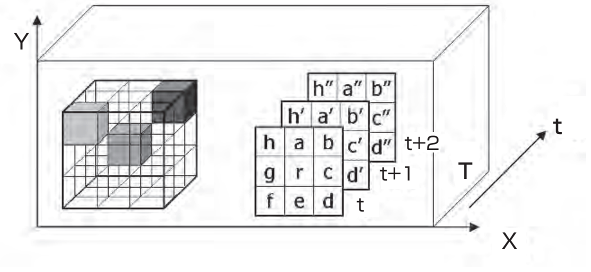


Fig. 7 Example of CHLAC mask (hr’b”) [5][11]

the masks over the solid frame *XYT*. The dimension is 279 for a gray-scale moving image and 251 for a binary moving image.

The feature extraction methodology using the integral features of HLAC (CHLAC) is a fundamental and general-purpose feature extraction method for "object shape (and movement)" and satisfies the required conditions R1 (shift-invariance) and R2 (frame-additivity). Employing these methods, the recognition object can always be captured and represented in a unified manner as a single point (a vector) $\boldsymbol{x}$ in an invariant feature space.

4.2 Discriminant feature extraction (MDA)

In the next stage of adaptive learning (satisfying R3), statistical discriminant feature extraction, various multivariate data analysis (MDA) techniques are applied as linear mappings (Fig.8). This refers to adaptively deriving a new feature $\boldsymbol{y}$ optimized for the given recognition task, as the weighted linear sum of the elements of the HLAC or CHLAC feature vector $\boldsymbol{x}$ (R3: adaptive trainability) (Fig.5) ; since the mappings are linear, this secures the required condition R2 (frame-additivity).

A similarity can be found in neural networks, but owing to its nonlinearity R2 is not preserved. In addition, it requires an iterative solution for optimization and takes much computation time. On the other hand, MDA has the advantage that by learning through examples, the weights optimal for the tasks are easily obtained in an analytically explicit form[Note3].

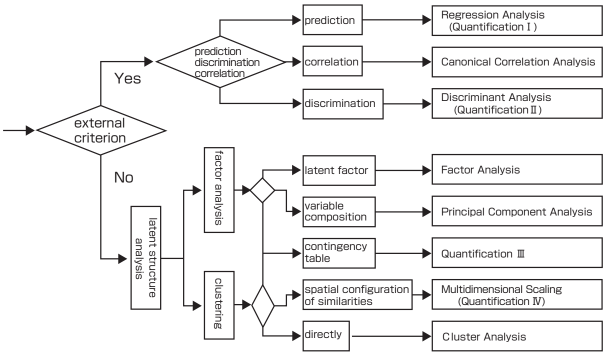


Fig. 8 Multivariate data analysis method (by objective)

Whether an external criterion exists corresponds to being supervised. A quantification method is for the case of qualitative data (Yes/No, 1/0).

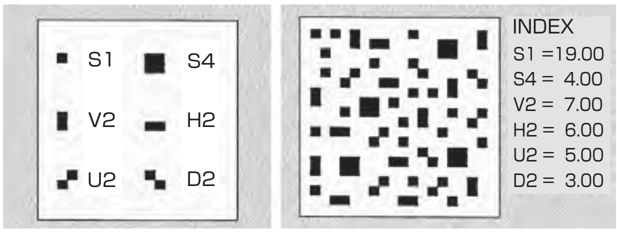


Fig. 9 Simultaneous recognition (/enumeration) of multiple objects[5][6]

4.3 Characteristics of the ARGUS recognition system

This formulation comprising these two stages of feature extraction does not require segmentation or positioning of the object, and is unique in not requiring any knowledge or model of the object. Thus, the formulation has a versatility that makes it applicable for various recognitions, measurements or enumerations of still and moving images. Moreover, since it basically performs only the multiply-accumulate operation, even CHLAC can run on a normal PC with an extremely high processing speed (2 msec/frame).

5 Application examples

5.1 Simultaneous recognition (/enumeration) of multiple objects

As an example of recognition of still images, an application for the enumeration task of simultaneously recognizing multiple objects is presented (Fig.9). This can be easily realized by utilizing factor analysis (FA), based on the shift-invariance

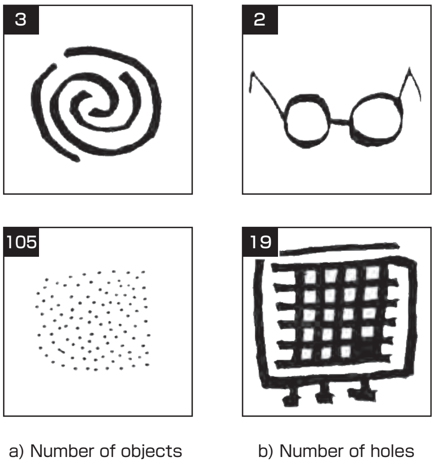


Fig. 10 Recognition (/enumeration) of topological characteristics[5][6]

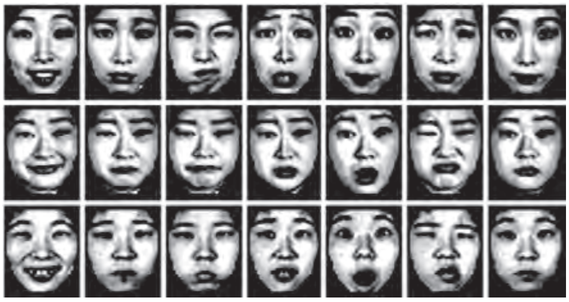


Fig. 11 Example of the JAFFE facial expression dataset (3 people)[14]

and the additive properties of HLAC features. Once each of the patterns on the left in Fig.9 are presented to the system, the system instantly responds to the test image (on the right) with the numbers y_i of each as $\boldsymbol{y} = (F'F)^{-1}F'\boldsymbol{x}$. This is by virtue of additivity, where the feature vector $\boldsymbol{x}$ for the entire right-hand diagram can be decomposed into the linear sum $\boldsymbol{x} = \sum_{i=1}^{6} y_i\boldsymbol{f}_i = [\boldsymbol{f}_1, \ldots, \boldsymbol{f}_6]\,\boldsymbol{y} = F\boldsymbol{y}$, which has as its coefficient the number of feature (factor) vectors $\boldsymbol{f}_i$ for each pattern.

5.2 Recognition (/enumeration) of topological characteristics

Next, as an example of recognition that is independent of shape, recognition outcomes for topological characteristics are given in Fig. 10. By learning from examples using multiple regression analysis (MRA), the system correctly answered the number of objects (a) or the number of holes (b)[4]. Interestingly, from the examples, the system learned the Euler number that underlies the basis of topology (number of points − number of lines + number of planes)[Note4] and used it for recognition.

5.3 Recognition of faces and facial expressions

HLAC is not limited to binary images and can be directly applied to gray-scale images as well. Face recognition was done as such an example[12][13]. By integrating with discriminant analysis (DA), the HLAC features extracted from each layer of a pyramid of images representing multi-resolution, even the simple classification method MDD[Note5] achieved a high recognition rate of more than 99 % among 119 people[13]. Furthermore, the method was applied to the difficult task of facial expression recognition for seven facial expressions by nine people (JAFFE Dataset[14], Fig.11). Using the MDD and the discriminant analysis that takes into account the position based weighting of HLAC features, a high recognition rate of more than 80 % was achieved[15].

5.4 Recognition of person and motion

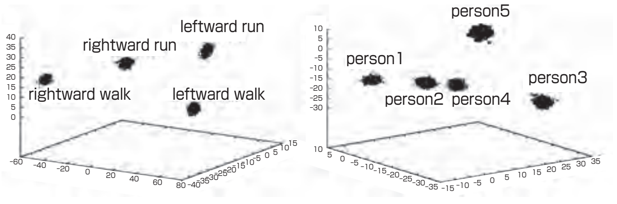
Using CHLAC features that are a natural expansion from HLAC features when moving images are considered, both object and movement can be recognized in a moving image. Videos of four motions (walking/running to the left/right) by five persons were converted to binary images based on frame difference and thresholding, and the CHLAC features were extracted. Fig.12 shows the results after applying discriminant analysis to person and motion, respectively[11]. Each cluster (category) is well grouped and separated,demonstrating the effectiveness of CHLAC features. Even with a simple classification method MDD, recognition rate of almost 100 % was obtained.

5.5 Recognition of gait

In recent years, the concept of "gait" has attracted attention as a key in the identification of individuals(terrorists, etc.) by surveillance cameras from a distance. Application of CHLAC and discriminant analysis together with the *k*-NN decision rule to the Gait Challenge Dataset (Fig.13) of 71 individuals compiled by the NIST in the United States has achieved the best performance in the world thus far, significantly surpassing the top five methods[16](Fig.14).

5.6 Abnormality detection

When there are multiple objects in an image, CHLAC has the additive property in which the sum of the features of each object becomes the features of the whole; therefore, the feature vector for usual (normal) motion will be



a) Motion discrimination b) Person discrimination

Fig. 12 Obtained discriminant feature spaces[11]

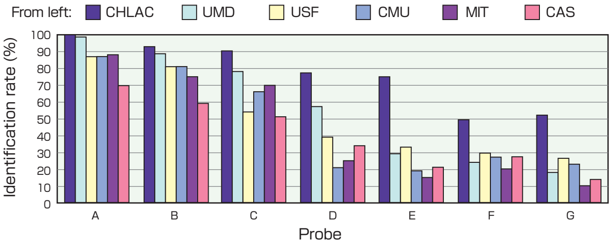


Fig. 14 Comparative experiment of "gait" recognition[5][6][16]

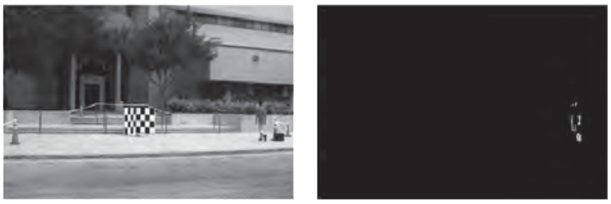


Fig. 13 Gait video and frame difference

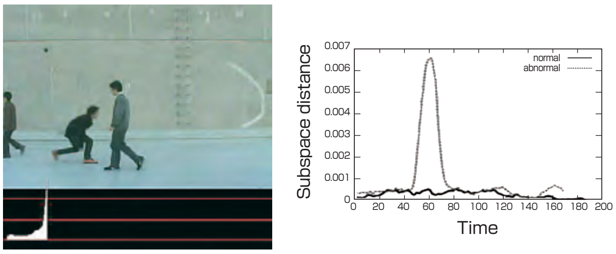


Fig. 15 Example of abnormality detection (Here, "falling down" is abnormal.)

distributed in a linear subspace (usual motion subspace) S_N in the feature space (with 251 dimensions). Accordingly, once S_N has been derived using principal component analysis (PCA) while learning on a regular basis (unsupervised), abnormal behavior does not require prior definitions, and can be detected and recognized immediately with high speed and accuracy, as a deviation (in distance or angle) from S_N[17] (Fig.15). Because of its additivity, capability for detecting abnormalities remains constant even for multiple persons (Fig.16).

This abnormality detection system is already put into practice with surveillance cameras in elevatorsNote6.

This system, where usual cases are learned as a statistical distribution in the CHLAC feature space and abnormalities are detected as deviations from such distribution (unusual), does not necessitate any model or knowledge of the objects. Accordingly, the system can be applied not only for abnormality detection from footage taken by surveillance and car-mounted cameras, but also for various other abnormality detection scenarios. For example, using the HLAC feature space for still images, it can be applied to various appearance inspections such as in the field of manufacturing semiconductor substrates (Fig.17).

Moreover, abnormality detection using HLAC is equally applicable in the medical field for various kinds of tissue examinations, especially in the pathological diagnosis of cancer. Cancer is a cell abnormality. The pathological diagnosis of cancer is conducted under a microscope by a pathologist who determines the degree of change found in the structure and the cells of organ tissues. However, this requires a wealth of experience and knowledge, and experienced pathologists are in short supply with their ever-increasing workload. Thus, there is great demand for system development to support pathologists, in the form of alleviating the burden of screening tests through automation,and preventing oversights through crosschecking. When this method was applied to actual lymph node metastasis in stomach cancer, it was possible to obtain analysis results that were close to those obtained by a experienced pathologist[18](Fig.18). Currently, we are collaborating with university hospitals and cancer centers with the goal of setting up a support system for pathological diagnosis.

5.7 Time series data analysis

In general, sensing data, not limited to images, is represented by N-dimensional (Ch.) time series data, $\{s_i(t)\}_{i=1}^{N}$, $t = 1, \ldots, M$. Although it is possible to consider these as an $N \times M$ two-dimensional matrix(image) and extract HLAC features, the order of the dimension (Ch.) subscript i is generally optional. Accordingly,if for example a combination of any three is taken arbitrarily, this gives

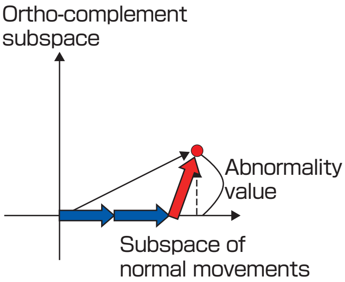


Fig. 16 Deviation from the subspace of usual motions

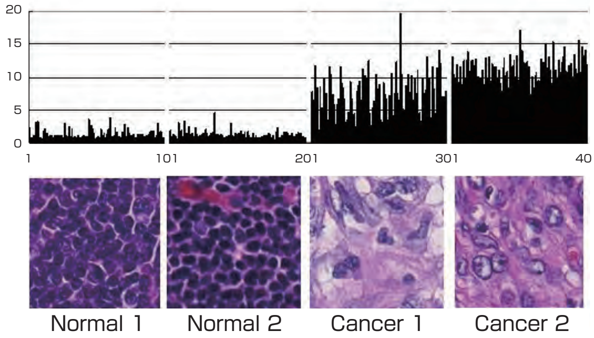


Fig. 18 Example of application to cancer detection (The upper figure shows abnormal values for each specimen.)[18]

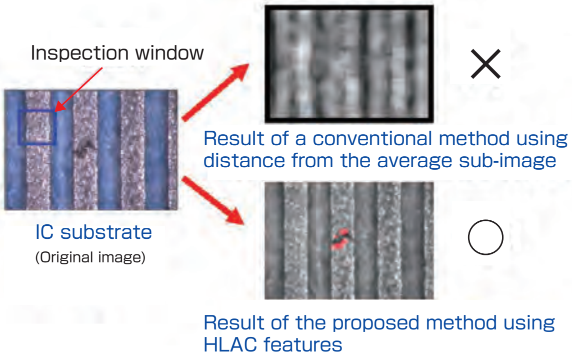

Inspection window
×
Result of a conventional method using distance from the average sub-image
IC substrate
(Original image)
○
Result of the proposed method using HLAC features

Fig. 17 Example of application to substrate inspection

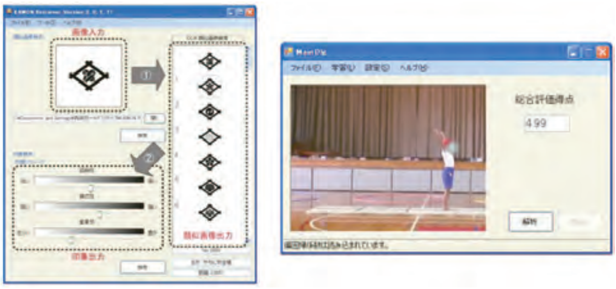


Fig. 19 Correspondence learning[24][25]

$K={}_NC_3$ two-dimensional ($3\times M$) matrices, and taking HLAC features from each 3 × 3, a feature vector with K × HLAC dimensions is obtained. By performing multivariate analysis on this (PCA, DA), analysis of such time series can be conducted (abnormality detection and discrimination). This method has been applied to abnormality detection in electrocardiograms[19] and in the analysis of the movement of a multi-fingered robot hand with multiple degrees of freedom[20].

Moreover, causal relationships, which can be interpreted as the asymmetric interrelationship (correlation) among time series data, are important in many fields. Granger Causality[21] has been proposed as an analysis index using a linear auto-regression model, but the present paper expands this model to a polynomial auto-regression model[22] (which therefore involves Higher-order Local Auto-Correlation features). Furthermore, by introducing a weighting function *w(t)* (a Causality Marker) to indicate the existence of a causal relationship, we have proposed a method that automatically extracts where a causal relationship exists[23].

5.8 Correspondence learning

Correspondence learning is connected to a wide range of common applications. Retrieval through impressions and interactive searching as well as automatic evaluation (prediction) become possible by approximating (canonical correlation analysis (CCA) or multiple regression analysis (MRA)) through learning the correspondence between the expression of a person's judgment or evaluation (external criteria) toward a pattern (still or moving images), such as the qualitative expression in the form of keywords or impression (sensitivity) vocabulary $\boldsymbol{y}^{\mathrm{Note7}}$ or rating y, and the feature vector expression (HLAC/CHLAC) $\boldsymbol{x}$ of a pattern. Figure 19 shows applications to retrieving of family crests by impressions (CCA)[24], and to the automatic evaluation of exercise (MRA)[25].

The former has been further applied to general image annotation and retrieval[26], and the latter to the automatic indexing of sports video images[27], as well as to the judgment of beef meat quality (BMS) based on ultrasound video images[28].

6 Effectiveness of the theoretical approach

This paper has thus far given an outline of an Adaptive Recognition for General-Use System (ARGUS) constructed to fulfill basic required conditions, based on feature extraction theory in pattern recognition. This paper has also discussed the system's application, focusing on a variety of practical applications in visual systems.

Unlike the scientific approach in physics and chemistry (elucidation of phenomena), in engineering applications, and in particular information technology, construction methods designed for realizing functionality are highly flexible and tend to become ad hoc and arbitrary. Thus, it is important to design proper and novel solutions from a theoretical perspective based on the fundamental requirement conditions of application demand.

By considering the fundamental framework of pattern recognition based on a theoretical standpoint, the method in this paper comprises a twostage method of Higher-order Local Auto-Correlation (HLAC/CHLAC), which is a geometrical invariant feature extraction, and multivariate data analysis, which is a statistical discriminant feature extraction. By using the latter, it is possible to learn from examples appropriate to the task. The method requires neither any model of the object nor prior knowledge, and the shape and movement of the object pattern are distinguished as points in discriminant feature space. Since segmentation of the object is also unnecessary and the computation is small with a fixed quantity of the sum of products, even moving images can be processed at far greater speeds than real-time operation on a normal PC. The features of this method are as follows:

- non-model base methodology → high versatility.
- basic initial features (HLAC/CHLAC) → applicable to a wide range of data formats.
- statistical learning (MDA) → task adaptability and increased accuracy.
- parallel sum of products operations → possible to process large amounts of data at high speed.

Almost as expected, through a variety of applications, it has outperformed the schemes that have been developed thus far. This can be attributed greatly to the method that is substantiated by theory, in particular the predominance of the Higher-order Local Auto-Correlation features and its essence. In contrast to being restricted to a two-point relationship of usual autocorrelation, by increasing to higher orders of threepoint relationship, the features obtained have become specific, e.g., curvature (convexity/concavity) rather than local straight-line direction for a contour in still images, and acceleration rather than velocity in moving images. These basic and essential initial features do not use an arbitrary iterative procedure or logical decisions (such as threshold processing and conditional branching, etc.). Rather, they use a multivariate data analysis technique and are integrated into new effective features in a parallel and comprehensive manner, forming a robust system with low information loss.

HLAC/CHLAC are fundamental general-purpose features, i.e., statistics (correlation and frequency) of spatio-temporally localized "patterns". In that sense, this constitutes a precedent for such trends as "from a model collation base to local feature statistics," which are representative of the recent HOG and SIFT features. In addition, not limited to images, it is widely applicable to the multi-channel time series data for audio and various kinds of sensor information and the like, as well as to general three-way data. Future goals include extension from quantitative data to qualitative (categorical) data, and the development of technique is already underway[29].

The application of this method is expected in a wide range of computer vision applications, such as automatic (unattended) video surveillance for intelligent security cameras, various appearance inspection systems, image annotation and retrieval, robot vision, motion analysis, and evaluation in sports and rehabilitation. At the moment, we are promoting its medical application through collaborative research with university hospitals and cancer centers, specifically toward an automatic inspection system for cancer using microscopic images. In addition, centering on an AIST-approved venture (United Technologies Institute), applications are being developed for the commercial viability of semiconductor substrate inspection and various kinds of visual inspection for agriculture and livestock fields, including inspection of rice quality and forecast of estrus and delivery in milk cows in local consortium projects.

The practical application of this method requires adjustments such as pre-processing and parameter tuning (correlation width). Future topics include automation of those settings, accumulating such knowledge base.

Acknowledgements

This study significantly progressed under the guidance of graduation and master's degree theses at the Department of Mechano-Informatics, the University of Tokyo, where I (the author) served concurrently. I hereby offer my appreciation to my dear students. I am especially indebted to the great contribution made by Dr. Takumi Kobayashi (currently in the Information Technology Research Institute, AIST) for the subsequent collaborative research and wish to offer my appreciation. I am also indebted to Dr. Takio Kurita (currently in Hiroshima University) for his work in the early HLAC experiments, and to various postdoctoral fellows as well as the Dr. Sakaue group and the Dr. Higuchi group for their contributions in recent years to the development of applications. Lastly, I would like to thank all of those concerned.

Notes

Note 1) Initially, it was intended to be called ARGUS (Adaptive Recognition for General Use System), after the giant of Greek mythology with a hundred eyes. In recent years, although HLAC/CHLAC has often been used as an abbreviation for this methodology, this actually refers to the first-stage feature extraction, and therefore is not appropriate. As such, the system/methodology as a whole will be referred to as ARGUS.

Note 2) HLAC features of a binary image are closely related to the image spectra due to the N +1th vector in perceptrons[10]. Here, combinations of "black" (1) and "white" (0) have been considered, and at first glance, it may appear that our approach that considers only "black" would not be sufficient, but it actually is. For example, ■□ , with $f_0 = f(\boldsymbol{r}) = 1$ and $f_1 = f(\boldsymbol{r}+\boldsymbol{a}_1) = 1$, and logically $f_0 \cdot f_1 = f_0 \cdot (1-f_1) = f_0 - f_0 \cdot f_1$, is represented in the range of the linear sum of feature values due to the masks (No.1 and No.3).

Note 3) This method was proposed[2][3] prior to the back propagationlearning method[7] in neural networks.

Note 4) HLAC features count the number of those topological geometry elements and their coefficients are adaptively determined in the second stage by multiple regression.

Note 5) Minimum Distance Decision: the method whereby the distance from the unknown input feature vector to the center of each class is measured, and the class with the shortest is identified.

Note 6) Helios Watcher (KK Hitachi Building Systems), http://www.hbs.co.jp/lineup/elevator/hw_outline.html

Note 7) A vector with elements of 1or 0 to express positive or negative response for each corresponding word.

References

[1] N. Otsu, T. Kurita and I. Sekita: *Pattern Recognition - Theory and Application, Behaviormetrics Series 12*, Asakura Shoten, Tokyo (1996) (in Japanese).

[2] N. Otsu: Mathematical studies on feature extraction in pattern recognition, *Research Report of Electrotechnical Laboratory*, No.818, 210 pages (1981) (in Japanese).

[3] N. Otsu, T. Shimada and S. Mori: Shape feature extraction using *N*th order auto-correlation mask, *IECE Technical Report*, PRL-78 (31) (1978) (in Japanese).

[4] 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).

[5] N. Otsu: Towards flexible and intelligent vision systems - from thresholding to CHLAC, *Proc. IAPR Conf. on Machine Vision Applications, Invited talk*, 430-439 (2005).

[6] N. Otsu: CHLAC approach to flexible and intelligent vision systems, *Proc. ECSIS and IEEE Symposium on Bio-inspired Learning and Intelligent Systems for Security (BLISS 2008), Invited talk*, 23-33 (2008).

[7] D. Rumelhart, G. Hinton and R. Williams: Learning representations by back-propagating errors, *Nature*, 323 (9), 533-536 (1986).

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

[9] J. Mclaughlin, J. Raviv: *N* th-order autocorrelations in pattern recognition, *Information and Control*, 12, 121-142 (1968).

[10] M. Misky and S. Papert: *Perceptrons*, The MIT Press (1969).

[11] T. Kobayashi and N. Otsu: Action and simultaneous multiple-person identification using Cubic Higherorder Local Auto-Correlation, *Proc. 17th Int. Conf. on Pattern Recognition (ICPR)*, 741-744 (2004).

[12] T. Kurita and N. Otsu: Face recognition method using higher order local autocorrelation and multivariate analysis, *Proc. 11th ICPR*, 213-216 (1992).

[13] F. Goudail, E. Lange, T. Iwamoto, K. Kyuma, N. Otsu: Face recognition system using local autocorrelations and multi-scale integration, *IEEE Trans. PAMI*, 18, 1024-1028 (1996).

[14] M. Lyons and S. Akamatsu: Coding facial expressions with gabor wavelets, *Proc. 3rd IEEE Int. Conf. on Automatic Face and Gesture Recognition (FG1998)*, 200-205 (1998).

[15] Y. Shinohara and N. Otsu: Facial expression recognition using fisher weight maps, *Proc. 6th IEEE Int. Conf. on Automatic Face and Gesture Recognition (FG2004)*, 499-504 (2004).

[16] T. Kobayashi and N. Otsu: A three-way auto-correlation based approach to human identification by gait, *Proc.6th IEEE Int. Workshop on Visual Surveillance*, 185-192(2006).

[17] T. Nanri and N.Otsu: Unsupervised abnormality detection in video surveillance, *Proc. IAPR Conf. on Machine Vision Applications*, 574-577 (2005).

[18] K. Iwata, et al: Cancer pathological diagnostic imaging support system using Higher-order Local Auto-Correlation feature method, *Proc. ViEW2009*, B-6H, 32-37 (2009) (in Japanese).

[19] H. Araki, M. Murakawa, T. Kobayashi, T. Higuchi, I. Kubota and N. Otsu: Abnormality detection in multichannel time-series data using Higher-order Local Auto-Correlation features, *Journal of the IEEJ*, 129 (7), 1305-1310 (2009) (in Japanese).

[20] R. Fukano, Y. Kuniyoshi, T. Otani, T. Kobayashi and N. Otsu:Acquisition of unknown object property for manipulation by a compliant multi-fingered hand, *Journal of Robotics and Mechatoronics*, 17 (6), 645-654 (2005).

[21] C. W. J. Granger: Investigating causal relations by econometric models and cross-spectral methods, *Econometrica*, 37, 424 (1969).

[22] K. Ishiguro, N. Otsu, M. Lungarella and Y. Kuniyoshi: Comparison of nonlinear granger causality extensions for low-dimensional systems, *Physical Review E*, 77 (3) (2008).

[23] K. Ishiguro, N. Otsu, M. Lungarella and Y. Kuniyoshi: Detecting direction of causal interactions between dynamically coupled signals, *Physical Review E*, 77 (2) (2008).

[24] T. Kobayashi, K. Morisaki and N. Otsu: Evaluation of image retrieval performance by using subjective impression, *Journal of the IEICE*, J91-D (4), 1025-1032 (2008) (in Japanese).

[25] Y. Morishita, T. Kobayashi, K. Morisaki and N. Otsu: A method of motion evaluation using time weights and external criteria, *Technology Research Report of the IEICE*, PRMU-107 (539), 371-376 (2008) (in Japanese).

[26] H. Nakayama, T. Harada, Y. Kuniyoshi and N. Otsu: High-performance image annotation and retrieval for weakly labeled images, *Proc. Pacific-Rim Conf. on Multimedia*, 601-610 (2008).

[27] F. Yoshikawa, T. Kobayashi, K. Watanabe and N. Otsu: Start and end point detection of weightlifting motion using CHLAC and MRA, *Proc. 1st Int. Workshop on Bio-inspired Human-Machine Interfaces and Healthcare Applications*, 44-50 (2010).

[28] T. Kobayashi, K. Watanabe, T. Higuchi, T. Miyajima and N. Otsu: Recognition of dynamic texture patterns using CHLAC features and linear regression, *International Journal of Database Theory and Application*, 2 (4), 13-26 (2009).

[29] T. Kobayashi and N. Otsu: Image feature extraction using gradient local auto-correlations, *Proc. European Conf. on Computer Vision* (ECCV), 346-358 (2008).

Author

Nobuyuki OTSU

Completed MSc course major in mathematical engineering at the Department of Mathematical Engineering and Information Physics of the University of Tokyo in March 1971. Joined the Electrotechnical Laboratory (ETL) in April of the same year. Was engaged in the research of pattern recognition theory and its application, especially feature extraction theory and image recogonition. Doctor of Engineering. Head of the Mathematical Informatics Section from April 1985, then Chief Senior Researcher in April 1990, was appointed Director of the Machine Understanding Division in April 1991. Developed the Real World Computing (RWC) project (1992-2001) and especially promoted the R&D of Real World Intelligence. Fellow of the AIST (National Institute of Advanced Industrial Science and Technology) since 2001. Held a concurrent post as Professor at the Cooperative Graduate School, University of Tsukuba from April 1992 through March 2010. From April 2001 to March 2007 also held a concurrent post as Professor at the Graduate School of Information Science and Technology of the University of Tokyo.

Discussions with Reviewers

1 Expansion of the theory and application to the industrial world

Question (Motoyuki Akamatsu, Human Technology Research Institute, AIST)

I understand from the fact that ARGUS was a robust method backed by theory, that the technique could be widely applied. As a *Synthesiology* paper about research based on such a theory, could you perhaps write about essential points or difficulties regarding the research of this theoretical basis? Also, after trying a variety of applications, could you possibly record if they generally went according to theory, or if not, whether you experienced difficulties? If it is the former case, it would greatly help the readers if you could explain why the theory went well.

Answer (Nobuyuki Otsu)

I have responded to the extent possible.

2 Selection of elemental technologies

Question (Kanji Ueda, AIST)

This paper, being *Type 2 Basic Research* for a theoretically based technique which applies to real problems, is of a type that had not yet appeared in *Synthesiology.* I would like to ask about the selection of elemental technologies for this theoretically based constitutive research. How did you choose components to achieve a practical target? Please explain whether they are just components derived by deduction from existing states, or if there are hypothetical components.

Question (Motoyuki Akamatsu)

In subchapter 4.1, you discuss how you developed an adaptive general-use image recognition system as a system satisfying the required conditions from R1 to R3, and how you adopted HLAC and CHLAC as a technique to extract feature values satisfying shift-invariance. In the process of their adoption, I believe there were other techniques considered as candidates. Could you write the rationale for how you came to the conclusion that, compared to those other techniques, HLAC is superior? Also, what is written here as a reason why HLAC was adopted, is the point that patterns are localized and their localized relative relationship is essential. I am afraid readers not specialized in this field may not immediately understand the relationship between focusing solely on local features and shift-invariance. It would be helfpful if you could add a little postscript .

Answer (Nobuyuki Otsu)

In a recognition system, the feature extraction from the object pattern is an important component in determining the performance. In contrast to choosing a variety in an ad hoc way (hypothetically so to speak or by trial and error), as has been the case up to now, higher-order local auto-correlation and multivariate data analysis were adopted as concrete components, which from a theoretical basis gives a two-stage framework comprising geometrical invariant feature extraction and statistical discriminant feature extraction, and which satisfies the basic required three conditions for achieving practical implementation objectives. In that sense, one can think of them as components derived in a deductive manner from theory, and also as hypothetical yet essential components that satisfy both theory and requirements.

There actually are not many alternatives for features that simultaneously satisfy the basic required conditions (especially R1 and R2), and yet are generic features not based on any model. As you pointed out, simply examining local features does not imply shift-invariance. Rather, because “relative” relationships are extracted as autocorrelation, this implies shift-invariance. I have supplemented the explanation as much as possible within the given space.

3 Requirements for vision systems

Question (Motoyuki Akamatsu)

As basic conditions required in a vision system, you listed “R1: shift-invariance, R2: frame-additivity, R3: adaptive trainability,” but the grounds for citing these was not clearly written. Could you please write the scenario in which these theoretical developments were selected? Also for geometrical invariance, I believe that other choices could also be considered, such as invariance with size, invariance with inclination, and relative position invariance between features. Moreover, for invariant feature extraction, it is mentioned that the functional is investigated which gives feature values invariant under geometrical transformations. Is it correct to understand that since this targets a vision system, geometrical invariance is an essentially important property? Finally, about frame-additivity, I do not think additivity will be satisfied in the case where there is overlapping, so is this a choice made mainly from processing time?

Answer (Nobuyuki Otsu)

The shift-invariance refers to invariance under a parallel shift. This does not mean that “the distance between camera and physical object hardly changes,” but rather that due to changes in the camera direction, the physical object undergoes a geometrical transformation which is a parallel shift within the screen frame, and its position changes, and that features which are invariant to such kinds of basic translation are essential in recognition. Of course, as you pointed out, other size (scale) transformations and rotations can be considered as invariant transformations, but what I am saying here is that the parallel shift (or position) invariance is the most fundamental. To avoid any misunderstanding, I have made a slight revision. The invariant feature extraction theory that seeks features invariant under geometrical transformations (functionals) is not something which is restricted to vision but also includes audio signals, and is a theory which we can generally consider as universal.

Frame-additivity, as you mentioned, does not strictly hold true for cases of overlapping, but I will risk asserting that it is important to leave the requirements as they are even in those cases. This, as you have pointed out, has implications from a processing time viewpoint, but the feature representation is a convenient one (linear) in terms of recognition (especially enumeration), and also the required condition to make the subsequent processing simple. I have supplemented the explanation.

4 The meaning of adaptive learning

Question (Kanji Ueda)

There could be several rules in using the word “adaptive learning,” but could you clarify its meaning in the context of this paper?

Answer (Nobuyuki Otsu)

To start, the prerequisite information in pattern recognition is not perfect. Based only on a finite number of examples given as learning samples, recognition is conducted on unknown test samples (an infinite number if possible). As you pointed out, there is certainly some ambiguity in the terminology “adaptive learning.” First, even if the pattern recognition is limited to the recognition object, there is adaption according to variations in the pattern. This is related to feature extraction and the learning process. Also, the adaptive learning in this paper is used in a meta-sense in that it is adaptive learning to a given recognition task. In the case of model-based learning, the model needs to be replaced when the task changes, whereas this method adapts to the task, with no model required at all and the components as they are, and it is optimally constructed (using weights) by learning through examples for the multivariate data analysis technique, the statistical feature extraction of the latter stage. I have thus revised this area to make it a little easier to understand.

5 Percentage of correct answers in pattern recognition

Question (Kanji Ueda)

Why does the percentage of correct answers not achieve 100 %? Or, in what kind of cases is 100 % possible? Recognizing fully that your research gives superior results compared to other researchers and researches until date, this is a question designed to deepen the discussion of the research as a *Synthesiology* paper.

Answer (Nobuyuki Otsu)

Real-world patterns, e.g., for an “a/i” in audio, or a “dog/cat” in an image, have diverse variations and noise, and feature (observed) values from them, such as frequency or color, are generally distributed stochastically. Using this visualization, even when categories are discriminative,the bases of those distributions near to as much extent as possible and could overlap at their frontal boundaries. Therefore, it is normal if 100 % is not achieved even for learning samples. Of course, the more the valid features are extracted in large numbers and integrated, the more close to 100 % it will approach asymptotically, However,it is more realistic to keep the feature extraction down to a finite number from a cost standpoint. It is a question of cost effectiveness.

If it is a simple identification problem, then apparently 100 % correct answer is possible. For example, in the identification and classification of 100 yen and 10 yen coins by using their feature values (e.g., diameter and weight), the fact that they are definitely different as a rule, allows for its implementation in vending machines (though sometimes there are erroneous recognitions). This paper has presented a scheme incorporating a general purpose approach aimed at more difficult, advanced recognition problems.

6 Application examples

Question (Motoyuki Akamatsu)

8 application examples were discussed in Chapter 5, and I understand that the argument as a general purpose system is based on such examples, However, in these examples, HLAC and CHLAC are the only ones that are common, and for discriminant feature extraction by multivariate data analysis, different techniques have been used, namely factor analysis, multiple regression analysis, discriminant analysis, *k*-NN classification, principal component analysis, AR model, and canonical correlation analysis among others. Although there are partial explanations such as what technique is optimum for each task, I look forward to an organized description of the basic thinking and theory behind the proper use of techniques depending on the task. I believe this would promote the understanding of the reader in terms of which technique to apply in solving their task.

Answer (Nobuyuki Otsu)

As you pointed out, I have used HLAC/CHLAC features as basic initial features (invariant features), and various multivariate data analyses for their optimum integration (linear weighted sum) corresponding to the task. As I think your concern is important in helping understand the readers who are unfamiliar with multivariate data analysis, I revised to include a correspondence table that had previously been omitted owing to limited space.