6 Pattern Properties, Invariance and Classification

Helmut Glünder

Introduction

Most often, pattern recognition is identified with what is called classification, i.e. with the process of making decisions about the membership of items to given classes of items. Of course, classes must be defined in terms of directly measurable properties of the items of interest. Classification means to compare individual data of items with the definitions of all classes and to decide which definition applies best to the data. In monocular vision, data acquisition from the static three-dimensional world is restricted to low-pass filtered and geometrically distorted central projections onto a two-dimensional surface (retina) which is spatially sampled by a finite number of non-linear sensor elements of limited sensitivity, as well as restricted dynamic and colour-spectral range. Thus, the accessible amount of information about static items in the field-of-view is completely contained in the signal values that are delivered by the ensemble of sensor elements. The so-called retinotopic arrangement of these values or components (picture elements or pixels) represents a pictorial pattern signal.

Differing from the initial view, one should state more appropriately that: "The main task in pattern recognition is to define pattern classes by means of configurations and suitable combinations of a given number of signal components."

Three idealized approaches to pattern recognition may be distinguished which essentially differ in how pattern classes are defined.

1. Template matching, in the restricted sense, deals with classes, each of which consists of a single known pattern signal. These classes are directly defined by specific configurations of all signal components, namely by the corresponding pictorial signals themselves.

2. Statistical pattern recognition is based on classes, each of which is constituted by a prototype pattern signal and versions which result from adding noise of known properties to the prototype. In order to avoid false classification, members of such classes must remain sufficiently similar in the statistical sense. The tolerable intraclass variations mainly depend on the task-specific interclass similarities. This kind of classes are most often defined by weighted sums of their members, or by those of representative samples. The resulting mean signals can be imagined as somewhat "blurred" versions of the corresponding pictorial prototype signals.

3. Geometrically invariant pattern recognition is based on classes, each of which ideally comprises only one pattern, i.e. a pattern signal and its well-defined geometrically transformed versions. Members of such classes show common properties in the deterministic sense. Although the geometric transformations may be confined to certain types, they are not assumed *a priori* to be restricted to certain ranges of their parameters. Under such highly variant conditions it is no longer possible to base acceptable class definitions on the mean signals of classes, since the resulting blur will mainly leave the mean value of each type of signal as the remaining property which is anything but a characteristic pattern feature. Real pattern classes in the above-mentioned sense can be solely defined by non-linear combinations of signal components.

In practice, combinations of these fundamental approaches are applied to pattern recognition problems. The choice of class definitions depends on parameters such as the number of signal components and classes, the noise statistics, the types and extent of geometric variance, and the percentage of tolerable false classification. As usual, economic reasons, i.e. limited resources, largely determine the character and scope of class definitions. In conjunction with restrictions imposed by the data acquisition process, class definitions cause categorial interpretations, i.e. "filtered views", of the actual world.

Until now, technical pattern recognition, such as character recognition, deals with problems that are characterized by typically small and predetermined numbers of classes, minor geometric variance and preferably moder-

ate noise. Thereby, costs can be held low compared with biological pattern recognition which deals with an enormous variety of potential classes, extreme geometric variance and sometimes considerable noise. The maximum number of signals consisting of n components, each of m levels, is $K = m^n$ which, for a modest 5×4 sensor-array, is more than a million binary signals ($m = 2$). The resulting tremendous amount of possible templates – and thus classes – especially for large arrays, is slightly reduced if invariant pattern recognition is considered. Unrestricted shift invariance, for instance, leads to a reduction factor of about $1/n$. More drastic reductions, however, are due to the fact that, in general, only those patterns which show pronounced spatial bindings are biologically relevant.

The following considerations start from the most basic form of classification that permits unequivocal signal reconstruction and which is applicable if all signals of interest are known, i.e. if generalization is not required. It will be demonstrated that one can eliminate this constraint if one accepts less restricted class definitions. The main part deals with methods of geometrically invariant classification that permit ideal or nearly ideal class definitions. Translation invariant versions of the approaches are discussed and network structures for their parallel implementation are introduced. As far as possible, statements are illustrated by simple examples. A discussion of the self-organization and learning procedures of networks lies beyond the scope of the intended investigation concerning the class definitions and costs of invariant classifier structures. Learning and adaptation procedures are documented, for instance in the books of Schürmann (1977), Rumelhart and McClelland (1986), and Pao (1989). In a final section, the results are summarized and assessed from technical and biological points of view. Problems associated with noisy signals, i.e. all aspects of statistical pattern recognition are not treated in detail. This field, however, is well represented in the literature (Sebestyen, 1962; Fukunaga, 1972; Schürmann, 1977). In the next section fundamental formalisms are explained, thereby introducing the inevitable terminology of (pictorial) pattern recognition.

Signal Representations and Signal Similarity

In order to circumvent most of the problems associated with image-data acquisition and preprocessing, all considerations are based on pictorial signal representations of $n = \mu_{\max} \times \nu_{\max}$ spatially discrete and real-valued signal components. Number n denotes the degrees of freedom of a signal, i.e. its dimensions. For components of m equally distributed (quantization) levels, the entropy of a signal representation is:

$$H_{\max} = n\,\mathrm{ld}(m) = \mathrm{ld}(K) \qquad \text{in [bit]} \tag{6.1}$$

Maximum entropy, or maximum number of possible signals K, is assumed constant when assessing or comparing pattern recognition approaches.

Correct sampling of real-world scenes generally requires analogue low-pass filtering (sampling theorem). Therefore, a single bright point in the world that may be up to a pixel in size, is blurred and consequently represented by several pixels of various grey values. In practice, this point spread is spatially restricted by the finite accuracy $1/m$ of the amplitude measurements. As a consequence a "correctly sampled binary signal" is a contradiction. Due to mostly unknown global variations in illumination, the signal mean, i.e. the mean intensity or grey value of an image, is of minor importance for pattern classification in biological, as well as machine vision. For this reason, the mean is often subtracted from the signal. The resulting bipolar image representations are advantageous with respect to amplitude dynamics whenever multiplicative comparisons (correlations) are applied. Even more convenient are isotropically band-pass filtered pictorial signals in which all parts of constant intensity are set to zero, and derived versions (Marr, 1982; Watt 1988), as well as contour representations. Despite these reasonable assumptions, the accompanying examples deal with signals that are distributions of ones and zeros. Furthermore, a toroidal image array is assumed, i.e. the top row of the pictorial representation is connected to the bottom row and its left side column to the right side one. Thereby, the effect of signal variance can be studied even within very small arrays. These idealizations are introduced to help with concentrating on the essential issues of the examples.

For the purpose of signal analysis or classification it is sometimes advantageous to consider a pattern signal $x(\mu,\nu)$ as a vector $\vec{x}$ in an n-dimensional and orthonormal vector space.

$$\begin{aligned}\vec{x} &= \left(x(1,1), x(1,2), \ldots x(\mu,\nu), \ldots x(\mu_{\max},\nu_{\max})\right)^{\mathrm{T}} \\ &= \left(x_1, x_2, \ldots x_i, \ldots x_n\right)^{\mathrm{T}} \qquad \text{with } n = \mu_{\max} \cdot \nu_{\max}\end{aligned} \tag{6.2}$$

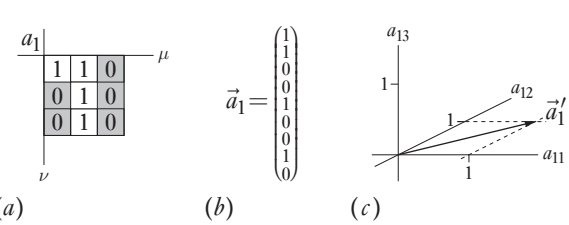

Fig. 6.1 *Pictorial pattern signal $a_1(\mu,\nu)$ in a 3×3 image-array (a), in vector notation (b), and its first three components in signal space (c).*

Every signal is represented by a single point in this n-dimensional space. If the components x_i result from basic measurements, i.e. if they are for example the values of pixels, then this space is called a signal space. It is called feature space if they are derived quantities.

For example, Fig. 6.1(a) shows the pictorial representation $a_1(\mu,\nu)$ of one of the $K = 512$ possible binary signals of $n = 3\times 3$ pixels. Figure 6.1(b) presents the corresponding nine-dimensional signal vector $\vec{a}_1$. Due to the problems associated with the drawing of a nine-dimensional space, merely a vector $\vec{a}_1' = (a_{11}, a_{12}, a_{13})^{\mathrm{T}}$ consisting of the first three components of vector $\vec{a}_1$ is sketched in Fig. 6.1(c).

The similarity of signals is evaluated through suitable comparison operations that are defined between the corresponding signal components of two signals, such as the squared difference, the absolute value of difference, or simply the product. For 0/1-signals the first two operations reduce to the logical NAND, and the third to the logical AND operation. Although not always optimum, multiplications are most often used and lead to simple mathematical formulations; they are considered here as well. Since it is desirable to characterize signal similarity by a single number, the sum of all elementary comparisons, i.e. products, must be computed. This yields the cross-correlation coefficient or inner product y' of two signal vectors $\vec{x}_1$ and $\vec{x}_2$.

$$\begin{aligned} y' &= \vec{x}_1^{\mathrm{T}} \cdot \vec{x}_2 = \sum_{i=1}^{n} x_{1i} \cdot x_{2i} \\ &= \sum_{\mu=1}^{\mu_{\max}} \sum_{\nu=1}^{\nu_{\max}} x_1(\mu,\nu) \cdot x_2(\mu,\nu) \end{aligned} \qquad (6.3)$$

for real-valued $\vec{x}_1$ and $\vec{x}_2$

Unfortunately, this measure of similarity suffers from what is called form/intensity crosstalk, i.e. pattern signals of rather different form can lead to large correlation coefficients provided their intensities are sufficiently high. This ambiguity can be avoided by normalizing the cross-correlation coefficients with respect to the product of their vector lengths (standard normalization).

$$y = y' / \left(|\vec{x}_1| \cdot |\vec{x}_2| \right) \qquad (6.4)$$

Unless stated otherwise, standard normalization is presupposed for all linear classifier stages that are considered in what follows. Two signals are maximally different (orthogonal signals) if $y = 0$, and identical if $y_a := y = 1$ (normalized autocorrelation coefficient) and vice versa. This becomes evident when considering these situations in the signal space, where cross correlation means the projection of one vector onto the other. If the vectors are orthogonal the projection is zero, if they are parallel it becomes maximum. In pattern recognition, comparisons

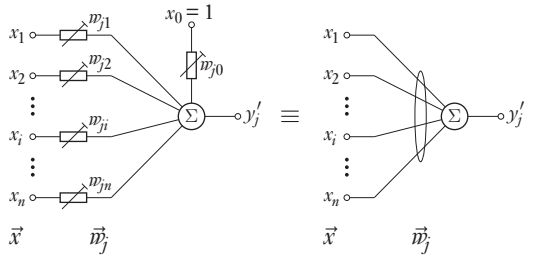

Fig. 6.2 *Circuit for the computation of cross-correlation coefficients* $y_j' = \vec{x}^{\mathrm{T}} \vec{m}_j$.

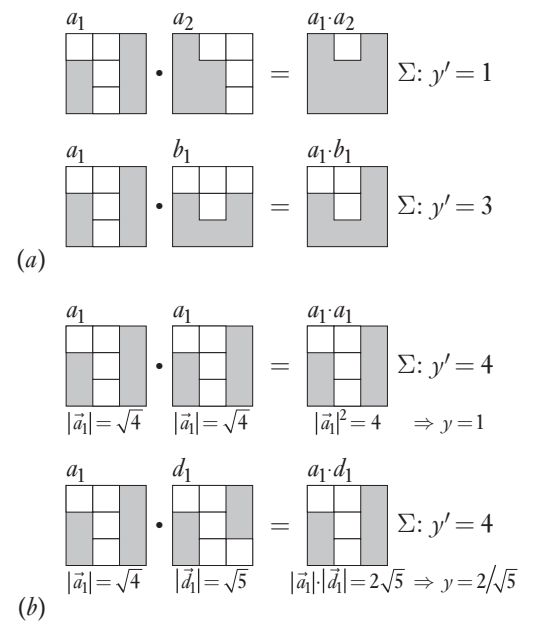

Fig. 6.3 *Computation of correlation coefficients* y' *in the pictorial domain. The advantage of normalized coefficients* y *is demonstrated in (b).*

are to be made between various n-dimensional signal vectors $\vec{x}$ and certain $(n+1)$-dimensional stored vectors $\vec{m}_j$. For this purpose, augmented $(n+1)$-dimensional signal vectors with an additional component $x_0 = 1$ are used. The sum of Equation 6.3 is then called a linear discriminant function

$$y_j' = m_{j0} + x_1 m_{j1} + x_2 m_{j2} + \ldots + x_i m_{ji} + \ldots x_n m_{jn} \qquad (6.5)$$

which can be computed in parallel within the structure shown in Fig. 6.2. At node Σ, the signal components x_i, weighted by the coefficients m_{ji}, are summed. Hence, the weighted interconnection scheme is an implementation of a stored vector $\vec{m}_j$.

Equations 6.3 and 6.4 are illustrated by correlating binary pattern signals directly in the pictorial domain. It can

be seen from Fig. 6.3(a) that the identical patterns – not pattern signals – a_1 and a_2 are more dissimilar than the signals a_1 and b_1 which are not translation equivalent. Their dissimilarity can also be expressed by the squared lengths of their difference vectors $|\vec{a}_1 - \vec{a}_2|^2 = 6$ and $|\vec{a}_1 - \vec{b}_1|^2 = 2$. This example gives a first impression of what is called the invariance problem. The necessity for normalization is demonstrated by the examples shown in Fig. 6.3(b). Normalization guarantees the unequivocal relation between the correlation coefficient $y = 1$ and the autocorrelation situation.

Template Matching and Deterministic Linear Classification

For reasons of simplicity it is assumed that, except for statistical pattern recognition, the frequencies of occurrence of every signal within a class, as well as those of all classes, are equally distributed.

Holistic template matching means to correlate a signal of interest $\vec{x}$ with k known and whole signals $\vec{m}_j$ that are stored in the recognizing system. They represent holistic templates and one may say that each of them defines a class with exactly one member. For the purpose of classification, the computation of the k correlation coefficients is followed by the detection of the autocorrelation coefficient y_a (exhaustive search). If it cannot be found, then the signal is recognized as being a member of the $k+1^{\text{st}}$ class of unknown pattern signals. Except for the latter, class definitions are ideal because every class is identical with its signal. In principle, this approach permits one to distinguish and to perfectly reconstruct the maximum number of K pattern signals. Template matching is a special case of linear classification.

Changing the decision criterion from the detection of the autocorrelation coefficient to maximum detection results in deterministic linear classification, for which the k templates represent the prototypes of k classes. Linear classification makes use of the similarities between a signal and the prototypes. The greatest similarity, expressed by the highest correlation coefficient, determines its class membership. Every pattern signal that results in a single maximum can now be classified, i.e. a class of unknown signals does not exist. Class definitions depend on the number of classes and on the similarity of their prototypes and they are the more restricted the more classes are to be considered. For $k \to K$ they approach the ideal class definitions. Owing to cross correlation which represents a linear form (cf. Equation 6.5), and to the decision criterion "maximum detection", each class is separated from all others by hyperplanes in the signal space. Maximum detection out of k correlation coefficients segregates the n-dimensional signal space, with its K possible signals, into k hypervolumes – similar to soap bubbles filling a closed box. Hence, classes are generally enclosed by differently shaped polyhedral hypersurfaces. In other words, the members assigned to every class cluster in the signal space. This property makes linear classification well-suited for the classification of noisy signals, where sufficiently similar signals are likely to belong to the same class. The linear classifier is optimum if all classes suffer from the same additive noise process. In this case of statistical pattern recognition the expected signals of the corresponding statistical signal distributions, i.e. their mean vectors or centroids in the signal space, are chosen as prototypes. Depending on the probability density function of the noise process, false classifications can occur. However, deterministic linear classification generally leads to class centroids that differ from the prototype vectors. Therefore, signal reconstruction may be carried out either according to the *a priori* prototypes, or with respect to the *a posteriori* determined centroids.

Holistic template matching, as well as deterministic linear classification with respect to k classes can be performed within networks of the type depicted in Fig. 6.4. Decision-making in a template matching classifier is per-

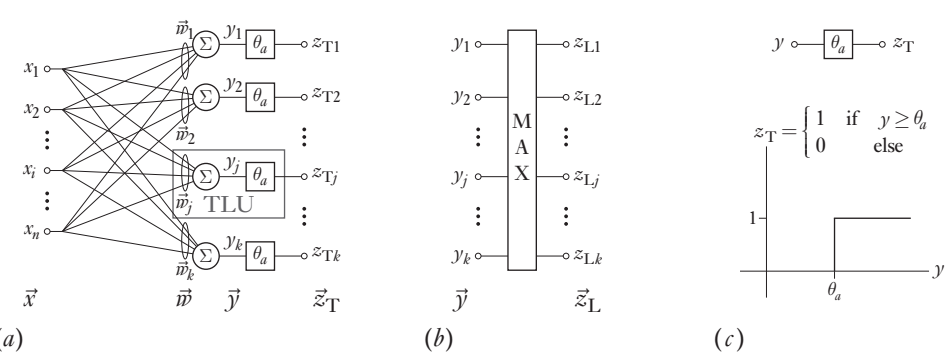

Fig. 6.4 *Structure for holistic template matching based on threshold logic units (TLUs) (a) and for general linear classification for which maximum detection (b) is applied, instead of ideal threshold characteristics (c).*

formed by k coincidence detectors that compare the correlation coefficients with the implemented autocorrelation coefficient. They can be replaced by ideal thresholds θ_a that are adjusted to the autocorrelation coefficient y_a (cf. Fig. 6.4(a)(c)). That is why the processing unit within the dashed frame in Fig. 6.4(a) is called a classic "threshold logic unit" (TLU). The much more complicated process of maximum detection (cf. Fig. 6.4(b)) that is performed by what is called "winner-take-all" or "maxnet" circuits (Feldman, 1982; Lippmann, 1987), involves operations of so-called polynomial order $q = k$ which means the nonlinear combination of all k correlation coefficients y_j (Uesaka, 1971; Hadeler, 1974).

For example, let us consider a classifier with $k = 3$ classes that are defined by the prototypes m_{a_1}, m_{b_1} and m_{c_1}. Figure 6.5 shows them (left column) together with four signals a_1, b_8, c_4 and d_1 that are to be classified (top row). The matrix contains the normalized correlation coefficients y_j. The classification results for holistic template matching and deterministic linear classification are shown below. For the latter, three of the signals are classified, while signal c_4 lies on a separating hyperplane. For the former approach, three of the pattern signals are assigned to the fourth class of unknown signals and thus they are described as being neither signal a_1, nor b_1, nor c_1. It is remarkable that signal b_8 is not assigned to the class defined by the prototype m_{b_1}, although both are identical patterns, i.e. they are translation equivalent signals.

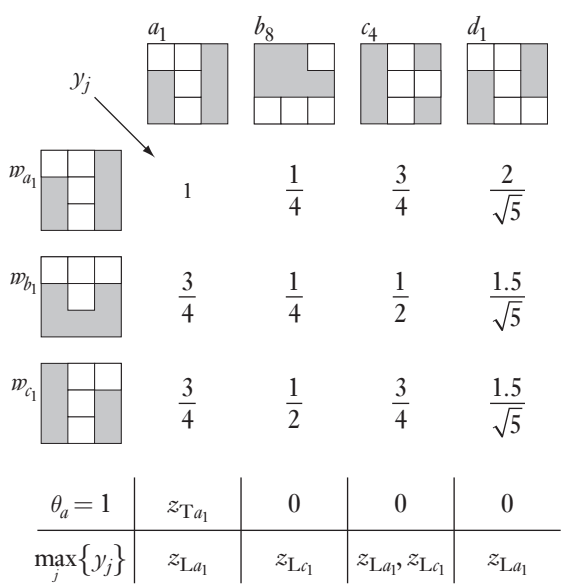

Fig. 6.5 *Assignment of four pattern signals (top row) to three classes (left column) by the classifier structures of Fig. 6.4(a)(b). Classification results are shown below (zero stands for unknown pattern).*

Geometrically Invariant Classification

After this introduction to basic mechanisms of linear signal classification, an important additional property is demanded which marks an essential difference between code deciphering and real pattern recognition, namely the invariance of classification under geometric transformations of the signals. Problems associated with reaching ideal definitions of invariant pattern classes are exemplified for unrestricted two-dimensional translational variance of pictorial pattern signals on a toroidal image array and three approaches are compared with respect to class definitions and effort. Corresponding considerations hold for unrestricted rotations and size variance that can be imagined as converted into two-dimensional translations by the so-called log-polar mapping (Brousil and Smith, 1967; Casasent and Psaltis, 1976). Comparable strategies allow for invariance and good to ideal class definitions under other types of geometric variance. The achievement of multiple invariance, for example under the group of the similitudes or even that of the affine transformations, is much more costly or leads to only moderate, and for many applications insufficient, class-definitions (Doyle, 1962; Moore and Parker, 1974; Casasent and Psaltis, 1976; Kröse, 1985; Giles and Maxwell, 1987; Glünder, 1987).

Based on the elementary examples shown in Fig. 6.3(a) and 6.5, one may speculate that cross correlative similarity measures alone do not suffice for invariant classification. Signals of the same invariant pattern class, i.e. so-called transformation equivalent signals, may even be orthogonal. Consequently, classes as defined by linear classifiers, and thus based on the similarity of their members, in general do not define invariant pattern classes. In signal space the lack of signal similarity means no more clustering of class members and, as a consequence for classification, more complicated, i.e. curved separating surfaces. The typically n members of a general translation invariant class which is defined on a toroidal image array of n pixels, are nicely distributed over an n-dimensional hypersphere – or hypercube for 0/1-signals – in signal space, with the class centroid at its centre. This implies that all members of an invariant class have the same distance $|\Delta\vec{x}_i|$ from their centroid $\vec{x}_c$ and that the members of classes with identical vector length $|\vec{x}_i|$ are intermingled on the same hypersphere. The difference vectors of a pattern

$$\Delta\vec{x}_i = \vec{x}_i - \vec{x}_c \qquad \text{with} \qquad x_c = \frac{1}{n}\sum_{i=1}^{n}\vec{x}_i \tag{6.6}$$

additionally show the same configuration in all n subspaces of equal dimensions. In other words, the signal distribution on the hypersphere "looks" the same when "viewed" from n orthogonal directions.

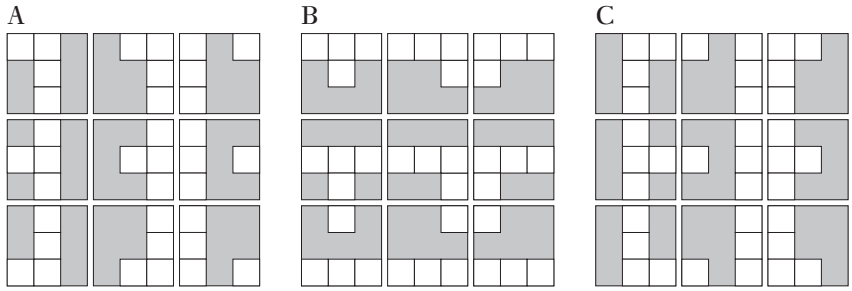

Fig. 6.6 *The nine members of three translation invariant pattern classes that are defined on a toroidal* 3×3 *image-array.*

Actually, some translation invariant pattern classes comprise less than n members, namely those exhibiting global shift invariance, such as periodic patterns and, most pronounced, structureless constants. Despite this fact, typically n members per ideal invariant pattern class are assumed for the following considerations.

For example, Fig. 6.6 shows the $n = 9$ members of each of the three translation invariant pattern classes A, B and C. These classes have the coinciding centroid vector $\vec{a}_c = \vec{b}_c = \vec{c}_c = \frac{4}{9}(1,1,1,1,1,1,1,1,1)^{\mathrm{T}}$ and, due to the equal length of their signal vectors $|\vec{a}_i| = |\vec{b}_i| = |\vec{c}_i|$, their members are dispersed over the same nine-dimensional hypersphere of radius $|\Delta\vec{a}_i| = |\Delta\vec{b}_i| = |\Delta\vec{c}_i| \approx 1.49$.

Except for trivial cases, linear classification is not suited for pattern classification that is invariant under unrestricted geometric transformations. For this purpose, more complicated separating surfaces are required which are commonly formulated as polynomials of sufficient degree and which result in polynomial discriminant functions. Instead of deforming separating hyperplanes, one can equally well apply the non-linearity to the signal components and subsequently linearly classify the signals in the thus generated feature space (Schürmann, 1977). Its ${}_p\tau$ coordinates are the terms of a complete polynomial of degree p that is computed from an n-dimensional signal vector $\vec{x}$.

$$ {}_p\tau = \binom{n+p}{p} \quad \text{or} \quad {}_p\tau \approx \frac{n^p}{p!} \quad \text{for} \quad n \gg p \tag{6.7} $$

Therefore, each TLU of the subsequent linear classifier receives $N = {}_p\tau$ input signals, i.e. N weighting coefficients must be determined for every class.

Polynomial classification appears well suited for the assessment of other approaches because discrimination of invariant classes, i.e. the definition of classes, depends on a single parameter, namely the polynomial degree p which is directly related to the costs of classification through Equation 6.7.

Direct Translation Invariant Classification

A straightforward solution for ideal translation invariant classification is the so-called invariant list classifier. It uses typically n holistic template-matching classifiers per class in parallel. Thereby, n subclasses are defined, each of which comprises exactly one member of a translation invariant pattern class. Each class-specific output is obtained by the summation of the n thresholded signals u_{ji} as depicted in Fig. 6.7(a) (grey inscriptions). The thresholds, that represent a non-linearity of theoretically infinite polynomial degree p, are adjusted to the autocorrelation coefficients (cf. Fig. 6.4(c)). Consequently, at best one single output of all kn TLUs can be activated. Each of the k linear substructures, comprising n holistic templates of the same form, represents a holistic matched filter and thus each output vector $\vec{y}_j$ is the correlation function of the signal vector $\vec{x}$ and the template vector $\vec{m}_j$. The costs for ideal invariant class definitions, i.e. for the possibility of unequivocal reconstruction of patterns irrespective of signal position, are $N \approx n^2$ weighting coefficients per class.

As with holistic template matching, only known patterns can be classified and it is therefore tempting to modify the invariant list classifier in the same way as template matching was generalized thus yielding the linear classifier (cf. Fig. 6.4). For this purpose, all threshold units in Fig. 6.7(a) are replaced by a single maximum detection unit that is followed by k class-specific summation nodes. Every class is composed of n more or less compact and isolated subspaces in signal space. Although the number of weighting coefficients equals that of the invariant list classifier, the computing effort is tremendously increased due to the maximum detection process which acts on typically $q = kn$ input signals. For tasks involving small and known numbers of invariant classes, redundant subclasses can be eliminated and manageable orders q may become attainable. In general, however, this approach is impracticable.

Up to now, two necessary and unfortunately independent conditions for translation invariant classification with ideal class definitions can be indicated:

1. Invariance demands for non-linear classification, i.e. at least for $p = 2$, which in turn does not yet permit any statements about class definitions.
2. Ideal class definitions are achieved through list classification which, due to the thresholds, is of polynomial degree $p \to \infty$. The costs are characterized by $N \approx n^2$ weighting coefficients per class.

Polynomial Approach to Pattern Decomposition

In contrast to the Aristotelian point of view, one can try to base classification on pattern descriptions that no longer consist of single correlation coefficients and their underlying holistic templates but are given in terms of compositions from pattern elements. These elements must be known to the recognizing system and are detected through cross correlation of pattern signals with the corresponding templates in conjunction with subsequent decisions. Such templates shall be called masks $\vec{h}_\chi$, in order to differentiate them from the holistic templates $\vec{m}_j$. As with list classification, each of the κ types of masks must be applied at every position of the toroidal image array, if unrestricted translation invariance is demanded. For every type of mask, the hereby introduced space invariant correlation filtering produces typically n correlation coefficients that constitute the filtered signal $\vec{y}_\chi$. Unlike normalized holistic matched filtering, the unambiguous detection of autocorrelation coefficients indicating the presence and the locations of certain pattern elements in non-binary signals, poses severe problems. Because the pattern elements are to be detected independently of the remaining pattern parts, holistic normalization according to Equation 6.4 is generally no longer applicable. Consequently, every single mask position must be individually normalized which requires κn normalization operations. For binary pattern signals, this tremendous effort is avoided if normalization of the filtered signals with respect to the mean of the binary

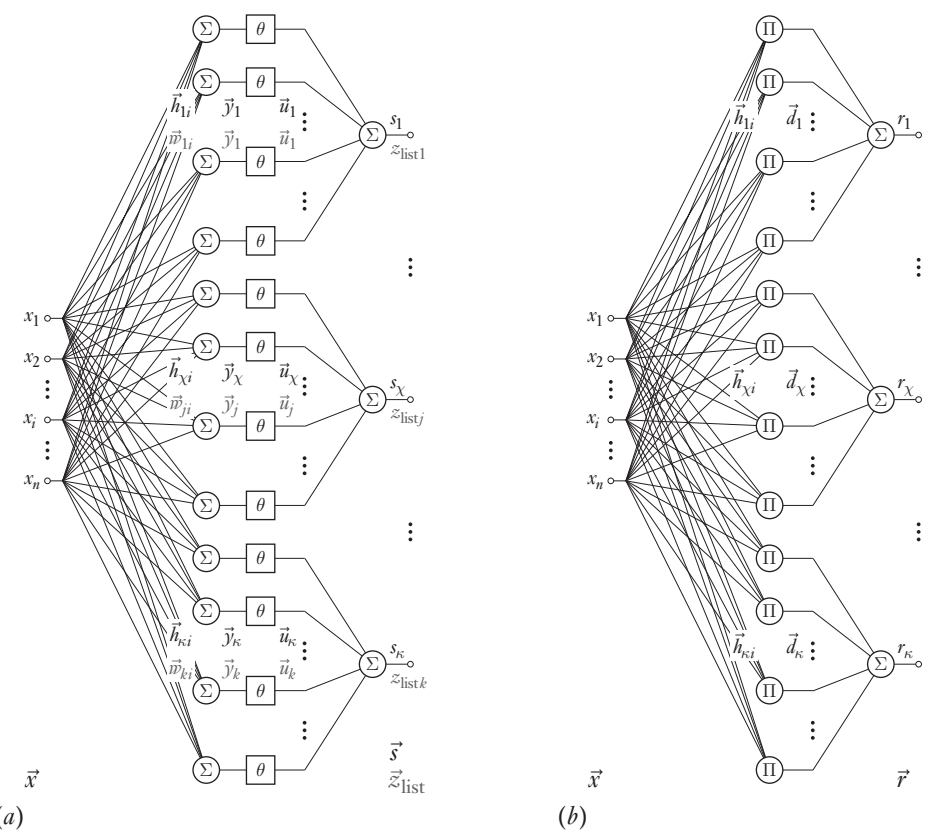

Fig. 6.7 *Network structures for list classification (grey inscriptions) and for polynomial pattern decomposition and "counting" (PDC) (black inscriptions) (a), as well as for multilinear pattern decomposition and "counting" (MDC) (b).*

masks is applied in conjunction with subsequent threshold detection. The filtered signals, and also the detected autocorrelation coefficients, still depend on the positions of the input signals, i.e. they are by no means invariant. Therefore, invariant measures, such as the number or sum of the autocorrelation coefficients that are contained in every filtered signal $\vec{y}_\chi$, must be determined. Accordingly, a set of κ types of masks $\vec{h}_\chi$ leads to a κ-dimensional vector $\vec{s}$ of translation invariant features s_χ which can be linearly classified.

Of course, the restriction to binary signals is not at all satisfying and, because the normalization problems are due to the binary character of the threshold decisions (cf. Fig. 6.4(c)), a generalization of the detection criterion to a non-linear function $u_{\chi i} = f(y_{\chi i})$ of non-negative derivative may be helpful. The counting of suprathreshold signals, i.e. the counting of "ones", is then replaced by the summation of the non-linearly weighted results from correlation filtering. If the transfer characteristic is formulated as a polynomial of degree p, then every invariant feature is a sum of typically n polynomials ${}_p u_{\chi i}$ produced from a linear discriminant function (cf. Equation 6.5).

$$ {}_p s_\chi = \sum_{i=1}^{n} {}_p u_{\chi i} = \sum_{i=1}^{n} \sum_{\rho=0}^{p} \alpha_\rho \cdot (y_{\chi i})^\rho \tag{6.8} $$

Therefore, this kind of feature extraction shall be called a polynomial approach to pattern decomposition and "counting" (PDC). Although arbitrarily valued mask coefficients are feasible, they do not make much sense without normalization of the resulting cross-correlation functions. For this reason, and in order to prevent the combinatorial explosion of mask types, 0/1-masks are considered which turns out to be an appropriate procedure (see the next section). The translation invariant feature vector generated thereby, is then fed into a linear classifier with standard normalization.

A parallel-computing structure for this kind of feature extraction is familiar from the list classifier shown in Fig. 6.7(a). The desired network is obtained by replacing the k holistic templates $\vec{m}_j$ by κ masks $\vec{h}_\chi$, and by replacing the thresholds by a single type of non-linear transfer characteristic, thus leading to generalized TLUs, and finally by recognizing the classification result $\vec{z}_{list}$ as the feature vector $\vec{s}$ (black inscriptions). The second processing stage consists of a linear classifier network (cf. Fig. 6.4) that acts on this feature vector. This series connection leads to a succession of two linear summation stages that can be combined if the vectors $\vec{u}_\chi$ are used as features for the classifier section and if one accepts the involved increase in the number of weighting coefficients by a factor n. The resulting two-stage network is known as a *Perceptron* structure with one hidden layer (Rosenblatt, 1962; Minsky and Papert, 1969; Uesaka, 1971, 1975; Lippmann, 1987).

For example, let us decompose the binary patterns A, B and C (see Fig. 6.6). Because the resulting features are translation invariant, it suffices to investigate one member of every class, e.g. a_1, b_1 and c_1. Four non-linearities are alternatively applied to the correlation coefficients: a threshold adjusted to $\theta_a = q$, with q being the number of the non-zero components of the mask vectors, as well as parabolic ${}_2u_{\chi i} = (y_{\chi i})^2$, cubic ${}_3u_{\chi i} = (y_{\chi i})^3$ and logarithmic ${}_{\ln}u_{\chi i} = \ln(1 + y_{\chi i})$ transfer characteristics. Only masks ${}^q h_\chi$ with $q = 2$ and $q = 3$ are considered. Figure 6.8(a) shows the pictorial representations of all ${}^2\kappa = 4$ translation invariant 0/1-masks ${}^2h_\chi$ that exist within the toroidal 3×3 image array. The corresponding ${}^3\kappa = 12$ types of translation invariant 0/1-masks ${}^3h_\chi$ are depicted in Fig. 6.8(b).

In Fig. 6.9 the filtered pattern signals ${}^2y_\chi$ are shown for the three not normalized pattern signals. The correspond-

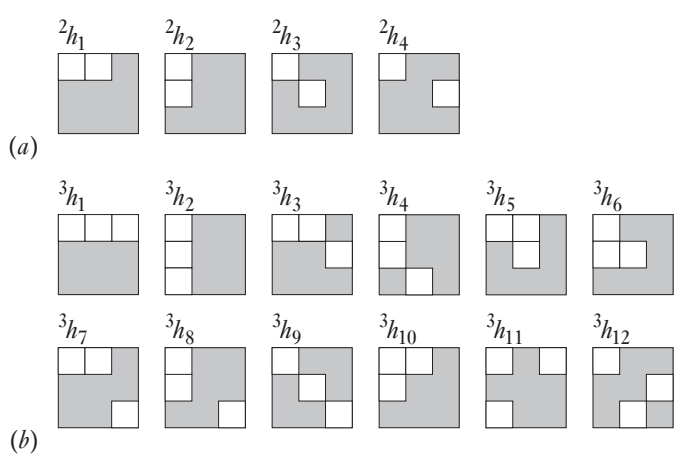

Fig. 6.8 *Pictorial representations of all translation invariant* 0/1*-masks that exist in a toroidal* 3×3 *image array and that contain two (a) and three (b) pixels of non-zero value.*

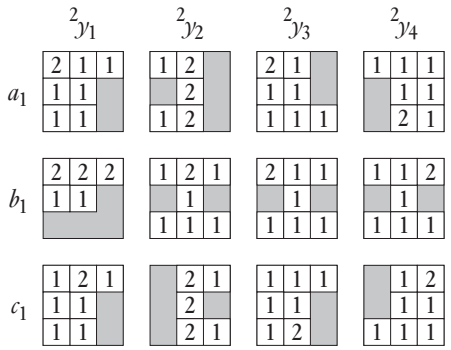

Fig. 6.9 *Correlation functions computed from three pattern signals and the masks of Fig. 6.8(a).*

	$^{2}_{\theta}\vec{s}$	$^{2}_{2}\vec{s}$	$^{2}_{3}\vec{s}$	$^{2}_{\ln}\vec{s}$	$^{3}_{\theta}\vec{s}$	$^{3}_{2}\vec{s}$	$^{3}_{3}\vec{s}$	$^{3}_{\ln}\vec{s}$	$^{2}\vec{r}$	$^{3}\vec{r}$
A	1	10	14	5.3	0	18	30	7.5	1	0
	3	14	26	4.7	3	30	84	6.2	3	3
	1	10	14	5.3	0	18	30	7.5	1	0
	1	10	14	5.3	0	22	42	6.9	1	0
					1	22	48	7.0		1
					0	22	42	6.9		0
					0	18	30	7.5		0
					1	22	48	7.0		1
					0	18	30	7.5		0
					0	22	42	6.9		0
					1	22	48	7.0		1
					0	18	30	7.5		0
B	3	14	26	4.7	3	30	84	6.2	3	3
	1	10	14	5.3	0	18	30	7.5	1	0
	1	10	14	5.3	1	22	48	7.0	1	1
	1	10	14	5.3	0	18	30	7.5	1	0
					1	22	48	7.0		1
					0	22	42	6.9		0
					0	22	42	6.9		0
					0	18	30	7.5		0
					0	18	30	7.5		0
					1	22	48	7.0		1
					0	22	42	6.9		0
					0	18	30	7.5		0
C	1	10	14	5.3	0	18	30	7.5	1	0
	3	14	26	4.7	3	30	84	6.2	3	3
	1	10	14	5.3	0	18	30	7.5	1	0
	1	10	14	5.3	1	22	48	7.0	1	1
					0	22	42	6.9		0
					1	22	48	7.0		1
					0	18	30	7.5		0
					0	22	42	6.9		0
					0	18	30	7.5		0
					1	22	48	7.0		1
					0	22	42	6.9		0
					0	18	30	7.5		0

Table 6.1 *Translation invariant feature vectors of three patterns which are computed through PDC with four different non-linearities (left section) and MDC (right section), both for polynomial orders two and three.*

ing four-dimensional feature vectors $^{2}_{\theta}\vec{s}$, for the threshold with $\theta_a = 2$, as well as $^{2}_{2}\vec{s}$, $^{2}_{3}\vec{s}$ and $^{2}_{\ln}\vec{s}$, for the other non-linearities, are represented in Table 6.1 in the left half of the left section. Obviously, pattern class A cannot be separated from class C, even if arbitrary non-linearities are applied! Class B, however, is linearly separable from the other two, either by feature $^{2}s_1$ or $^{2}s_2$, i.e. independent of the applied non-linearity. The right half of the left section of Table 6.1 contains the 12-dimensional feature vectors $^{3}_{\theta}\vec{s}$, for the threshold $\theta_a = 3$, as well as $^{3}_{2}\vec{s}$, $^{3}_{3}\vec{s}$ and $^{3}_{\ln}\vec{s}$, for the other transfer characteristics. Again class B is easily separated from the two other classes, e.g. by the first three components of all four feature vectors. The classes A and C cannot be separated on the basis of the parabolic feature vector $^{3}_{2}\vec{s}$. The segregation of all three classes is possible by one of the following features: $^{3}_{3}s_4$, $^{3}_{\ln}s_4$, $^{3}_{3}s_8$ or $^{3}_{\ln}s_8$.

Finally, the ultimate masks, namely the holistic templates $^{4}h_{a_1} = w_{a_1}$ and $^{4}h_{c_1} = w_{c_1}$, are tested for their suitability to typify the translation invariant classes A and C. As it can be seen from Fig. 6.10, the dichotomy is not achieved with a parabolic non-linearity, even for this optimum condition. However, linear separation is guaranteed for the other non-linearities and all those of higher polynomial degree than two. Obviously, thresholding results in template matching known from list classification.

A fundamental question concerns the conditions for unequivocal pattern descriptions which, in conjunction with a template matching second stage and standard normalization, allow for ideal class definitions. It is not obvious, whether they can be achieved without holistic templates and, if so, how many masks of what type are required. At least the description of patterns by the numbers of their parts, which seems to imply the loss of knowledge about the relative positions of the parts (loss of pattern coherence), casts some doubts on this goal. Empirical investigations (cf. the preceding examples) reveal that ideal translation invariant class definitions are not achieved if the number q of non-zero components of the mask vectors $^{q}\vec{h}_{\chi}$ is less than three, or if the polynomial degree p of the non-linearity is less than three, even if all $^{2}\kappa$ types of 0/1-masks that exist in the image array, are applied. As mentioned, the dimensionality κ of the feature vector determines the number of weighting coefficients per class of the subsequent classifier stage. Hence, with

$$^{q}\kappa \approx \frac{1}{n}\binom{n}{q} \quad \text{or} \quad ^{q}\kappa \approx \frac{n^{(q-1)}}{q!} \quad \text{for} \quad n \gg q \tag{6.9}$$

and leaving the number of the binary mask coefficients out of consideration, the necessary condition for translation invariant classification with ideal class definitions can now be reformulated:

$$p = 3 \quad \text{and} \quad N \approx \frac{n^2}{6} \quad \text{for} \quad n \gg 3 \tag{6.10}$$

Although more specific than the statements given at the end of the previous section, this condition does not yet clarify whether $p = 3$ and $q = 3$ are sufficient for ideal pattern reconstructions.

$a_1 \text{ corr } {}^4h_{a_1} = a_1 \text{ corr } a_1$

4	1	1
3	1	1
3	1	1

${}^4y_{a_1}$

$a_1 \text{ corr } {}^4h_{c_1} = a_1 \text{ corr } c_1$

3	1	2
3		2
3		2

${}^4y_{c_1}$

$c_1 \text{ corr } {}^4h_{a_1} = c_1 \text{ corr } a_1$

3	2	1
3	2	
3	2	

$c_1 \text{ corr } {}^4h_{c_1} = c_1 \text{ corr } c_1$

4	1	1
3	1	1
3	1	1

(*a*)

	${}^4_\theta\vec{s}$	${}^4_2\vec{s}$	${}^4_3\vec{s}$	${}^4_{\ln}\vec{s}$
A:	$\begin{pmatrix}1\\0\end{pmatrix}$	$\begin{pmatrix}40\\40\end{pmatrix}$	$\begin{pmatrix}124\\106\end{pmatrix}$	$\begin{pmatrix}8.54\\8.15\end{pmatrix}$
C:	$\begin{pmatrix}0\\1\end{pmatrix}$	$\begin{pmatrix}40\\40\end{pmatrix}$	$\begin{pmatrix}106\\124\end{pmatrix}$	$\begin{pmatrix}8.15\\8.54\end{pmatrix}$

(*b*)

Fig. 6.10 *Cross- and autocorrelation functions of two signals (a) and the resulting PDC-features (b). Linear classification is possible except for the parabolic non-linearity.*

Multilinear Approach to Pattern Decomposition

It must be concluded from the empirically gained condition $p = q = 3$ that only a single term of each polynomial ${}^3_3u_{\chi i}$ (cf. Equation 6.8) is actually responsible for ideal translation invariant class definitions, namely the so-called trilinear term ${}^3d_{\chi i}$ which comprises those $q = 3$ signal components that are selected by mask ${}^3h_{\chi i}$. For every feature ${}^3_3s_\chi$ this corresponds to a sum of n well-defined trilinear terms, i.e. to a special trilinear form ${}^3r_\chi$. The corresponding q-linear form is defined by

$$ {}^qr_\chi = \sum_{i=1}^{n} {}^qd_{\chi i} = \sum_{i=1}^{n} \prod_{\substack{\sigma=1 \\ h\neq 0}}^{q} h_{\chi i\sigma} \cdot x_{\chi i\sigma} \tag{6.11} $$

where $x_{\chi i\sigma}$ is one of those q components of a signal vector $\vec{x}$ that are selected and possibly weighted by the q non-zero components $h_{\chi i\sigma}$ of mask vector ${}^q\vec{h}_{\chi i}$. For 0/1-masks that are presupposed in what follows, such special kinds of q-linear forms are identical with autocorrelation coefficients of order q of a signal $\vec{x}$.

Obviously, pattern elements are also defined by the multiplication schemes that are expressed by 0/1-masks ${}^qh_\chi$. They are represented through multilinear terms, i.e. by products of signal components. In other words, non-zero terms indicate the presence of the corresponding elements (coincidence detection). It is remarkable that this kind of element detection is completely different from mask matching, because neither cross correlations of masks and signal, nor subsequent decisions take place. In order to obtain unrestricted translation invariant pattern descriptions, each multiplication scheme must be applied at every position in the image array and all terms stemming from the same scheme must be summed (generalized counting) which results in multilinear forms. Just as with the polynomial approach (cf. Equation 6.8), normalization of the pattern signals is not necessary. This kind of feature extraction shall be called a multilinear approach to pattern decomposition and "counting" (MDC). Autocorrelation coefficients of order $q \geq 2$ form a translation invariant κ-dimensional feature vector $\vec{r}$ that can be classified by a normalized linear classifier.

For example, the computation of the translation invariant feature ${}^3_3s_8$ (cf. Table 6.1) involves $n = 9$ polynomials ${}^3_3u_{8i}$ of third degree and order that are listed below. They are defined by the mask ${}^3h_8$ (see Fig. 6.8(b)) and, in the simplest form, by the cubic non-linearity ${}_3u_{\chi i} = \left(y_{\chi i}\right)^3$:

$$ \begin{aligned} {}^3_3u_{81} &= (x_1 + x_4 + x_9)^3 \\ &= x_1^3 + x_4^3 + x_9^3 + 6x_1x_4x_9 + \\ &\quad 3\left(x_1^2x_4 + x_1^2x_9 + x_4^2x_1 + x_4^2x_9 + x_9^2x_1 + x_9^2x_4\right); \end{aligned} $$

$$ \begin{aligned} {}^3_3u_{82} &= (x_2 + x_5 + x_7)^3; & {}^3_3u_{83} &= (x_3 + x_6 + x_8)^3; \\ {}^3_3u_{84} &= (x_4 + x_7 + x_3)^3; & {}^3_3u_{85} &= (x_5 + x_8 + x_1)^3; \\ {}^3_3u_{86} &= (x_6 + x_9 + x_2)^3; & {}^3_3u_{87} &= (x_7 + x_1 + x_6)^3; \\ {}^3_3u_{88} &= (x_8 + x_2 + x_4)^3; & {}^3_3u_{89} &= (x_9 + x_3 + x_5)^3; \end{aligned} $$

Therefore, the trilinear form that is contained in the feature polynomial ${}^3_3s_8$ is

$$ \begin{aligned} {}^3r_8 = {} & x_1x_4x_9 + x_2x_5x_7 + x_3x_6x_8 + x_4x_7x_3 + \\ & x_5x_8x_1 + x_6x_9x_2 + x_7x_1x_6 + x_8x_2x_4 + x_9x_3x_5 \end{aligned} $$

which, for the pattern signal a_1, turns out to consist only of the single term ${}^3d_{85} = x_5x_8x_1 = 1$ and, for the pattern signal c_1, is even zero. This difference is indeed responsible for the difference of the corresponding feature values ${}^3_3s_8$ that are shown in Table 6.1.

Figure 6.7(b) shows a network for the parallel computation of multilinear forms. Compared with Fig. 6.7(a) (black inscriptions), the summation nodes are replaced by multiplication nodes and the non-linear transfer characteristics are omitted, i.e. the TLUs are substituted by product units (Glünder, 1986; Giles and Maxwell, 1987; Durbin and Rumelhart, 1989). Owing to the form of Equation 6.11, these structures belong to the category of so-called $\Sigma\Pi$-networks (Rumelhart *et al.*, 1986). The considerations concerning the subsequent linear classifier network that were made in the last section, apply here as well.

By Condition 6.10 it is not yet stated that $q = 3$ is a sufficient condition for ideal translation invariant pattern descriptions. However, because this crucial parameter is responsible for the number of features and thus determines the costs of invariant classification, the finding reported in what follows is of the utmost importance. It has been proven by several authors, firstly by McLaughlin and Raviv (1968), for binary pattern signals by Minsky and Papert (1969), and for complex-valued signals by Lohmann and Wirnitzer (1984), that any pattern of finite extent is perfectly described by its complete triple-autocorrelation function, except for its translational position (translation invariance). This function comprises all triple-autocorrelation coefficients that are pure trilinear forms but also all autocorrelation coefficients that are made up by sums of mixed terms, as well as the coefficient which is the sum of the cubes of all signal components. Actually, ideal invariant class definitions require pattern descriptions either by the whole triple-autocorrelation function or, alternatively, by all ${}^3\kappa$ trilinear and ${}^2\kappa$ bilinear autocorrelation coefficients, both in conjunction with subsequent template matching and standard normalization. In general, this holds if no *a priori* information (for example, about the size of a pattern) is available. In not considering the binary mask coefficients, the necessary and sufficient condition for translation invariant classification with ideal class definitions is thus given by:

$$(q = 2 \text{ and } q = 3) \Rightarrow N \approx \frac{n^2}{6} \quad \text{for } n \gg 3 \qquad (6.12)$$

This result expresses the primary role polynomial order q plays in ideal invariant classification and it demonstrates that the polynomial degree p is important only insofar as multilinear terms are to be generated through polynomials, e.g. by generalized TLUs. Autocorrelation can be generalized in order to formulate and analyze features that are invariant under various other geometric transformations (Glünder, 1987).

Although the considerations about pattern recognition from triple-autocorrelations do not generally hold for signals on toroidal image arrays, such pattern descriptions are now to be exemplified for the pattern classes A, B and C (cf. Fig. 6.6). In the right section of Table 6.1, the complete translation invariant feature vectors ${}^2\vec{r}$ and ${}^3\vec{r}$ for the three classes are given. The former are bilinear forms, computed according to the ${}^2\kappa = 4$ types of multiplication schemes shown in Fig. 6.8(a), and the latter are trilinear forms that result from the ${}^3\kappa = 12$ configurations depicted in Fig. 6.8(b). Again, classes A and C are not separable on the basis of the feature vector ${}^2\vec{r}$, because their members are related by point inversion and thus have identical second order autocorrelation functions. However, all three patterns are linearly separable if the feature vector ${}^3\vec{r}$ is used, but no longer with respect to a single component of the feature vector, as it was the case with the results given in the right half of the left section of Table 6.1.

Discussion and Assessment

Three approaches to unrestricted translation invariant classification have been compared. The criterion was the effort per class for ideal class definitions, i.e. for the feasibility of unequivocal pattern reconstructions from the classification results. The effort was specified by the number N of adjustable weighting coefficients for each class that are necessary for correct classification. Somewhat surprisingly, the minimum costs for the approaches turn out to be of the same order, namely $N \approx n^2$, for n-dimensional pattern signals on a toroidal image array. The same holds for the normalization effort. Although the general polynomial classifier with $p = 3$, and therefore about n^3 weights (see Equation 6.7), appears to be more costly, it equals the MDC-approach if appropriate grouping is applied to the polynomial terms.

On the basis of this outcome, it would be false to conclude that all three approaches are equally well suited in practice. Invariant list classification, for instance, is only applicable if the pattern signals are exactly those stored as templates, because, due to the enormous costs of maximum detection ($q = kn$), its generalization turns out to be impractical in the case of many classes. Therefore, this method is restricted to well-defined technical classification problems and it is not suited for biological pattern recognition. The PDC- and MDC-approaches do not suffer from this short-coming as their features can simply be classified by a general linear classifier ($q = k$). Owing to the many polynomial terms that constitute features in PDC-systems, processing structures must cope with the involved dynamic range which, depending on the kind of signal representation, can be very large. On the other hand, signal processing must be highly accurate in order to represent the crucial trilinear terms that are "riding" on sums of otherwise unimportant polynomial terms. In this regard, MDC-systems are more economic and robust (cf. Fig. 6.11).

Another aspect of assessment is the reduction of costs if classification of restricted invariance or based on not ideally defined classes is of concern. Under such conditions, two alternative strategies must additionally be considered, both of which are related to list classification. Since the effort of classification depends directly on the dimensionality of the signals, its reduction is an effective means for cost reduction. Linear reduction of dimensions uses cross correlation with a minimum number of templates that maximally decorrelate the patterns under consideration. The resulting correlation coefficients must

then be non-linearly classified. The other approach uses adequately defined subclasses. Again, cross correlation is applied but this time with templates that produce a feature space in which all signals of interest are split into a minimum number of linearly separable subclasses. After subclassification, the results are grouped in order to obtain the class-specific output signals (cf. Fig. 6.7(a)). In the case of few classes, a considerable decrease of the costs can be achieved by both methods. Because this reduction relies on the clever choice of highly task-specific templates, these approaches are mainly suited for applications where the number of classes is predetermined. Otherwise, any additional invariant class necessitates the change of all templates. Owing to this lack of flexibility, these methods are biologically implausible although they have become standard in technical domains.

Restriction of signal variance permits a proportional decrease in the number – not the types – of templates, masks, or multiplication schemes that serve for list, PDC- and MDC-based classification respectively. This is possible, because (translation) invariance is exclusively due to the appropriate summation of non-linearly weighted correlation coefficients, or multilinear terms that correspond to a certain type of template, mask, or scheme, and which are extracted at different (translational) positions, according to the extent of the geometric variance. Pitts and McCulloch (1947) called this procedure the averaging of different types of functionals that are evaluated from the transformed signals, over the groups of the underlying transformations. These authors also recognized that the number of types of functionals which are necessary in order to completely characterize a signal is extremely large, and that the nervous system uses perhaps less than complete information for the recognition of shapes. It should be remembered that every single "decomposition and counting"-feature is invariant, i.e. a reduction in the number of features does not affect the invariance but the precision of the pattern descriptions. In this sense, good invariant descriptions even require certain types of variance, namely all those geometric transformations that relate the pixels of a pattern. That is why two-dimensional translations, as well as expansions in conjunction with rotations are well suited for ideal pattern characterizations. Obviously, a reduction of the number of template types is not applicable to invariant list classification. However, for both decomposition approaches, a restriction, for instance by a factor n, to all ${}^{2}\kappa \approx n/2$ possible features, does not cause a severe loss in descriptive power. With this specially restricted feature vector, patterns that have identical second order autocorrelation functions can no longer be distinguished, for instance those related by point inversion.

Unfortunately, the decrease in descriptive power that is caused by fewer features, happens more rapidly for the PDC- than for the MDC-features, at least if non-negative signals are considered. This is due to the principal fact that a PDC-feature cannot unambiguously signal the presence or absence of a form element in patterns which in turn is a consequence of the normalization problem. Owing to the coincidence-detection character of multiplication, MDC-features are ideally suited for such decisions (cf. Fig. 6.11), especially if they are computed from sparsely coded signal representations, i.e. from those that contain only a few non-zero values. Even if coincidence detection is taken literally, i.e. if it results in binary decisions about the presence of inner pattern bindings, the invariant description as well as the perfect reconstruction of the binary versions of patterns is still possible. (For the latter task all ${}^{3}\kappa$ features are required but not the ${}^{2}\kappa$ bilinear forms.) PDC-systems lack this valuable property because pattern information is essentially contained in the amplitudes of polynomials.

To give an example. It is demonstrated that the PDC-approach does not permit the unambiguous and translation invariant detection of single form elements in non-binary and unipolar patterns. For this purpose the grey-valued patterns E and F – both of the same mean – the mask ${}^{3}h_{10}$ (cf. Fig. 6.8(b)), as well as the cubic non-linearity ${}_{3}u_{\chi i} = \left(y_{\chi i}\right)^{3}$ are considered. Although pattern E contains the form element ${}^{3}h_{10}$ and pattern F does not, they both lead to the same and considerably high feature value ${}^{3}_{3}s_{10}$ (Fig. 6.11(a)). The corresponding MDC-features reflect this pattern difference and they are of much lower value (Fig. 6.11(b)).

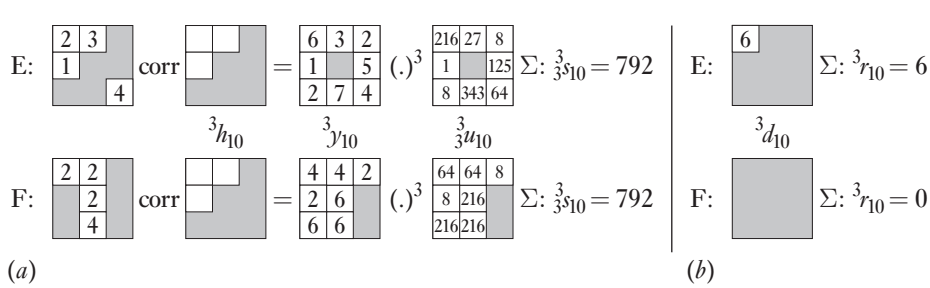

Fig. 6.11 *Detection of the form element* ${}^{3}h_{10}$ *(see Fig. 6.8(b)) in two patterns by means of PDC (a) and MDC (b).*

Extraction of Invariant Features by (Artificial) Neural Networks

From the previous considerations one can conclude that the extraction of invariant features by PDC- and MDC-approaches meets essential requirements of biological pattern recognition, namely the flexibility concerning both, the type and extent of invariance, and the precision of the pattern descriptions. On the other hand, the role classification plays in biological pattern recognition is not clear at all and it is questionable whether a properly normalized linear classifier stage actually follows feature extraction.

It is evident that PDC-approaches, such as *Perceptrons*, are based on the model neurone that was introduced by McCulloch and Pitts (1943) for other reasons. This model represents a TLU that was later generalized by the introduction of various other non-linear transfer characteristics. A generalized TLU-model neurone computes the weighted sum of all presynaptic activities and produces axonal impulse rates that are proportional to the non-linearly weighted somatic potential. Because real neurones have several thousand synapses and, assuming realistic time constants at the destination neurones, can only represent less than 50 distinct levels at their axonal outputs (Barlow, 1963), accurate signal processing becomes difficult under the generalized TLU-paradigm, even if only a small percentage of the inputs are active. Opposed to the neuroanatomical evidence, it was demonstrated here that universal PDC-systems optimally consist of masks that act on very few pixels which in turn would call for "neurones" with such few inputs.

MDC-systems require computing units with many non-linearly interacting inputs and with a single and substantially linear output. Such $\Sigma\Pi$-model neurones perform conjunctive, e.g. multiplying operations on neighbouring synaptic inputs and sum the multilinear terms thus generated. The discovery of dendritic conductance changes that depend non-linearly on the local dendritic membrane potential, i.e. in general on potentials that are produced by more than one synaptic input, indicates a possible physiological mechanism that can serve for coincidence detection (Dingledine, 1983, MacDermott and Dale, 1987). Because non-linear processing, i.e. the selection of signals, takes place at the input of such units, overload is no longer a problem, provided the signals are sparsely coded. Owing to these capabilities, a single $\Sigma\Pi$-model neurone evaluates a multilinear form of moderate polynomial order that consists of several hundred to a thousand terms. In other words, it can compute, for instance, a translation invariant feature from a 30×30 image array. Features from larger arrays can be obtained by summation of such locally computed features if the local areas have an overlap of at least the extent of the underlying multiplication scheme. It can be shown that ideal pattern descriptions by MDC-features are exclusively based on all those trilinear and bilinear terms that contain the products between the values of the two most distant pattern pixels. Consequently, large patterns cannot perfectly be characterized through local operations (cf. similar conclusions by Minsky and Papert, 1969). Finally it is noteworthy that coincidence detection between few synaptic inputs can also be regarded as a consequence of self-structuration that starts from a signal representation of certain *a priori* frequencies of occurrence of the signal components (Phillips *et al.*, 1984). The probability of a coincidence is the product of the *a priori* probabilities of the contributing events, provided they are statistically independent, which in turn is a question of spatial signal coding. In other words, the chance for a coincidence decreases exponentially with the number of the events involved. If self-organization on the basis of synaptic modifications requires the frequent success of coincidence (local dendritic version of Hebb's postulate, cf. Kelso *et al.*, 1986; Nicoll *et al.*, 1988), then there is little chance for the formation of coincidence detectors with many inputs.

In conclusion, MDC-systems are advantageous if invariance and flexibility are demanded from a pattern recognition system. Mainly for computational reasons and for their unequivocal definition even of single features, MDC-systems are superior to PDC-systems. The MDC-principle is well-established in the field of commercial character recognition (Schürmann, 1977) – mostly in the form of parabolic classifiers – and there is evidence that it plays a role in neural pattern processing too.

Acknowledgements

I thank E. Pöppel for his interest and friendly support, H. Platzer for many helpful discussions and advice, and J. Schürmann for providing his lecture notes and unpublished reports on polynomial classification and network structures. Essential parts of this contribution were conceived while on leave at the "École Nationale Supérieure des Télécommunications de Bretagne". The support of the "Département Mathématique et Système de Communication" and especially of A. Hillion is acknowledged.

References

Barlow, H. B. (1963). The information capacity of nervous transmission. *Kybernetik*, **2**, 1.

Brousil, J. K. and Smith, D. R. (1967). A threshold logic network for shape invariance. *IEEE Trans. Electronic Computers*, **EC-16**, 818–828.

Casasent, D. and Psaltis, D. (1976). Position, rotation and scale invariant optical correlation. *Appl. Opt.*, **15**, 1795–1799.

Dingledine, R. (1983). N-methyl-aspartate activates voltage-dependent calcium conductance in rat hippocampal pyramidal cells. *J. Physiol.*, **343**, 385-405.
Doyle, W. (1962). Operations useful for similarity-invariant pattern recognition. *J. ACM*, **9**, 256-267.
Durbin, R. and Rumelhart, D. E. (1989). Product units: a computationally powerful and biologically plausible extension to backpropagation networks. *Neural Computation*, **1**, 133-142.
Feldman, J. A. (1982). Dynamic connections in neural networks. *Biol. Cybern.*, **46**, 27-39.
Fukunaga, K. (1972). *Introduction to Statistical Pattern Recognition*. New York, NY: Academic Press.
Giles, C. L. and Maxwell, T. (1987). Learning, invariances, and generalization in high-order neural networks. *Appl. Opt.*, **26**, 4972-4978.
Glünder, H. (1986). Neural computation of inner geometric pattern relations. *Biol. Cybern.*, **55**, 239-251.
Glünder, H. (1987). Invariant description of pictorial patterns via generalized autocorrelation functions. In *ASST '87*. ed. Meyer-Ebrecht, D. pp. 84-87. Berlin: Springer Verlag.
Hadeler, K. P. (1974). On the theory of lateral inhibition. *Kybernetik*, **14**, 161-165.
Kelso, S. R.; Ganong, A. H. and Brown, T.H. (1986). Hebbian synapses in hippocampus. *Proc. Natl. Acad. Sci. USA*, **83**, 5326-5330.
Kröse, B. J. A. (1985). A structure description of visual information. *Patt. Recogn. Lett.*, **3**, 41-50.
Lippmann, R. P. (1987). An introduction to computing with neural nets. *IEEE ASSP Magazine*, **4**, 4-22.
Lohmann, A. W. and Wirnitzer, B. (1984). Triple correlations. *Proc. IEEE*, **72**, 889-901.
MacDermott, A. B. and Dale, N. (1987). Receptors, ion channels and synaptic potentials underlying the integrative actions of excitatory amino acids. *TINS*, **10**, 280-284.
McCulloch, W. S. and Pitts, W. H. (1943). A logical calculus of ideas immanent in nervous activity. *Bull. Math. Biophys.*, **5**, 115-133.
McLaughlin, J. A. and Raviv, J. (1968). Nth-order autocorrelations in pattern recognition. *Information and Control*, **12**, 121-142.
Marr, D. (1982). *Vision. A Computational Investigation into the Human Representation and Processing of Visual Information*. San Francisco, CA: Freeman.
Minsky, M. L. and Papert, S. A. ([1]1969, [2]1988). *Perceptrons. An Introduction to Computational Geometry*. Cambridge, MA: MIT Press.
Moore, D. J. H. and Parker D. J. (1974). Analysis of global pattern features. *Pattern Recognition*, **6**, 149-164.
Nicoll, R. A., Kauer, J. A. and Malenka, R. C. (1988). The current excitement in long-term potentiation. *Neuron*, **1**, 97-103.
Pao, Y.-H. (1989). *Adaptive Pattern Recognition and Neural Networks.* Reading, MA: Addison-Wesley.
Phillips, C. G., Zeki, S. and Barlow, H. B. (1984). Localization of function in the cerebral cortex. *Brain*, **107**, 328-361.
Pitts, W. and McCulloch W. S. (1947). How do we know universals. The perception of auditory and visual forms. *Bull. Math. Biophys.*, **9**, 127-147.
Rosenblatt, F. (1962). *Principles of Neurodynamics. Perceptrons and the Theory of Brain Mechanisms*. Washington, DC: Spartan Books.
Rumelhart, D. E. and McClelland, J. L. (1986). *Parallel Distributed Processing 1*. Cambridge, MA: MIT Press.
Rumelhart, D. E., Hinton, G. E. and McClelland, J. L. (1986). A general framework for parallel distributed processing. In *Parallel Distributed Processing 1*. eds. Rumelhart, D. E. and McClelland, J. L. pp. 45-76. Cambridge, MA: MIT Press.
Schürmann J. (1977). *Polynomklassifikatoren für die Zeichenerkennung*, Munich: Oldenbourg Verlag.
Sebestyen, G. S. (1962). *Decision-making Processes in Pattern Recognition*. New York, NY: Macmillan.
Uesaka, Y. (1971). Analog perceptrons: on additive representations of functions. *Information and Control*, **19**, 41-65.
Uesaka, Y. (1975). Analog perceptron: its decomposition and order. *Information and Control*, **27**, 199-217.
Watt, R. (1988). *Visual Processing: Computational, Psychophysical, and Cognitive Research*. Hove, UK: Erlbaum.

As David G. Stork (1994) remarks in his review of *Vision and Visual Dysfunction*, "The one noteworthy editorial or production weakpoint in the Series [...] concerns the mathematical typography, which is erratic, and does not always rise to the level one expects from typeset books in this price range. [...] Such widespread cavalier mathematical typography from CRC Press [the US-American Publisher of the book series] [...] is frankly a disappointment." Of course, contributions extensively dealing with formal topics suffer most from sloppy typesetting, misprints and bad printing quality, as it happens to be the case in Chapter 6 of Volume 14. With the here presented electronic reprint, most of these deficiencies and a few errors, for which the author is responsible, are corrected. Except essentially for the example in Figure 6.10 that is made more explicit, the original content is left unchanged. The most obvious typographic change concerns the vector notation with the originally intended arrows instead of semibold characters.

Gavin Brelstaff (1992) states in his review of *Pattern Recognition by Man and Machine*: "Surprisingly, given the volume's title, no chapter provides a broad review of standard pattern recognition techniques—as described in texts such as those by Duda and Hart (1973), Rosenfeld and Kak (1976), and Tou and Gonzales (1974). Chapter 6 comes closest in content, where Glünder discusses techniques for classifying pictorial patterns. He briefly defines template matching, statistical pattern recognition, and geometrical invariant pattern recognition, before concentrating on a particular problem: how linear classifiers can be rendered translationally invariant. With such classifiers it should be possible to recognise patterns without first locating them. His uncompromisingly mathematical treatment leads to a discussion of how invariant features might be extracted by neural networks."

Brelstaff G. (1992) Pattern recognition by man and machine (Book Review). *Perception*, **21**: 559-560.
Stork D. G. (1994) Vision and visual dysfunction (Book Review). *Journal of Cognitive Neuroscience*, **6**: 180-183.

VISION AND VISUAL DYSFUNCTION
VOLUME 14

Pattern Recognition by Man and Machine

Edited by

Roger J. Watt

Dept of Psychology
University of Stirling
Stirling, UK

M
MACMILLAN
PRESS
Scientific & Medical

© The Macmillan Press Ltd 1991

All Rights reserved.

Apart from any fair dealing for the purposes of research or private study, or criticism or review, as permitted under the Copyright, Designs and Patents Act, 1988, this publication may only be reproduced, stored or transmitted, in any form or by any means, with the prior permission in writing of the publishers, or in the case of reprographic reproduction in accordance with the terms of licences issued by the Copyright Licensing Agency. Enquiries concerning reproduction outside those terms should be sent to the publishers at the undermentioned address.

Any person who does any unauthorized act in relation to this publication may be liable to criminal prosecution and civil claims for damages.

First published 1991 by
THE MACMILLAN PRESS LTD
Houndmills, Basingstoke, Hampshire RG21 2XS
and London
Companies and representatives
throughout the world

Typeset in Monophoto Ehrhardt by August Filmsetting, Haydock, St Helens, UK
Printed and bound in Great Britain
by William Clowes, Beccles and London

British Library Cataloguing in Publication Data
Patter recognition by man and machine
1. Man. Visual perception. Pattern recognition.
2. Computers. Pattern recognition.
I. Watt, R. J. II. Series
152.1′432
621.39.9

ISBN 0–333–49635–3
ISBN 0–333–52713–5 set.