DISCRIMINATIVE METRIC DESIGN FOR PATTERN RECOGNITION

Hideyuki WATANABE, Tsuyoshi YAMAGUCHI, and Shigeru KATAGIRI

ATR Interpreting Telecommunications Research Laboratories
2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-02, Japan
E-mail: watanabe@itl.atr.co.jp

ABSTRACT

This paper proposes a new approach, named *Discriminative Metric Design (DMD)*, to pattern recognition. DMD optimizes discriminant functions with the Minimum Classification Error/Generalized Probabilistic Descent method (MCE/GPD) such that intrinsic features of each pattern class can be represented efficiently. Resulting metrics accordingly lead to robust recognizers. DMD is quite general. Several existing methods, such as Learning Vector Quantization and the Continuous Hidden Markov Model, are defined as its special cases. The paper specially elaborates the DMD formulation for the quadratic discriminant function, and clearly demonstrates its utility in a speaker-independent Japanese vowel recognition task.

1. INTRODUCTION

In most pattern recognizers, feature extraction is not necessarily appropriately linked with the recognition decision; this often complicates the over-learning problem or the robustness problem. In light of this, this paper proposes a new method, named *Discriminative Metric Design (DMD)*, which allows one to alleviate the problem by designing discriminant functions that can effectively represent intrinsic features of its corresponding class.

DMD is quite general and can be applied to various types of recognizers as well as a wide range of recognition tasks. The paper is specifically devoted to DMD implementation for a *static* (fixed-dimensional) pattern recognizer using a fundamental, quadratic discriminant function and its evaluation in a speaker-independent Japanese five-vowel recognition task.

DMD relies on a recently-developed, general discriminative learning methodology, the Minimum Classification Error/Generalized Probabilistic Descent method (MCE/GPD) [3, 4, 5]. It has been shown that this MCE/GPD can be considered a more general version of recent recognizer design algorithms including Learning Vector Quantization (LVQ) [4]. The development of DMD greatly widens the scope with regard to this relationship. In the paper, a discussion is made on the relationship between DMD and several important algorithms, such as LVQ, the Learning Subspace Method (LSM) [7], Discriminative Feature Extraction (DFE) [1], and an MCE/GPD-trained kernel function recognizer [4].

2. DISCRIMINATIVE METRIC DESIGN

2.1. Statistical Pattern Recognition

Let us consider the problem of classifying a d-dimensional input pattern $\boldsymbol{x} \in \mathcal{R}^d$ into one of K classes $\{C_s\}_{s=1}^K$. We assume that the dimensionality of the pattern is as high as in many pattern recognition tasks. Our decision rule is as follows:

$$C(\boldsymbol{x}): \quad C(\boldsymbol{x}) = C_i \quad \text{if} \quad i = \arg\min_s g_s(\boldsymbol{x}), \qquad (1)$$

where $C(\boldsymbol{x}) : \mathcal{R}^d \mapsto \{C_s\}_{s=1}^K$ is the recognition operation and $g_s(\boldsymbol{x})$ is the discriminant function that indicates the degree to which $\boldsymbol{x}$ belongs to C_s. The ultimate goal here is to achieve discriminant functions that can minimize the recognition error probability. In reality, however, despite many approaches, achieving this goal has been rather difficult due to the limited amount of available resources such as design samples.

2.2. The Concept of the New Approach

In most cases, the discriminant function is simply based on heuristics and on some kind of scientific knowledge indirectly related to error minimization. Such functions are never guaranteed to lead to a *robust* recognition that is accurate over unknown samples. One way to remedy this inadequacy is to design each discriminant function so as to represent the intrinsic features of its corresponding class efficiently.

Our solution, DMD, is illustrated in Fig. 1. DMD forms an *individual metric* for each class, and also de-

0-7803-2431-5/95 $4.00

signs this metric and its corresponding similarity measure consistently with the minimum error objective. This design can consequently increase the design robustness: Each class membership is evaluated in its corresponding class-feature space where features relevant to recognition are emphasized and information irrelevant to recognition is suppressed.

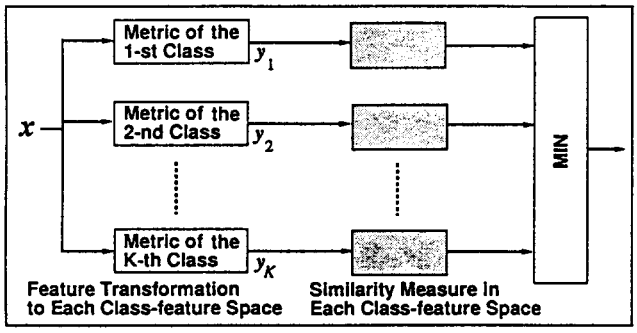

Figure 1: Pattern Recognizer Based on DMD

2.3. The definition of Metric

We define in this paper each class metric as the following linear transformation $\mathcal{L}_s$ from the original pattern space $\mathcal{X}$ to the s-th class feature space $\mathcal{Y}_s$:

$$\begin{aligned} \boldsymbol{y}_s &= \mathcal{L}_s(\boldsymbol{x}) = \boldsymbol{\Phi}_s \boldsymbol{V}_s^T \boldsymbol{x} \quad (s = 1, 2, ..., K) && (2) \\ \boldsymbol{\Phi}_s &= \text{diag}(\ \varphi_{s,1}\ \varphi_{s,2}\ \cdots\ \varphi_{s,d}\ ) && (3) \\ \boldsymbol{V}_s &= [\ \boldsymbol{v}_{s,1}\ \boldsymbol{v}_{s,2}\ \cdots\ \boldsymbol{v}_{s,d}\ ], \quad \boldsymbol{V}_s^T \boldsymbol{V}_s = \boldsymbol{I}, && (4) \end{aligned}$$

where the superscript ^{T} denotes the matrix transposition, diag( ) stands for the diagonal matrix, and $\boldsymbol{I}$ is the identity matrix. Each $\boldsymbol{V}_s$ forms the orthonormal base of the s-th class feature space, and $\varphi_{s,i}$ is the weight that indicates the contribution to the axis $\{\boldsymbol{v}_{s,i}\}$; so the smaller the weight is, the less the component on the corresponding axis can be utilized. Accordingly, each parameter set $(\boldsymbol{\Phi}_s, \boldsymbol{V}_s)$ comprises a feature space intrinsic to each class $\boldsymbol{C}_s$.

The similarity measure on the linear-transformed space $\mathcal{Y}_s$ can be arbitrarily specified. Then, we specially consider the case of using a Euclidean distance measure on $\mathcal{Y}_s$. Each discriminant function comes simply to a quadratic one (which is fundamental to many pattern classifiers) as follows:

$$\begin{aligned} g(\boldsymbol{x}; \boldsymbol{\Theta}_s) &= ||\mathcal{L}_s(\boldsymbol{x}) - \mathcal{L}_s(\boldsymbol{r}_s)||^2 \\ &= (\boldsymbol{x} - \boldsymbol{r}_s)^T \boldsymbol{V}_s \boldsymbol{\Phi}_s^2 \boldsymbol{V}_s^T (\boldsymbol{x} - \boldsymbol{r}_s) && (5) \\ \boldsymbol{r}_s &= [\ r_{s,1}\ r_{s,2}\ \cdots\ r_{s,d}\ ]^T && (6) \\ \boldsymbol{\Theta}_s &= \{\boldsymbol{r}_s, \boldsymbol{\Phi}_s, \boldsymbol{V}_s\}, && (7) \end{aligned}$$

where $\boldsymbol{r}_s$ represents the reference vector of class $\boldsymbol{C}_s$.

2.4. The Formulation of DMD

A key concept of DMD is to fully optimize the entire metric of each discriminant function so that each class-feature space can effectively and efficiently represent each class identity which is essential for accurate recognition. In the quadratic discriminant function case, the DMD formalization is elaborated as shown in the following paragraph.

Similar to MCE/GPD, the adjustment (training) mechanism of DMD is based on gradient search optimization. Adjusting $\boldsymbol{r}_s$ and $\boldsymbol{\Phi}_s$ is fairly easy. It can be done in the same way as is done for the original MCE/GPD. However, $\boldsymbol{V}_s$ is difficult to adjust because of its orthonormal constraint ($\boldsymbol{V}_s^T \boldsymbol{V}_s = \boldsymbol{I}$). To overcome this difficulty, $\boldsymbol{V}_s$ is represented by the multiplication of Jacobi-rotation matrices [2] as follows:

$$\boldsymbol{V}_s = \boldsymbol{U}_{1,2}(\theta_{s,1,2}) \boldsymbol{U}_{1,3}(\theta_{s,1,3}) \cdots \boldsymbol{U}_{d-1,d}(\theta_{s,d-1,d}), \quad (8)$$

where the $d \times d$ matrix $\boldsymbol{U}_{p,q}(\theta)$ $(p < q)$ is an orthogonal matrix such that entries (p,p) and (q,q) are $\cos\theta$, entry (p,q) is $\sin\theta$, entry (q,p) is $-\sin\theta$, and the other entries are 1 on diagonal and 0 on non-diagonal. Note here that the parameter set $\boldsymbol{\theta}_s = \{\theta_{s,1,2} ... \theta_{s,d-1,d}\}$ does not need to hold to the above constraint. This matrix decomposition thus allows us to adjust $\boldsymbol{V}_s$ through the simple adjustment of $\boldsymbol{\theta}_s$. Accordingly, for a design sample $\boldsymbol{x}_t$ $(\in \boldsymbol{C}_k)$, the full DMD updating rule is completed by using the chain rule of derivatives and is given as follows:

$$\begin{aligned} \boldsymbol{p}_s^{(t)} &= \boldsymbol{p}_s^{(t-1)} + \Delta \boldsymbol{p}_s^{(t)} && (9) \\ \Delta \boldsymbol{p}_s^{(t)} &= 2\varepsilon_t \ell_k'(\boldsymbol{x}_t; \boldsymbol{\Theta}^{(t-1)}) \rho_{k,s}(\boldsymbol{x}_t; \boldsymbol{\Theta}^{(t-1)}) \\ &\quad \times \boldsymbol{W}_s^{(t-1)} \boldsymbol{\Phi}_s^{(t-1)^2} \boldsymbol{W}_s^{(t-1)^T} \boldsymbol{z}_s^{(t)} && (10) \\ \varphi_{s,i}^{(t)} &= \varphi_{s,i}^{(t-1)} + \Delta \varphi_{s,i}^{(t)} && (11) \\ \Delta \varphi_{s,i}^{(t)} &= -2\varepsilon_t \ell_k'(\boldsymbol{x}_t; \boldsymbol{\Theta}^{(t-1)}) \rho_{k,s}(\boldsymbol{x}_t; \boldsymbol{\Theta}^{(t-1)}) \\ &\quad \times \varphi_{s,i}^{(t-1)} |\boldsymbol{w}_{s,i}^{(t-1)^T} \boldsymbol{z}_s^{(t)}|^2 \ (i = 1, 2, ..., d) && (12) \\ \theta_{s,p,q}^{(t)} &= \theta_{s,p,q}^{(t-1)} + \Delta \theta_{s,p,q}^{(t)} && (13) \\ \Delta \theta_{s,p,q}^{(t)} &= -2\varepsilon_t \ell_k'(\boldsymbol{x}_t; \boldsymbol{\Theta}^{(t-1)}) \rho_{k,s}(\boldsymbol{x}_t; \boldsymbol{\Theta}^{(t-1)}) \\ &\quad \times \boldsymbol{z}_s^{(t)^T} \boldsymbol{A}_{s,p,q}^{(t-1)} \boldsymbol{z}_s^{(t)} && (14) \\ \boldsymbol{A}_{s,p,q}^{(t-1)} &= \left(\frac{\partial \boldsymbol{W}_s}{\partial \theta_{s,p,q}} \right)^{(t-1)} \boldsymbol{\Phi}_s^{(t-1)^2} \boldsymbol{W}_s^{(t-1)^T} && (15) \\ &\quad (p = 1, 2, ..., d-1; q = 2, 3, ..., d; p < q), \end{aligned}$$

where the suffix t denotes the iteration step number $(t = 1, 2, ...)$, ε_t (> 0) is called the learning rate satisfying $\sum_{t=1}^{\infty} \varepsilon_t \to \infty$ and $\sum_{t=1}^{\infty} \varepsilon_t^2 < \infty$, $\ell_k'(\boldsymbol{x}; \boldsymbol{\Theta})$ (> 0) denotes the derivative (with respect to the misclassification measure) of the smooth loss function [3], $\rho_{k,s}(\boldsymbol{x}; \boldsymbol{\Theta})$

denotes the derivative (with respect to the s-th discriminant function) of the misclassification measure [3] satisfying $\rho_{k,k} > 0$ and $\rho_{k,j} < 0$ $(j \neq k)$, and the other variables are given by

$$\boldsymbol{p}_s^{(t)} = \boldsymbol{V}_s^{(0)T} \boldsymbol{r}_s^{(t)} \quad (16)$$

$$\boldsymbol{z}_s^{(t)} = \boldsymbol{V}_s^{(0)T} \boldsymbol{x}_t - \boldsymbol{p}_s^{(t-1)} \quad (17)$$

$$\boldsymbol{W}_s^{(t)} = \boldsymbol{V}_s^{(0)T} \boldsymbol{V}_s^{(t)} = \left[\boldsymbol{w}_{s,1}^{(t)} \; \boldsymbol{w}_{s,2}^{(t)} \; \ldots \; \boldsymbol{w}_{s,d}^{(t)} \right] \quad (18)$$

$$\boldsymbol{W}_s = \boldsymbol{U}_{1,2}(\theta_{s,1,2}) \boldsymbol{U}_{1,3}(\theta_{s,1,3}) \cdots \boldsymbol{U}_{d-1,d}(\theta_{s,d-1,d}). \quad (19)$$

The initial values of $\{\theta_{s,p,q}^{(0)}\}$ are all 0. Properties of the updating convergence are discussed in [3].

3. EXPERIMENTS

To evaluate DMD, we conducted a five-class, fixed-dimensional vowel pattern recognition experiment in a speaker-independent mode. Vowel tokens were extracted from 520 isolated words spoken by 70 speakers (36 males and 34 females), and digitized at a sampling rate of 12 kHz. The center fragment of each vowel segment was selected using a 20 ms Hamming window and converted into a recognizer input pattern consisting of 32 LPC cepstral coefficients. Note that each pattern sample was a single frame cepstral vector.

Recognition error rates were computed by the following procedure:

For $n = 1$ to 5 {

Select 10 speakers randomly for *unknown set* $\Omega_u^{(n)}$ (about 1500 samples) from the whole set Ω of 70 speakers;
Let $\bar{\Omega}_u^{(n)}$ be the remained set of 60 speakers;

For $m = 1$ to 5 {

Select 10 speakers randomly for *validation set* $\Omega_v^{(n,m)}$ (about 1500 samples) from $\bar{\Omega}_u^{(n)}$;
Let $\Omega_d^{(n,m)}$ be the remained *design set* of 50 speakers (about 7500 samples);
Train the recognizer $\Theta^{(n,m)}$ using $\Omega_d^{(n,m)}$;
Compute the error rate $P_{ev}^{(n,m)}$ of $\Theta^{(n,m)}$ for the validation set $\Omega_v^{(n,m)}$;

}

Select the *best* recognizer $\Theta^{(n,m^*)}$ where $P_{ev}^{(n,m^*)} = \min_m P_{ev}^{(n,m)}$;
Compute the error rate $P_{ed}^{(n)}$ and $P_{eu}^{(n)}$ of $\Theta^{(n,m^*)}$ for the design set $\Omega_d^{(n,m)}$ and the unknown set $\Omega_u^{(n,m)}$, respectively;

}

Compute the *averaged* error rate $P_{ed} = (1/5)\sum_{n=1}^{5} P_{ed}^{(n)}$ and $P_{eu} = (1/5)\sum_{n=1}^{5} P_{eu}^{(n)}$ for the design and unknown sets, respectively.

For comparison purposes, we also used three types of recognizers: 1) a quadratic discriminant-based DMD recognizer, 2) a Mahalanobis distance recognizer, and 3) multi-template (reference) LVQ recognizers [5]. The LVQ system used the Euclidean distance for its discriminant function. All of the parameters in the DMD-based recognizer were initialized using the Mahalanobis distance; i.e., in each class $(s = 1, 2, ..., K)$, $\boldsymbol{r}_s^{(0)}$ was the sample mean vector, $\boldsymbol{V}_s^{(0)}$ was the set of eigenvectors of the sample covariance matrix, and each $\varphi_{s,i}^{(0)}$ was the inverse of the square root of the eigenvalue.

Table 1 summarizes the (averaged) recognition error rates for these three systems. The DMD-based recognizer achieved a higher recognition performance for unknown patterns than the Mahalanobis distance-based one or the LVQ-based ones. Moreover, interestingly, the DMD-based system performed much better over the unknown sets than did the LVQ system with several templates, while the LVQ system performed best over the design set. This result demonstrates that DMD contributes toward increasing the robustness through a suitable design of the class-feature space.

Table 1: Recognition error rates for a Japanese five-vowel task

	design set	unknown set
DMD	4.00%	10.84%
Mahalanobis-distance	8.66%	16.36%
LVQ (1 template)	10.71%	13.73%
LVQ (8 templates)	5.48%	13.87%
LVQ (16 templates)	3.78%	14.76%

4. RELATIONSHIPS BETWEEN DMD AND OTHER TECHNIQUES

The DMD implementation for the quadratic discriminant function has important implications for other recognizer design techniques.

Performing the well-known Principal Component Analysis (PCA) in each class can be a simple solution for finding each metric. In PCA, the eigenvectors associated with the large eigenvalues of the sample covariance matrix represent the intra-class statistical variation factors. To reduce the influence of such variation

factors on recognition decisions, in other words, to normalize this type of variation, each weighting parameter $\varphi_{s,i}$ is usually set to the value of the parameter that is inversely proportional to the i-th eigenvalue. The Mahalanobis distance that was used in the previous section is equivalent to the case where the Euclidean distance is used in each class-feature space. This PCA-based design, however, estimates the parameters of each class independently and does not consider the influence of different classes; this does not necessarily reduce the recognition errors. This insufficiency has been demonstrated in the experimental results above.

Recently, demonstrations have been made of continuous Gaussian HMM speech recognizers based on MCE/GPD [6], which have achieved highly accurate recognition results. In most of these applications, diagonal covariance matrices were used: the original GPD adjustment rule was applicable only to this type of simple form matrix. In contrast, the DMD adjustment rule enables the full adaptation of full covariance matrices; this will improve the recognition performance compared to usual mixture Gaussian HMMs with diagonal matrices which essentially correspond to multi-template classifiers using a limited, simplified distance measure.

It is obvious that the continuous HMM recognizer is a general version of the RBF recognizer and the Likelihood Network recognizer [4]. Therefore, DMD also enables the full adjustment of these types of Neural Network-based systems.

The linear transformation considered in this paper can be viewed as a feature extraction process. This point reminds us of the close relation between DMD and the DFE that jointly optimizes both the feature extraction and classification processes for the purpose of minimum error [1]. It is actually obvious that DFE can be considered to be a special case of DMD. The difference between these two is that DFE uses a common metric over all of the classes while DMD designs an individual metric for each class.

DMD is also related to LSM in the sense that each class possesses its own feature space [7]. However, DMD is clearly distinct from LSM in several aspects. For instance, LSM computes as a discriminant function an orthogonal projection onto each class feature subspace, which restricts each input pattern to a member of a linear space. This restriction is valid only for scale-invariant patterns such as power spectra, and not for patterns based on log spectra or linear prediction. In contrast, DMD does not have this restriction, which is indicative of its higher applicability. Another difference between these two methods is found in the way they handle feature axes; LSM treats the set of axes as a sort of "template" of the corresponding class, while DMD treats it as a metric, i.e., a mapping from the original pattern space to each class-feature space.

5. CONCLUSION

This paper proposed a novel approach to pattern recognition, named Discriminative Metric Design (DMD), which fully designs the metric of each class discriminant function in a manner consistent with recognition error minimization. The experimental results in a vowel recognition task clearly demonstrated its high utility. Moreover, a comparison study of the relationships between DMD and several other recognition methods provided quite a useful basis for future theoretical analysis and a clear perspective on feature representation. It is lastly worth noting that one can easily apply the DMD formulation presented in this paper to *dynamic* (variable-durational) patterns by using a state transition structure like an HMM.

6. REFERENCES

[1] A. Biem and S. Katagiri, "Feature extraction based on minimum classification error/ generalized probabilistic descent method", *Proc. ICASSP 93*, vol. 2, pp. 275-278, Apr., 1993.

[2] G. H. Golub and C. F. Van Loan, *Matrix Computations*, The Johns Hopkins University Press, 1989.

[3] B.-H. Juang and S. Katagiri, "Discriminative learning for minimum error classification", *IEEE Trans. Signal Processing*, vol. 40, No. 12, pp. 3043-3054, Dec. 1992.

[4] S. Katagiri, C.-H. Lee, and B.-H. Juang, "Discriminative Multi-Layer Feed-Forward Networks", in *Proc. 1991 IEEE Workshop on Neural Networks for Signal Processing*, pp. 11-20, Princeton, NJ, Sept. 1991.

[5] S. Katagiri, C.-H. Lee, and B.-H. Juang, "New discriminative training algorithms based on the generalized probabilistic descent method", in *Proc. 1991 IEEE Workshop on Neural Networks for Signal Processing*, pp. 299-308, Princeton, NJ, Sept. 1991.

[6] W. Chou, B.H. Juang, and C.H. Lee, "Segmental GPD training of HMM based speech recognizer", *Proc. ICASSP 92*, vol. 1, pp. 473-476, 1992.

[7] E. Oja, *Subspace Methods of Pattern Recognition*, Research Studies Press, 1983.