Discriminative Feature Selection by Optimal Manifold Search for Neoplastic Image Recognition

Hayato Itoh[1], Yuichi Mori[2], Masashi Misawa[2], Masahiro Oda[1], Shin-Ei Kudo[2], and Kensaku Mori[1,3,4]

[1] Graduate School of Informatics, Nagoya University, Japan
[2] Digestive Disease Center, Showa University Northern Yokohama Hospital, Japan
[3] Information Technology Center, Nagoya University, Japan
[4] Research Center for Medical Bigdata, National Institute of Informatics, Japan

Abstract. An endocytoscope provides ultramagnified observation that enables physicians to achieve minimally invasive and real-time diagnosis in colonoscopy. However, great pathological knowledge and clinical experiences are required for this diagnosis. The computer-aided diagnosis (CAD) system is required that decreases the chances of overlooking neoplastic polyps in endocytoscopy. Towards the construction of a CAD system, we have developed texture-feature-based classification between neoplastic and non-neoplastic images of polyps. We propose a feature-selection method that selects discriminative features from texture features for such two-category classification by searching for an optimal manifold. With an optimal manifold, where selected features are distributed, the distance between two linear subspaces is maximised. We experimentally evaluated the proposed method by comparing the classification accuracy before and after the feature selection for texture features and deep-learning features. Furthermore, we clarified the characteristics of an optimal manifold by exploring the relation between the classification accuracy and the output probability of a support vector machine (SVM). The classification with our feature-selection method achieved 84.7% accuracy, which is 7.2% higher than the direct application of Haralick features and SVM.

Keywords: Feature selection, manifold learning, texture feature, convolutional neural network, endocytoscopic images, automated pathological diagnosis

1 Introduction

An endocytoscope was recently developed as a new endoscopic imaging modality for minimally-invasive diagnosis. Endocytoscopy enables direct observation of the cells and their nuclei on the colon wall at a 500-time-maximum ultramagnification as shown in 1. However, great pathological knowledge and clinical experiences are necessary to achieve accurate endocytoscopy. Automated pathological diagnosis is required to prevent overlooking neoplastic lesions to support

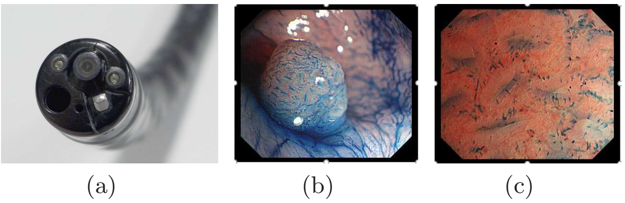


Fig. 1. (a) Endocytoscopy (CF-H290ECI, Olympus, Tokyo). (b) Conventional endoscope observation of a polyp by an endocytoscope. (c) Ultramagnified view by an endocytoscope. Small blue spots represent cell nuclei. In (b) and (c), a polyp's surface is stained by methylene blue.

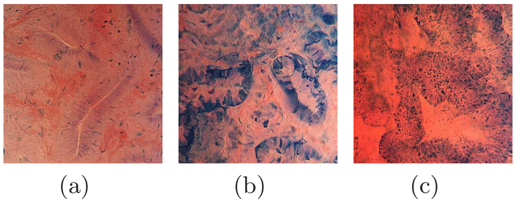


Fig. 2. Typical examples of endocytoscopic images for neoplastic- and non-neoplastic polyps. (a) and (b), and (c) are categorised to neoplastic and non-neoplastic polyps.

physicians [19, 18, 12]. This automated pathological diagnosis is achieved by robust two-category image classification. Figure 2 shows typical examples of neoplastic and non-neoplastic endocytoscopic images. The differences between the two categories are observed as differences of textures as shown in Fig. 2.

Robust two-category classification is a fundamental problem in pattern recognition, since multi-category classification is also based on a two-category classification concept. A robust classification can be achieved by an optimal pipeline of feature extraction, feature selection, and the classification of the selected features. A recent approach in image pattern recognition adopt deep learning architectures [16, 15, 23] as a full pipeline from feature extraction to the classification of extracted features. This deep learning approach can achieve robust multi-category classification with sufficiently large training dataset. However, deep learning fails to find optimal parameters with a small dataset. In medical image classification, collecting a large amount of training data with sufficient patient cases is difficult. Therefore, we have to tackle this problem with medical image classification, especially for a new medical modality, using a handcraft feature and a robust classifier. Previous works [19, 18, 12, 26, 25] adopted texture-based feature extraction and support-vector-machine classification without feature selection.

Principal component analysis (PCA) is a fundamental methodology for analysis and compression in data processing [14]. PCA is also used for feature se-

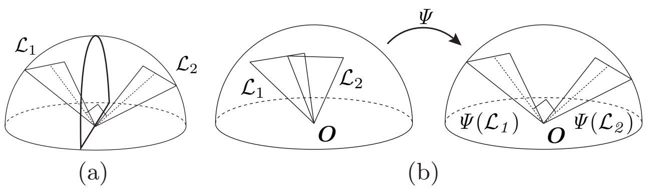


Fig. 3. Interpretation of relation between category subspace by canonical angles between them: (a) Linear separation by hyper plane for ideal features; (b) Before and after feature selection.

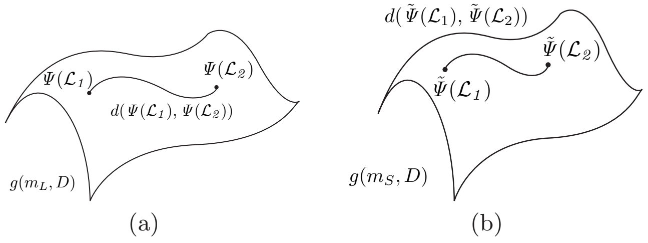

$\Psi(\mathcal{L}_1)$
$\Psi(\mathcal{L}_2)$
$d(\Psi(\mathcal{L}_1), \Psi(\mathcal{L}_2))$
$g(m_L, D)$
(a)
$d(\tilde{\Psi}(\mathcal{L}_1), \tilde{\Psi}(\mathcal{L}_2))$
$\tilde{\Psi}(\mathcal{L}_2)$
$\tilde{\Psi}(\mathcal{L}_1)$
$g(m_S, D)$
(b)

Fig. 4. Grassmann distance between two category subspaces. (a) and (b) show Grassmann distance between category subspaces in D-dimensional space for different feature selections ψ and $\tilde{\psi}$. In (a), Grassmann distance for m_L-dimensional subspace is longer than the one for m_S-dimensional subspace in (b). An optimal manifold gives largest distance between two-category subspaces for classification.

lection [20, 2], through which finds a small number of principal components to represent the patterns in a category. This feature selection is useful and optimal for the representation of the distribution of a pattern with respect to the mean square root error. However, this feature selection is not optimal for the classification of patterns in different categories. The common or similar principal components among patterns in different categories can lead to incorrect classification. Fukunaga and Koontz proposed feature selection methods using PCA for clustering [7]. Their method removes the features shared by two categories from the features. The validity of their method is experimentally presented with phantom data, where the means of each category patterns are known. Fukui et al. proposed a constraint mutual subspace method [5] with which they tried to remove the common subspace among patterns of different categories. However, this methods was only designed for the mutual subspace method [17].

We propose a new feature-selection method for the linear classification of neoplastic and non-neoplastic lesions on endocytoscopic images. This selection method is achieved by searching for an optimal manifold for classification. Figure 3 summarises the concept of our feature-selection method. Triangles $\mathcal{L}_1$ and $\mathcal{L}_2$ represent the linear category subspaces spanned by the features of each category. The ideal discriminative feature gives a hyper-plane between the features of the

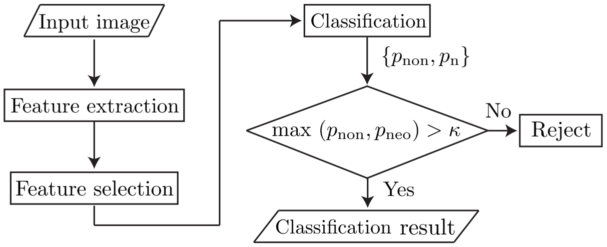


Fig. 5. Processing flow of classification of an endocytoscopic image. We use texture features for feature extraction. Classification is achieved by linear support vector machine with probability estimation. In the classification, we utilise rejection option with estimated probabilities. To selection the discriminative features, we propose a new selection method.

two categories, as shown in Fig. 3(a). This ideal feature extraction gives robust classification by a linear classifier. The difference between the two category subspaces can be represented by the canonical angle between two subspaces. If a feature contains worthless elements, it gives a common subspace for two categories with small canonical angles. These worthless elements can lead to classification errors. The overlap region of the triangles in Fig. 3(b) shows the common subspace between two categories. For feature selection, we have to find a map Ψ that maximises the canonical angles between two-category subspaces on a manifold as shown in Fig. 3(b). By projection with Ψ, we obtained a discriminative features as shown in Fig. 3(a). To find linear map Ψ, we used feature normalisation [7] and the Grassmann distance [8, 4, 10, 22].

Feature normalisation clarifies the importance of each eigenvector for category representations in PCA [7]. The Grassmann distance represents the difference of two linear subspaces by canonical angles [4, 8, 1, 22, 10, 24]. We obtained discriminative features by selecting the eigenvectors of the normalised features, which give the maximum Grassmann distance between two category subspaces as shown in Fig. 4. Our proposed method can analyse the extracted features and improve the classification accuracy. Furthermore, we integrated our proposed method to the processing flow of the classification shown in Fig. 5 that output the probabilities of each category for practical applications. In such practical applications as computer-aided diagnosis, accurate probabilities for each category are helpful information for physicians during a diagnosis. We evaluated the proposed method and its classification procedure to the classify neoplastic and non-neoplastic colorectal endocytoscopic images in numerical experiments.

2 Mathematical Preliminaries

2.1 Linear Subspace of a Pattern

We have a set $\{\boldsymbol{x}_i\}_i^N$ of N extracted D-dimensional features from observed patterns $\{\mathcal{X}_i\}_{i=1}^N$ of a category with the condition $d \ll N$. These extracted features span $\mathcal{L} = \mathrm{span}(\boldsymbol{x}_1, \dots, \boldsymbol{x}_N)$ for a category. We call this linear subspace a category subspace, which is approximated by

$$\mathcal{L} \approx \mathrm{span}(\boldsymbol{y}_1, \dots, \boldsymbol{y}_m), \tag{1}$$

where $m \leq D$. In such standard approaches as subspace methods [11, 28, 20, 17], we obtain bases $\boldsymbol{y}_i, i = 1, 2, \dots, m$ from the following eigendecomposition

$$\boldsymbol{M}\boldsymbol{y} = \eta \boldsymbol{y}, \quad \boldsymbol{M} = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{x}_i \boldsymbol{x}_i^\top. \tag{2}$$

For this eigendecomposition, an eigenvector $\boldsymbol{y}_i$ correspond to an eigenvalue η_i with conditions $\eta_1 \geq \eta_2 \geq \dots \eta_D$ and $\eta_i \geq 0$. We project an input feature vector $\boldsymbol{x}$ to a linear subspace $\mathrm{span}(\boldsymbol{y}_1, \dots, \boldsymbol{y}_m)$ by $\boldsymbol{Y}^\top \boldsymbol{x}$, where $\boldsymbol{Y} = [\boldsymbol{y}_1, \boldsymbol{y}_2, \dots, \boldsymbol{y}_m]$.

2.2 Grassmannian and Canonical Angle

The Grassmannian manifold (Grassmannian) $\mathcal{G}(m, D)$ is the set of m-dimensional linear subspaces of $\mathbb{R}^D$ [8]. An element of $\mathcal{G}(m, D)$ can be represented by an orthogonal matrix $\boldsymbol{Y}$ of size D by m, where $\boldsymbol{Y}$ is comprised of the m basis vectors for a set of patterns in $\mathbb{R}^D$.

Let $\boldsymbol{Y}_1$ and $\boldsymbol{Y}_2$ be the orthogonal matrices of size $D \times m$. Canonical angles (principal angles) $0 \leq \theta_1 \leq \dots \leq \theta_m \leq \frac{\pi}{2}$ between two subspaces $\mathcal{L}_1 = \mathrm{span}(\boldsymbol{Y}_1)$ and $\mathcal{L}_2 = \mathrm{span}(\boldsymbol{Y}_2)$ are defined by

$$\begin{aligned} \cos\theta_k = \max_{\boldsymbol{y}_{1,k} \in \mathrm{span}(\boldsymbol{Y}_1)} \max_{\boldsymbol{y}_{2,k} \in \mathrm{span}(\boldsymbol{Y}_2)} \boldsymbol{y}_{1,k}^\top \boldsymbol{y}_{2,k}, \\ \text{s.t. } \boldsymbol{y}_{1,k}^\top \boldsymbol{y}_{1,i} = 0, \ \boldsymbol{y}_{2,k}^\top \boldsymbol{y}_{2,i} = 0, \end{aligned} \tag{3}$$

where $i = 1, 2, \dots, k-1$.

For two linear subspaces $\mathcal{L}_1$ and $\mathcal{L}_2$, we have projection matrices $\boldsymbol{P} = \boldsymbol{Y}_1 \boldsymbol{Y}_1^\top$ and $\boldsymbol{Q} = \boldsymbol{Y}_2 \boldsymbol{Y}_2^\top$. We also have a set of canonical angles $\{\theta_k\}_{i=1}^m$ between these two linear subspaces with conditions $\theta_1 \leq \theta_2 \leq \dots . \theta_K$. We obtain the canonical angles from the solution of the eigendecomposition problem

$$\boldsymbol{P}\boldsymbol{Q}\boldsymbol{P}\boldsymbol{u} = \lambda \boldsymbol{u} \text{ or } \boldsymbol{Q}\boldsymbol{P}\boldsymbol{Q}\boldsymbol{u} = \lambda \boldsymbol{u}. \tag{4}$$

The solutions of these eigendecomposition problems are coincident [17]. For the practical computation of canonical angles, we have singular value decomposition

$$\boldsymbol{Y}_1^\top \boldsymbol{Y}_2 = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top, \tag{5}$$

where $\boldsymbol{\Sigma} = \mathrm{diag}(\cos\theta_1, \cos\theta_2, \dots, \cos\theta_\mathrm{m})$ is the diagonal matrix, and $\boldsymbol{U}, \boldsymbol{V} \in \mathbb{R}^{m\times m}$ are the orthogonal matrices. Note that we have the following relation,

$$\lambda_i = \cos^2\theta_i \tag{6}$$

for eigenvalues $\lambda_k,\ i = 1, 2, \dots m$ in the eigendecomposition in Eq. (4). Canonical angles are used to define the geodesic distance on a Grassmannian.

2.3 Grassmannian Distances

For the two points represented by $\boldsymbol{Y}_1$ and $\boldsymbol{Y}_2$ on a Grassmann manifold, we have the following seven distances on the Grassmannian,

1. $d_p(\boldsymbol{Y}_1, \boldsymbol{Y}_2)$: projection metric
2. $d_\mu(\boldsymbol{Y}_1, \boldsymbol{Y}_2)$: mean distance
3. $d_\mathrm{min}(\boldsymbol{Y}_1, \boldsymbol{Y}_2)$: minimum canonical angle
4. $d_\mathrm{max}(\boldsymbol{Y}_1, \boldsymbol{Y}_2)$: maximum canonical angle
5. $d_{BC}(\boldsymbol{Y}_1, \boldsymbol{Y}_2)$: Binet-Caucy metric
6. $d_g(\boldsymbol{Y}_1, \boldsymbol{Y}_2)$: geodesic distance
7. $d_c(\boldsymbol{Y}_1, \boldsymbol{Y}_2)$: Procrustes (chordal) distance

These seven distances are defined using the canonical angles between two linear subspaces. The projection metric and mean distance are defined by

$$d_p(\boldsymbol{Y}_1, \boldsymbol{Y}_2) = \left(\sum_{i=1}^{m} \sin^2\theta_i\right)^{1/2} = \left(m - \sum_{i=1}^{m}\lambda_i\right)^{1/2} \tag{7}$$

and

$$d_\mu(\boldsymbol{Y}_1, \boldsymbol{Y}_2) = \frac{1}{m}\sum_{i=1}^{m} \sin^2\theta_i = 1 - \frac{1}{m}\sum_{i=1}^{m}\lambda_i. \tag{8}$$

Furthermore, the sines of the maximum and minimum canonical angles

$$d_\mathrm{min}(\boldsymbol{Y}_1, \boldsymbol{Y}_2) = \sin\theta_1 = (1-\lambda_1)^{1/2}, \tag{9}$$

$$d_\mathrm{max}(\boldsymbol{Y}_1, \boldsymbol{Y}_2) = \sin\theta_\mathrm{m} = (1-\lambda_m)^{1/2}, \tag{10}$$

are also used as distances on a Grassmannian. Moreover, the Binet-Caucy distance is defined by

$$d_{BQ}(\boldsymbol{Y}_1, \boldsymbol{Y}_2) = \left(1 - \mathit{\Pi}_{i=1}^{m}\cos^2\theta_i\right) = \left(1 - \mathit{\Pi}_{i=1}^{m}\lambda_i\right), \tag{11}$$

where the product of $\cos\theta_i$ represents the similarity between two linear subspaces. Using canonical angles, we have geodesic distance

$$d_g(\boldsymbol{Y}_1, \boldsymbol{Y}_2) = \left(\sum_{i=1}^{m}\theta_i^2\right)^{1/2} = \left(\sum_{i=1}^{m}\left(\arccos\lambda_i^{1/2}\right)^2\right)^{1/2} \tag{12}$$

on a Grassmann manifold. We have two definitions of the Procrustes (chordal) distance,

$$d_c(\boldsymbol{Y}_1, \boldsymbol{Y}_2) = \min_{\boldsymbol{R}_1, \boldsymbol{R}_2 \in O(m)} \|\boldsymbol{Y}_1 \boldsymbol{R}_1 - \boldsymbol{Y}_2 \boldsymbol{R}_2\|_{\mathrm{F}} = 2 \left(\sum_{i=1}^{m} \sin^2(\theta_i/2) \right)^{1/2}, \quad (13)$$

where $\| \cdot \|_{\mathrm{F}}$ is the Frobenius norm.

2.4 Normalisation of Features

We project the extracted features onto the discriminative feature space for accurate classification. We extract a set of features $\{\boldsymbol{x}_i \in \mathbb{R}^D\}_{i=1}^N$ from a set of images with condition $D \ll N$. Let $\boldsymbol{\mu} = \mathrm{E}(\boldsymbol{x}_i)$ be the mean of the features. By setting $\bar{\boldsymbol{x}}_i = \boldsymbol{x}_i - \boldsymbol{\mu}$, we obtain a set of centred features $\{\bar{\boldsymbol{x}}_i\}_{i=1}^N$. We assume each image belongs to either category $\mathcal{C}_1$ or $\mathcal{C}_2$. Therefore, set $\{\bar{\boldsymbol{x}}_i\}_{i=1}^N$ is divided into two sets $\{\boldsymbol{x}_i^{(1)}\}_{i=1}^{N_1}$ and $\{\boldsymbol{x}_i^{(2)}\}_{i=1}^{N_2}$, where N_1 and N_2, respectively, represent the number of images in the first and second categories.

We define the autocorrelation matrices in the centred feature space as

$$\boldsymbol{A}_1 = \frac{1}{N_1} \boldsymbol{X}_1 \boldsymbol{X}_1^\top, \ \boldsymbol{A}_2 = \frac{1}{N_2} \boldsymbol{X}_2 \boldsymbol{X}_2^\top, \quad (14)$$

where $\boldsymbol{X}_1 = [\boldsymbol{x}_1^{(1)}, \boldsymbol{x}_2^{(1)}, \dots, \boldsymbol{x}_{N_1}^{(1)}]$ and $\boldsymbol{X}_2 = [\boldsymbol{x}_1^{(2)}, \boldsymbol{x}_2^{(2)}, \dots, \boldsymbol{x}_{N_2}^{(2)}]$, for the two categories. We define the covariance matrix of all the features as

$$\boldsymbol{C} = P(\mathcal{C}_1) \boldsymbol{A}_1 + P(\mathcal{C}_2) \boldsymbol{A}_2, \quad (15)$$

where we set $P(\mathcal{C})_1 = \frac{N_1}{N_1 + N_2}$ and $P(\mathcal{C})_2 = \frac{N_2}{N_1 + N_2}$. Fukunaga [7] used the autocorrelation matrix of all the features instead of $\boldsymbol{C}$ in Eq. (15). Fukunaga [6] used covariance matrices $\boldsymbol{C}_1$ and $\boldsymbol{C}_2$ for the two categories instead of $\boldsymbol{A}_1$ and $\boldsymbol{A}_2$ in Eq. (15). In this manuscript, we adopt covariance matrix $\boldsymbol{C}$ for all features to remove the common features of the two categories. Here, autocorrelation matrices $\boldsymbol{A}_1$ and $\boldsymbol{A}_2$ include the gaps of the means between all the features and each category. The covariance matrix gives the following eigendecomposition problem $\boldsymbol{C}\boldsymbol{V} = \boldsymbol{V}\boldsymbol{\Xi}$, where $\boldsymbol{\Xi} = \mathrm{diag}(\xi_1, \xi_2, \dots, \xi_D)$ consists of eigenvalues ξ_i for $i = 1, 2, \dots, D$ the condition $\xi_1 \geq \xi_2 \geq \cdots \geq \xi_D$. The eigendecomposition results derive a whitening matrix $\boldsymbol{W} = \boldsymbol{\Xi}^{-\frac{1}{2}} \boldsymbol{V}^\top$.

Using this whitening matrix and Eq. (15), we obtain the following relation

$$\boldsymbol{W}\boldsymbol{C}\boldsymbol{W}^\top = \boldsymbol{W} P(\mathcal{C}_1) \boldsymbol{A}_1 \boldsymbol{W}^\top + \boldsymbol{W} P(\mathcal{C}_2) \boldsymbol{A}_2 \boldsymbol{W}^\top = \tilde{\boldsymbol{A}}_1 + \tilde{\boldsymbol{A}}_2 = \boldsymbol{I}, \quad (16)$$

where $\boldsymbol{I}$ is an identity matrix. The solutions of the eigenvalue problems

$$\tilde{\boldsymbol{A}}_j \boldsymbol{\phi}_i^{(j)} = \lambda_i^{(j)} \boldsymbol{\phi}_i^{(j)}, \quad (17)$$

where we set $j = 1, 2$, give the bases of two category subspaces. From Eqs. (16) and (17), we have

$$\tilde{\boldsymbol{A}}_2 \boldsymbol{\phi}_i^{(2)} = (\boldsymbol{I} - \tilde{\boldsymbol{A}}_1) \lambda_i \boldsymbol{\phi}_i^{(2)}. \quad (18)$$

This leads $\tilde{\boldsymbol{A}}_1 \boldsymbol{\phi}_i^{(2)} = (1 - \lambda_i^{(2)}) \boldsymbol{\phi}_i^{(2)}$. These relation give the following relations

$$\boldsymbol{\phi}_i^{(2)} = \boldsymbol{\phi}_{D-i+1}^{(1)} \tag{19}$$

and

$$\lambda_i^{(2)} = 1 - \lambda_{D-i+1}^{(1)}. \tag{20}$$

Equations (19) and (20) show that both eigenvalue problems give the same set of eigenvectors, and corresponding eigenvalues. Note that the two sets of eigenvalues are reversely ordered. The eigenvalue orders in Eq. (20) satisfy

$$1 \geq \lambda_1^{(1)} \geq \lambda_2^{(1)} \geq \cdots \geq \lambda_D^{(1)} \geq 0, \tag{21}$$

$$0 \leq 1 - \lambda_1^{(1)} \leq 1 - \lambda_2^{(1)} \leq \cdots \leq 1 - \lambda_D^{(1)} \leq 1. \tag{22}$$

These relations imply that the eigenvectors corresponding to the large eigenvalues of $\tilde{\boldsymbol{A}}_1$ contribute to represent the subspace for $\mathcal{C}_1$, although they only make minor contribution to the representation for $\mathcal{C}_2$. Therefore, we obtained discriminative features by projecting the features to a linear subspace given by $\mathrm{span}(\boldsymbol{\phi}_1^{(1)}, \boldsymbol{\phi}_1^{(2)}, \boldsymbol{\phi}_2^{(1)}, \boldsymbol{\phi}_2^{(2)}, \dots, \boldsymbol{\phi}_d^{(1)}, \boldsymbol{\phi}_d^{(2)})$. We discuss how to decide number d in the next section.

2.5 Linear Classification with Rejection Option

We use a linear support vector machine (SVM) as a classifier. SVM classifies an input feature vector $\boldsymbol{x} \in \mathbb{R}^D$ by $\mathrm{sing}(f(\boldsymbol{x}))$, where $f(\cdot)$ is a decision function. The parameters and the hyperparameters of this decision function are optimised by a training procedure with training data. We can estimate the probability of belonging to each category with the optimised decision function. For two categories with label $\mathcal{L} \in \{0, 1\}$, we can approximately estimate the probabilities

$$P(\mathcal{L} = 1 | \boldsymbol{x}) \approx P(A, B, f(x)) = \frac{1}{1 + \exp(Af(x) + B)}, \tag{23}$$

where A, B are the parameters in Platt's method [21] for category i. $P(\mathcal{L} = 0 | \boldsymbol{x})$ is obtained by $1 - P(\mathcal{L} = 1 | \boldsymbol{x})$. After the training for decision function, we obtain A, B by maximum likelihood estimation with the training dataset. In our method, we represent the non-neoplastic and neoplastic categories by 0 and 1. SVM can output the probabilities [21] p_{non} and p_{nneo} that satisfy $p_{\mathrm{non}} + p_{\mathrm{neo}} = 1$ and $p_{\mathrm{non}}, p_{\mathrm{non}} \in [0, 1]$, for non-neoplastic and neoplastic images. We adopt the rejection option to remove low confident classification [2]. The rejection option discards classifications with low probabilities such that $\max\,(p_{\mathrm{non}}, p_{\mathrm{neo}}) < \kappa$, where criteria κ was decided from preliminary experiments.

3 Feature-Selection Method

We have a set $\{\mathcal{X}_i\}_{i=1}^N$ of three-channel images. We extract a feature vector $\boldsymbol{x}_i \in \mathbb{R}^D$ from each image $\mathcal{X}_i$. As in the same manner of Section 2.4, we divide

Algorithm 1: Feature-selection method for training data

```
Input: Two sets of feature vectors $\{\boldsymbol{x}_i^{(1)}\}_{i=1}^{N_1}$ and $\{\boldsymbol{x}_i^{(2)}\}_{i=1}^{N_2}$,
       criteria $\tau_k = 1.0 - 0.01 * k,\ k = 1, 2, \dots, 50$, a small value $\varepsilon$.
Output: Projected features $\{\breve{\boldsymbol{x}}_i^{(1)}\}_{i=1}^{N_1}$ and $\{\breve{\boldsymbol{x}}_i^{(2)}\}_{i=1}^{N_2}$, matrices $\boldsymbol{P}^*$ and $\boldsymbol{W}$.
1. Compute two sets of eigenvectors $\{\boldsymbol{\phi}_i^{(1)}\}_{i=1}^{d}$ and $\{\boldsymbol{\phi}_i^{(2)}\}_{i=1}^{d}$.
2. For all $\tau_k$:
  2-1. Select eigenvectors $\{\boldsymbol{\phi}_i^{(1)}\}_{i=1}^{m_k}$ and $\{\boldsymbol{\phi}_i^{(2)}\}_{i=1}^{m_k}$ that
       correspond to eigenvalues larger than $\tau_k$.
  2-2. Construct a matrix $\boldsymbol{P}^{(k)} = [\boldsymbol{\phi}_1^{(1)}, \boldsymbol{\phi}_1^{(2)}, \dots, \boldsymbol{\phi}_{m_k}^{(1)}, \boldsymbol{\phi}_{m_k}^{(2)}]^\top$.
  2-3. Project all features by $\breve{\boldsymbol{x}}_i^{(j)} = \boldsymbol{P}^{(k)}\boldsymbol{x}_i^{(j)}$, where $i = 1, 2, \dots, N_j,\ j = 1, 2$.
  2-4. Compute eigenvectors $\boldsymbol{Y}_1^{(k)}$ and $\boldsymbol{Y}_2^{(k)}$ for two
       categories by solving the eigenproblem in Eq. (2).
  2-5. Compute Grassmann distance $d(\boldsymbol{Y}_1^{(k)}, \boldsymbol{Y}_2^{(k)})$ between two category subspaces.
  2-6. If $|d(\boldsymbol{Y}_1^{(k)}, \boldsymbol{Y}_2^{(k)}) - d(\boldsymbol{Y}_1^{(k-1)}, \boldsymbol{Y}_2^{(k-1)})| < \varepsilon$, set $\boldsymbol{P}^* = \boldsymbol{P}^{(k)}$, and iteration break.
3. Return projected feature vectors $\{\breve{\boldsymbol{x}}_i^{(j)} | \breve{\boldsymbol{x}}_i^{(j)} = \boldsymbol{P}^*\boldsymbol{W}\boldsymbol{x}_i^{(j)} \forall i, j\}$, and $\boldsymbol{P}^*$ and $\boldsymbol{W}$.
```

$\{\boldsymbol{x}_i\}_{i=1}^{N}$ to two sets $\{\boldsymbol{x}_i^{(1)}\}_{i=1}^{N_1}$ and $\{\boldsymbol{x}_i^{(2)}\}_{i=1}^{N_2}$, where N_1 and N_2 represent the number of images in the first and second categories.

We find a linear map $\Psi(\boldsymbol{x}) = \boldsymbol{P^*W(x - \mu)}$ by solving

$$\arg\min_{\boldsymbol{P}}(\mathrm{rank}(\boldsymbol{P}\boldsymbol{P}^\top)) \quad \text{s.t.} \quad \max_{\boldsymbol{P}}(d(\Psi(\boldsymbol{Y}_1), \Psi(\boldsymbol{Y}_2))) \tag{24}$$

where $d(\cdot, \cdot)$ represents one of the seven Grassmann distances. In this equation, $\max(\cdot)$ returns one or more matrices that give the maximum distance, and $\min(\cdot)$ selects one matrix with the minimum rank from the matrices. The solution is an orthogonal matrix $\boldsymbol{P}^* = [\boldsymbol{\phi}_1^{(1)}, \boldsymbol{\phi}_1^{(2)}, \boldsymbol{\phi}_2^{(1)}, \boldsymbol{\phi}_2^{(2)}, \dots, \boldsymbol{\phi}_m^{(1)}, \boldsymbol{\phi}_m^{(2)}]^\top$, which is comprised of $2m$ eigenvectors. Algorithm 1, which summarises a procedure to find the $\boldsymbol{P}^*$ using the training data, also returns the selected features for the training data.

4 Numerical Experiments

We evaluated our feature-selection method by applying the image classification of neoplastic and non-neoplastic images of a colon polyp. We used images of the magnified surfaces of polyps captured by an endocytoscope (CF-H290ECI, Olympus, Tokyo) with IRB approval. The neoplastic and non-neoplastic labels of the images ware annotated by expert physicians. The number of images of neoplastic and non-neoplastic polyps is summarised in Table 1. We extracted two kinds of features from these images: the Haralick features [9] and the convolutional neural network (CNN) features [15]. In this section, we first compare the classification accuracy before and after the feature selection for the Haralick features and next compare the classification accuracy before and after the feature selection of the combined Haralick and CNN features. We finally analysed the

Table 1. Dataset details: 14,840 training and 4,126 test images.

Category	♯ training	♯ test	♯ total
Neoplasia	7,800	1,925	9,725
Non-neoplasia	7,040	2,201	9,241

Table 2. Dimension and classification accuracy for extracted features with SVM.

	Haralick	CNN	Haralick+CNN
Dimension	312	576	888
Accuracy [%]	77.3	75.6	78.3

Table 3. Dimension and classification accuracy of selected features of Haralick feature for each Grassmann distance.

	d_g, d_p, d_c	$d_{\mathrm{BC}}, d_{\max}, d_{\min}$	d_μ
Dimension	312 ($\tau = 0.5$)	60 ($\tau = 0.99$)	96 ($\tau = 0.77$)
Accuracy [%]	77.5	76.4	78.0

Table 4. Dimension and classification accuracy of selected features of combination of Haralick and CNN features for each Grassmann distance.

	d_g, d_p, d_c	$d_{\mathrm{BC}}, d_{\max}$	$d_{\min}, d_\mu$
Dimension	888 ($\tau = 0.5$)	74 ($\tau = 0.99$)	102 ($\tau = 0.97$)
Accuracy [%]	79.7	78.4	77.7

relation between the output probability and the accuracy with rejection option. The relation clarifies the validity of our feature-selection method as a manifold optimisation. In the final analysis, we also show the performance of two-category classification with a rejection option.

In these evaluations, we used a SVM [27] for the classifications and trained it by the training data and the best hyperparameters. We obtained the best hyperparameters of the SVM by five-fold cross validation with the training data. For a practical computation of SVM, we used libsvm [3]. We note that kernel SVM with a radius basis function gives less classification accuracy than the linear SVM for the endocytoscopic images in our preliminary experiments. The classification accuracy of the original Haralick, CNN and their combined features without feature selection is summarised in Table 2.

4.1 Haralick Feature

The Haralick feature represents the texture information measured by fourteen statistical categories of a gray-level co-occurrence matrix. In this case, we used thirteen categories for eight directions with three scales. To compute the statistics, we used contrast normalisation for each local region for the achievement of robustness against illumination changes. We then extracted 312-dimensional Haralick feature vector for an image, each element of which is normalised to a range of $[0, 1]$. We applied the proposed feature-selection method to the normalised Haralick features.

Figures 7(a), and (b) and (c) show the eigenvalues of the whitened autocorrelation matrices for two categories, and the Grassmann distances. Table 3 summarises the dimensions and classification accuracy for each feature selection with respect to seven Grassmann distances.

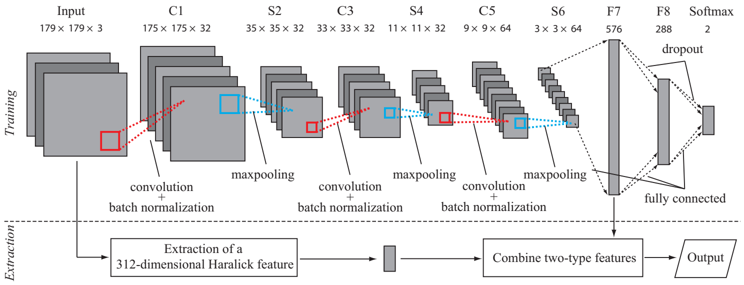


Fig. 6. Architecture of convolutional neural network for feature extraction.

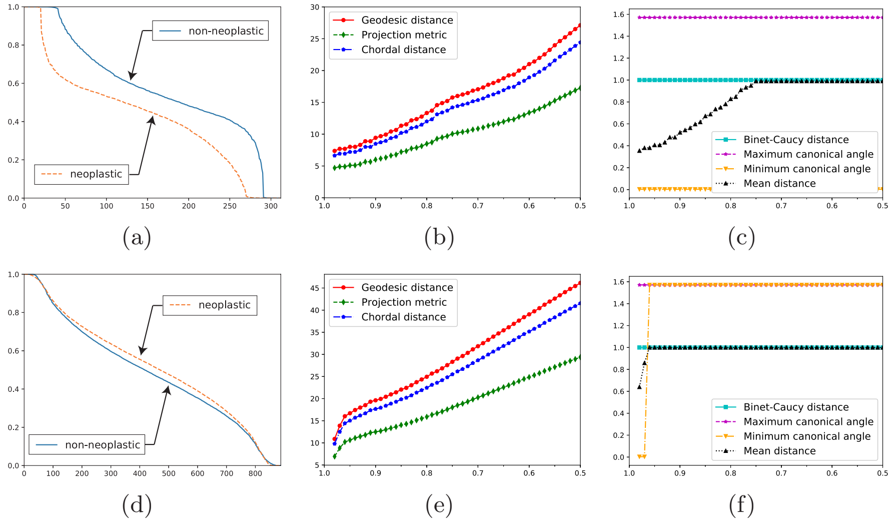
4.2 Combination of Haralick and CNN features

We combined the Haralick and CNN features into a single feature vector. Figure 6 illustrates the architecture of our CNN. Its settings were decided by preliminary experiments [12], where we compared the several parameter settings with AlexNet [15], VGG Net [23], and their modified versions. Our shallow architecture gave the best classification accuracy among them. Before the extraction of the CNN features, we trained the CNN with training data from scratch. We applied batch normalisation and drop out to convolutional layers and full connected layers for the training. We used the values in full connection layer F7 in Fig. 6 for the feature extraction. Each element of the extracted CNN feature was normalised to a range of $[0, 1]$. For the CNN implementation, we used the Caffe platform [13] and and combined the normalised Haralick and normalised CNN features as 888-dimensional column vectors.

We applied the proposed method to these normalised combined features. Figures 7(d), and (e) and (f) show the eigenvalues of the whitened autocorrelation matrices for two categories, and the Grassmann distances. Table 4 summarises the dimensions of the selected feature and classification accuracy for the selection with respect to seven Grassmann distances.

4.3 Analysis of an Optimal Manifold

We evaluated the classification accuracy with respect to the divided range of the output probabilities. The relation between the classification accuracy and the output probability represents the characteristics of the optimal manifold given by our feature selection. Figure 8(a) illustrates the distributions of the output probabilities for the original Haralick features and the selected features. Table 5 summarises the accuracy for each range of the output probabilities for the original Haralick features and the selected features. We also evaluated the classification accuracy with the rejection option of $\kappa \in \{0.50, 0.55, \dots, 0.90\}$. The rejection rate is the ratio of the rejected images in the test data. The classification results with the rejection option are summarised in Fig. 8(b). We respectively



(a) (b) (c)

(d) (e) (f)

Fig. 7. Eigenvalues and Grassmann distance in feature selection: Upper and lower rows represent results for Haralick features and combined features, respectively. (a) and (d) show eigenvalues of $\tilde{\boldsymbol{A}}_1$ and $\tilde{\boldsymbol{A}}_1$ for non-neoplastic and neoplastic images. Horizontal and vertical axes represent indices of eigenvalues and eigenvalues. (b),(c),(e) and (f) summarise Grassmann distance after projection with respect to $\boldsymbol{P}$ given by a criteria τ. Horizontal and vertical axes represent τ and Grassmann distance.

obtained classification accuracy of 82.1%, 82.9% and 84.7% for the original Haralick feature, the selected features of the Haralick and the combined features with the same rejection rate of about 20%. Note that the percentage of inappropriate images in practical diagnosis is close to 20-30%. In these cases, we set κ to 0.70, 0.60 and 0.72 for them. Figure 9 shows examples of rejected images in the classification.

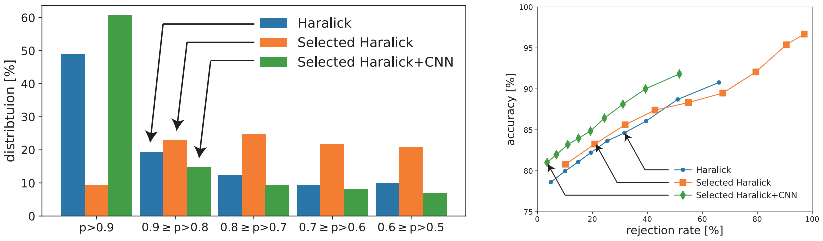
5 Discussion

The curves shown in Fig. 7(a) imply that a small number of eigenvectors is discriminative for classification, since almost all the eigenvalues are close to 0.5. Figures. 7(b) and (c) imply that d_μ gives the largest distance with the fewest selected eigenvectors. The accuracy is improved after the selection based on d_μ as shown in Tab. 3. The dimension of the selected features is 30% of the dimension of an original Haralick feature.

The curves shown in Fig. 7(d) imply that there are no particular discriminative eigenvectors, since they are distributed almost uniformly from zero to one. In Table 4, d_g, d_p, and d_c give the largest distance with all the eigenvectors. In this case, features are just whitened and used for classification. In both the

Table 5. Classification accuracy with respect to output probability. This table summarises classification accuracy for the original Haralick, the selected Haralick and the Haralick+CNN features. Classification accuracy is computed for each range of output probability p, where p represents $\max(p_{\mathrm{non}}, p_{\mathrm{neo}})$. Classification accuracy of selected Haralick is almost coincident to mean of each range of output probability.

	Original Haralick	Selected Haralick	Selected Haralick+CNN
$p > 0.9$	88.7%	95.4%	90.0%
$0.9 \geq p > 0.8$	74.3%	87.1%	72.0%
$0.8 \geq p > 0.7$	68.7%	84.7%	64.2%
$0.7 \geq p > 0.6$	60.5%	72.5%	61.0%
$0.6 \geq p > 0.5$	53.2%	58.4%	50.0%



(a) (b)

Fig. 8. Analysis of optimal manifold. (a) Distributions of output probabilities for original and selected features: Figure illustrates probability distributions of the original Haralick, the selected Haralick and the Haralick+CNN features. Vertical axis represents percentage of output probabilities for each range. Horizontal axis represent divided rages. In this figure, p is given as $\max(p_{\mathrm{non}}, p_{\mathrm{neo}})$. (b) Receiver operating characteristic (ORC) curves for classification accuracy and rejection rate: Vertical and horizontal axes represent classification accuracy and rejection rate for original Haralick, selected Haralick and Haralick+CNN features.

cases of the Haralick and the combined features, our proposed method found discriminative features and improved the classification accuracy.

The results summarised in Fig. 8(b) indicate that the rejection option improved the classification accuracy.The rejection option correctly removed the low-confident classification in both the cases of the Haralick and the combined features. Table 5 also supports the validity of the rejection option. We can observe low classification accuracy for low output probabilities. Figure 9 shows that the rejected images are inappropriate due to bad observation conditions: over staining, bad illumination, and a lack of discriminative texture.

The mean distance gives the best feature selection for the Haralick feature. In this case, the distributions of the output probabilities were characteristic. We have low distribution for $p > 0.9$ and medium distribution for $0.6 \geq p > 0.5$. The highest distribution is given for $0.8 \geq p > 0.7$. For these distributions, the output

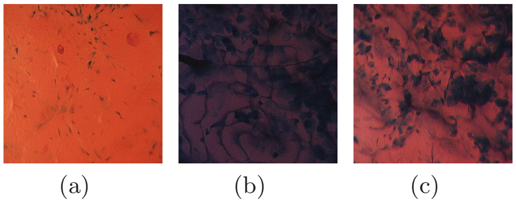


Fig. 9. Examples of rejected endocytoscopic images in classification: (a) insufficient texture for classification; (b) bad (too dark) illumination; (c) over staining of a polyp. These images were labelled as inappropriate for practical diagnosis. In practice, medical doctors also recognised them as inappropriate.

probability approximated the classification accuracy well, as shown gray in in Table 5. This characteristic of the relation between classification accuracy and output probability suggests the validity of the obtained manifold. In the case of the combined features, we did not observe the same characteristic even though the selected combined features achieved the highest classification accuracy.

6 Conclusions

We proposed a feature-selection method that improves the classification accuracy of two categories by an optimal manifold search for classification. We experimentally evaluated the proposed method by comparing the classification accuracy before and after feature selection for about 19,000 endocytoscopic images. The experimental results showed the validity of our proposed method with the improvement of classification accuracy. Furthermore, we experimentally demonstrated the validity of the obtained optimal manifold by exploring the relation between output probability and classification accuracy. The output probability is helpful information for practical diagnosis. Moreover, we achieved robust classification with our feature-selection method and a linear classifier with a rejection option. The classification accuracy was 84.7% with a rejection rate of 20%, in which the classification accuracy was 7.2 % higher than the classification with the original Haralick feature and SVM.

Acknowledgements

Parts of this research were supported by the Research on Development of New Medical Devices from the Japan Agency for Medical Research and Development (No. 18hk0102034h0103), and MEXT KAKENHI (No. 26108006, No. 17H00867).

References

1. Absil, P.A., Mahony, R., Sepulchre, R.: Optimization Algorithms on Matrix Manifolds. Princeton University Press (2009)
2. Bishop, C.M.: Pattern Recognition and Machine Learning. Springer (2006)
3. Chang, C.C., Lin, C.J.: LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology **2**, 27:1–27:27 (2011)
4. Edelman, A., Arias, T., Smith, S.: The geometry of algorithms with orthogonality constraints. SIAM J. Matrix Anal. Appl. **20**(2), 303–353 (1998)
5. Fukui, K., Maki, A.: Difference subspace and its generalization for subspace-based methods. IEEE Transactions on Pattern Analysis and Machine Intelligence **37**(11), 2164–2177 (2015)
6. Fukunaga, K.: *Introduction to Statistical Pattern Recognition (second edition)*. Academic Press (1990)
7. Fukunaga, K., Koontz, W.L.G.: Application of the Karhunen-Loéve expansion to feature selection and ordering. IEEE Transactions on Computers **C-19**(4), 311–318 (1970)
8. Hamm, J., Lee, D.: Grassmann discriminant analysis: A unifying view on subspace-based learning. Proc. International Conference on Machine Learning pp. 376–383 (2008)
9. Haralick, R.M., Shanmugam, K., Dinstein, I.: Textural features for image classification. IEEE Transactions on Systems, Man, and Cybernetics **SMC-3**(6), 610–621 (Nov 1973). https://doi.org/10.1109/TSMC.1973.4309314
10. Harandi, M., Sanderson, C., Shen, C., Lovell, B.: Dictionary learning and sparse coding on Grassmann manifolds: An extrinsic solution. Proc. The IEEE International Conference on Computer Vision pp. 3120–3127 (2013)
11. Iijima, T.: Theory of pattern recognition. Electronics and Communications in Japan pp. 123–134 (1963)
12. Itoh, H., Mori, Y., Misawa, M. Oda, M., Kudo, S.E., Mori, K.: Cascade classification of endocytoscopic images of colorectal lesions for automated pathological diagnosis. Proc. SPIE Medical Imaging (in Press) (2018)
13. Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T.: Caffe: Convolutional architecture for fast feature embedding. arXiv preprint arXiv:1408.5093 (2014)
14. Jollife, I.T.: Principal Component Analysis. Springer (2002)
15. Krizhevsky, A., Sutskever, I., Hinton, G.: Imagenet classification with deep convolutional neural networks. Proc. International Conference on Neural Information Processing Systems **1**, 1097–1105 (2012)
16. Lecun, Y., et al.: Gradient-based learning applied to document recognition. Proc. IEEE **86**(11), 2278–2324 (1998)
17. Maeda, K.: From the subspace methods to the mutual subspace method. In: *Computer Vision*, vol. 285, pp. 135–156. Springer (2010)
18. Mori, Y., Kudo, S.E., Chiu, P., Singh, R., Misawa, M., Wakamura, K. Kudo, T., Hayashi, T., Katagiri, A., Miyachi, H., Ishida, F., Maeda, Y., Inoue, H., Nimura, Y., Oda, M., Mori, K.: Impact of an automated system for endocytoscopic diagnosis of small colorectal lesions: an international web-based study. Endoscopy **48**, 1110–1118 (2016)
19. Mori, Y., Kudo, S.E., Wakamura, K., Misawa, M., Ogawa, Y., Kutsukawa, M., Kudo, T., Hayashi, T., Miyachi, H., Ishida, F., Inoue, H.: Novel computer-aided diagnostic system for colorectal lesions by using endocytoscopy. Gastrointestinal Endoscopy **81**, 621–629 (2015)

20. Oja, E.: *Subspace Methods of Pattern Recognition*. Research Studies Press (1983)
21. Platt, J.C.: Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In: Advances in large margin classifier. pp. 61–74. MIT Press (1999)
22. Shigenaka, R., Raytchev, B., Tamaki, T., Kaneda, K.: Face sequence recognition using Grassmann distances and Gassmann kernels. Poc. International Joint Conference on Neural Networks pp. 1–7 (2012)
23. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. Proc. International Conference on Learning Representations (2015)
24. Slama, R., Wannous, H., Daoudi, M., Srivastava, A.: Accurate 3D action recognition using learning on the Grassmann manifold. Pattern Recognition **48**(2), 556 – 567 (2015)
25. Tamaki, T., Sonoyama, S., T., H., Raytchev, B., Kaneda, K., Koide, K., Yoshida, S., Mieno, H., Tanaka, S.: Computer-aided colorectal tumor classification in NBI endoscopy using CNN features. Proc. Korea-Japan joint workshop on Frontiers of Computer Vision (2016)
26. Tamaki, T., Yoshimuta, J., Kawakami, M., Raytchev, B., Kaneda, K., Yoshida, S., Takemura, Y., Onji, K., Miyaki, R., Tanaka, S.: Computer-aided colorectal tumor classification in NBI endoscopy using local features. Mediacal Image Analysis **17**, 78–100 (2013)
27. Vapnik, V.N.: *Statistical Learning Theory*. Wiley (1998)
28. Watanabe, S., Pakvasa, N.: Subspace method of pattern recognition. Proc. of the 1st International Joint Conference of Pattern Recognition (1973)