A Bayesian Network Classifier and Hierarchical Gabor Features for Handwritten Numeral Recognition

JaeMo Sung, Sung-Yang Bang, Seungjin Choi [*]

Department of Computer Science
Pohang University of Science and Technology
San 31 Hyoja-dong, Nam-gu, Pohang 790-784, Korea

Abstract

We present a method of handwritten numeral recognition, where we introduce hierarchical Gabor features (HGFs) and construct a Bayesian network classifier that encodes the dependence between HGFs. We extract HGFs in such a way that they represent different levels of information which are structured such that the lower the level is, the more localized information they have. At each level, we choose an optimal set of 2-D Gabor filters in the sense that Fisher's linear discriminant (FLD) measure is maximized and these Gabor filters are used to extract HGFs. We construct a Bayesian network classifier that encodes hierarchical dependence among HGFs. We confirm the useful behavior of our proposed method, comparing it with the naive Bayesian classifier, k-nearest neighbor, and an artificial neural network, in the task of handwritten numeral recognition.

Key words: Bayesian networks, Gabor filters, Handwritten numeral recognition, Hierarchical models.

1 Introduction

The problem of pattern recognition consists of two important parts which are feature extraction and classification. A variety of methods of feature extraction

[*] Corresponding author. Tel.: +82-54-279-2259; Fax: +82-54-279-2299
Email: emtidi@postech.ac.kr (J. Sung), sybang@postech.ac.kr (S. -Y. Bang), seungjin@postech.ac.kr (S. Choi)
URL: http://www.postech.ac.kr/~seungjin (S. Choi)

To appear in Pattern Recognition Letters *12 July 2005*

have been studied and among those, some statistical features based on principle component analysis (PCA) (Joliffe, 1986; Oja, 1988) and independent component analysis (ICA) (Hyvärinen et al., 2001; Choi et al., 2005), and Gabor filters (Gabor, 1946), drew extensive attraction. Apart from those, kernel-based methods (Yang et al., 2004) and wavelet-based multiresolution methods (Zhang et al., 2004; Sastry et al., 2004; Li and Shawe-Taylor, 2005) have been developed to extract features. Popular classifiers include naive Bayesian classifiers, k-nearest neighbor classifiers, artificial neural networks, fuzzy classifiers, and support vector machines (SVMs) (for example see (Duda et al., 2001; Liu et al., 2003) and references therein).

For handwritten numeral recognition, various methods have been developed. These include a neural network with PCA-based features (Zhang et al., 2001; Cao et al., 1997), a self-organizing map with fuzzy rules (Chi et al., 1995), tolerant rough set (Kim and Bang, 2000), and so on. It was shown that Gabor features were somewhat robust to the noise and could model the receptive field characteristics of simple cells in the primary visual cortex (Porat and Zeevi, 1988; Daugman, 1980). Gabor filters were also used in feature extraction from handwritten numerals (Hamamoto et al., 1998; Shustorovich, 1994). So far, most of methods have considered the feature extraction and classification, separately. However, it is desirable to consider both feature extraction and classification simultaneously, in order to extract useful features and to construct a better classifier which incorporates with features.

The extensive survey of recognition performance for large handwritten digit database was reported in (Liu et al., 2003) through many kinds of features and classifiers. However, it excluded the Gabor features and the Bayesian network classifier, which we will explore. In this paper, we present a method of exploiting feature extraction and constructing a classifier simultaneously, in the task of handwritten numeral recognition. To this end, we first introduce hierarchical Gabor features (HGFs) which represent different levels of information which are structured such that the lower the level is, the more localized information they have. At each level, we choose an optimal set of 2-D Gabor filters in the sense that Fisher's linear discriminant (FLD) measure (Duda et al., 2001) is maximized and these Gabor filters are used to extract HGFs. Then, we construct a Bayesian network classifier that encodes hierarchical dependence among HGFs. We confirm the useful behavior of our proposed method, comparing it with the naive Bayesian classifier, k-nearest neighbor, and an artificial neural network, in the task of handwritten numeral recognition.

The rest of this paper is organized as follows. Next section briefly overviews 2-D Gabor filters and presents a method of extracting HGFs. Sec. 3 explains how to construct a hierarchical Bayesian network classifier which encodes dependence in HGFs. Experimental results in handwritten numeral recognition, are shown in Sec. 4. Finally conclusions are drawn in Sec. 5.

2 Hierarchical Gabor Features

Hierarchical features are expected to represent different levels of information where the lower the level is, the more localized the information is. To this end, we first consider 2-D Gabor filters in a hierarchy with several levels. Then for each level we select a set of Gabor filters with an optimal frequency in the sense that FLD measure is maximized. Finally, we extract HGFs from these optimal Gabor filters.

2.1 Gabor Filters

2-D Gabor filters have been widely used in computer vision and image processing, due to its usefulness in representing images in an efficient manner. In general, the 2-D Gabor filter centered at $(0,0)$ in the spatial domain is defined as

$$G(x,y,\xi_x,\xi_y,\sigma_x,\sigma_y,\theta) = \frac{1}{\sqrt{\pi\sigma_x\sigma_y}}\,\mathrm{e}^{-\frac{1}{2}\left[\left(\frac{R_1}{\sigma_x}\right)^2+\left(\frac{R_2}{\sigma_y}\right)^2\right]}\,\mathrm{e}^{i(\xi_x x+\xi_y y)}, \qquad (1)$$

where $R_1 = x\cos\theta + y\sin\theta$ and $R_2 = -x\sin\theta + y\cos\theta$, ξ_x and ξ_y are spatial frequencies, σ_x and σ_y are the standard deviations of an elliptical Gaussian along the x and y axes, and θ denotes an orientation.

Physiological findings revealed that simple and complex cells in primary visual cortex usually have an elliptical Gaussian envelope with an aspect ratio of $1.5 \sim 2.0$ and have the plane wave's propagating direction along the short axis of the elliptical Gaussian envelope (Daugman, 1985; Lee, 1996). These finding suggest the relation

$$\xi_x = \omega\cos\theta \quad \text{and} \quad \xi_y = \omega\sin\theta, \qquad (2)$$

where $\omega = \sqrt{{\xi_x}^2 + {\xi_y}^2}$.

Incorporating these relations into the form of Gabor filter (1), leads to

$$G(x,y,\omega,\sigma,r,\theta) = \frac{1}{\sqrt{\pi r}\,\sigma}\,\mathrm{e}^{-\frac{1}{2}\left[\left(\frac{R_1}{\sigma}\right)^2+\left(\frac{R_2}{r\sigma}\right)^2\right]}\,\mathrm{e}^{i\,\omega R_1}, \qquad (3)$$

where $\sigma = \sigma_x$. This Gabor filter is centered at $(0,0)$ in the spatial domain. In addition, it has the elliptical Gaussian envelope with an aspect ratio of $r = \sigma_y/\sigma_x$ and has the plane wave's propagation direction along the x-axis, which is the short axis of Gaussian envelope. The Gabor filter centered at (x', y'), is simply represented by $G(x' - x, y' - y, \omega, \sigma, r, \theta)$.

Given an input image I, the response, z, of the Gabor filter G centered at (x', y'), is computed as the convolution,

$$z = \sum_{x} \sum_{y} I(x, y) G(x' - x, y' - y, \omega, \sigma, r, \theta) \,, \tag{4}$$

where $I(x, y)$ is the intensity value of the image I at (x, y).

Since the complex-valued Gabor filter in (1) consists of real component (even symmetric function) and imaginary component (odd symmetric function), the response z also consists of real response and imaginary response. We can use the real response or imaginary response of z as a feature. Also, the magnitude of z can be used as a feature. While the imaginary component of the Gabor filter is zero mean, the real component is not zero mean and has the d.c. response. This d.c. response can be sensitive to the size and thickness of numeral in an image, which is irrelevant information for the recognition. Therefore, the real response or the magnitude including d.c. information may lead to poor recognition of handwritten numeral. Through experiments, we found that the imaginary response of Gabor filter gives better recognition performance than the real response or the magnitude on the handwritten numeral database, which we used. Therefore, we employ the only imaginary part of z as a feature.

2.2 Hierarchical Gabor Features

2-D Gabor filters involve several parameters such as centers, frequencies, orientations, and standard deviations. Thus various combinations of these parameters produce a set of Gabor filters, some portion of which might be useful in the task of pattern recognition. An open issue is concerned with a way of selecting an optimal (in some sense) set of Gabor filters which produce features providing the best classification performance. Here we propose a method of constructing a set of Gabor filters in a hierarchical way with Gabor filters at each level being selected such that FLD measure is maximized.

First, we determine the center of a Gabor filter in the spatial domain using the 9-sub-sampling decomposition in a hierarchy with L levels so that some sub-sampling points of neighbor sampling points are shared (see Fig. 1). Those shared sub-sampling points in level l play a role to capture the correlations between Gabor features extracted from neighbor sampling points in the upper level $l-1$ in the hierarchical Bayesian network classifier, which we will present next section. Levels in a hierarchy are numbered from the top to the bottom, that is, the level 1 is the top level and the level L is the bottom level. Starting from single point located at the center of an image in the top level, the sample point is gradually decomposed into 9 sub-sample points from the top to the

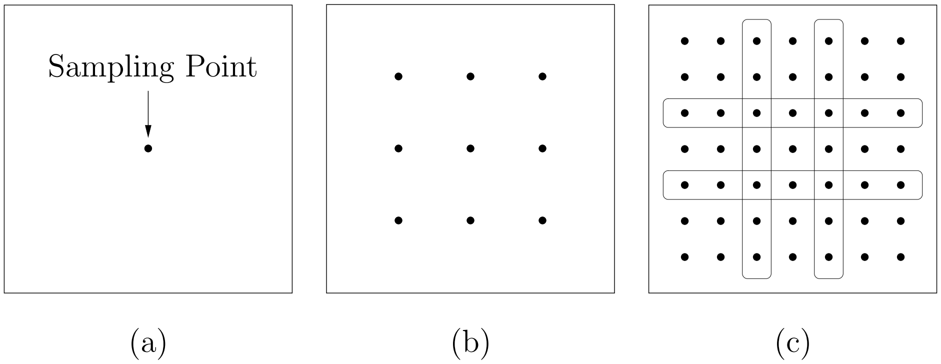

Fig. 1. An illustration of the 9-sub-sampling decomposition is shown. Starting from a single point at the top level (level 1) in (a), the number of sample points gradually increases in order to pick up more localized features. Sample points at level 2 and level 3 are shown in (b) and (c), respectively. In the level 3, the sampling points in rectangles indicate the sub-sampling points shared with adjacent neighbor centers, some of which correspond to sampling points in level 2.

bottom level. For instance, the top level contains a single sample point which is served as a center of the Gabor filter that covers the whole area of an image. The next lower level contains 9 sample points, each of which corresponds to the center of a Gabor filter which covers local area of an image. In this way, the lower the level is, the more localized area is covered by an Gabor filter. Note that after top level the sampling points located in a neighbor share some sub-sample points. In the example of 3 level hierarchy in Fig. 1, we actually have 1 sampling point in level 1, 9 sampling points in level 2, and 49 sampling points in level 3. In general, n-sub-sampling decomposition is possible, however, we employ the 9-sub-sampling decomposition in this paper.

Once sample points at each level are determined, a set of Gabor filters is defined at each sample point corresponding to the center of a Gabor filter in the spatial domain. Denote by $p^{ls} = (x^{ls}, y^{ls})$ the sth sample point in the level l for $l = 1, \dots, L$ and $s = 1, \dots, N_l$, where N_l is the number of sample points in the level l. For example, $N_1 = 1$, $N_2 = 9$, and $N_3 = 49$.

Let $\Omega = \{\omega_1, \dots, \omega_K\}$ be a set of K frequencies and $\Theta = \{\theta_1, \dots, \theta_D\}$ be a set of D orientations. Given a sample point p^{ls} and an orientation $\theta_j \in \Theta$, we define a set of Gabor filters, which is located at (x^{ls}, y^{ls}) in the spatial domain and have K-frequencies in Ω, such as

$$G_j^{ls} = \left\{ G_{j1}^{ls}, \dots, G_{jK}^{ls} \right\} , \tag{5}$$

where $G_{ji}^{ls} = G_{ji}^{ls}(x, y) = G\left(x^{ls} - x, y^{ls} - y, \omega_i, \sigma^{ls}, r, \theta_j\right)$ defined in (3).

To extract the hierarchical information with these Gabor filter sets, we determine the standard deviation in such a way that the average of distances

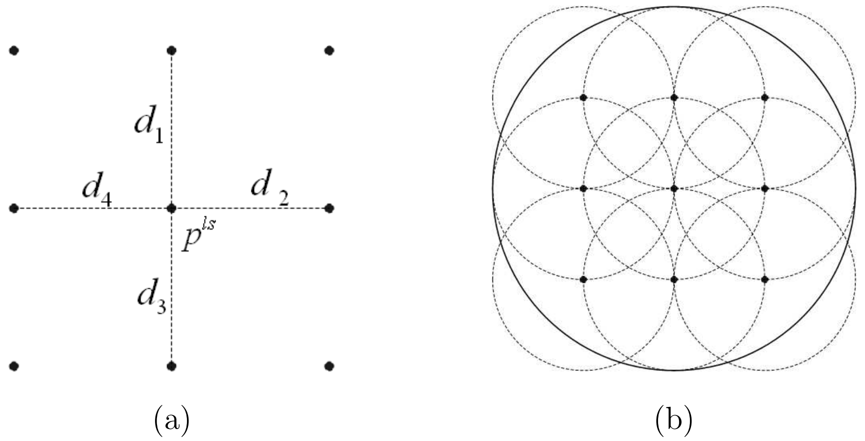

Fig. 2. A pictorial illustration for regions covered by Gabor filters with standard deviations determined by (7): (a) Distances between the point in the center, p^{ls}, and its 4 neighbor points, are denoted by $d_1, \ldots, d_4$, the average of which is d^{ls}; (b) Circles represent the regions of Gabor filters. A big circle whose center is located at p^{ls} is an area covered by a Gabor filter in the upper level and 9 smaller circles represent more localized areas covered by Gabor filters whose centers are located at 9 sub-sample points in the current level.

between p^{ls} and its four neighbor sample points, denoted by d^{ls}, satisfies the following relation:

$$\exp\left\{-\frac{1}{2}\left(\frac{d^{ls}}{\sigma^{ls}}\right)^2\right\} = \frac{1}{2}, \tag{6}$$

which leads to

$$\sigma^{ls} = \frac{d^{ls}}{\sqrt{2\log 2}}. \tag{7}$$

Figs. 2 and 3 illustrate the regions covered by Gabor filters with standard deviations determined by (7). The Gabor filter in the top level take care of a global region, then Gabor filters extract more localized information in the lower level since σ^{ls} decreases gradually as moving from the top to the bottom level. The lower the level is , the more detailed the information can be extracted by these Gabor filter sets. Gabor filters in the same level take care of regions that are sufficiently overlapped in the spatial domain, hence the loss of information is insignificant.

Remaining parameters controlling the shape of a Gabor filter are frequencies which are crucial in the design of a Gabor filter and have high influence on the recognition performance Hamamoto et al. (1998). Gabor filters with all K frequencies may result in irrelevant features which could decrease the efficiency of a classifier and do not necessarily improve the recognition performance. Therefore, the selection of optimal Gabor filters with an efficient frequency is required to extract the relevant features for recognition. We can directly select optimal Gabor filters among candidates based on the recognition rate. However, a few drawbacks exist with such a method. The first is that the num-

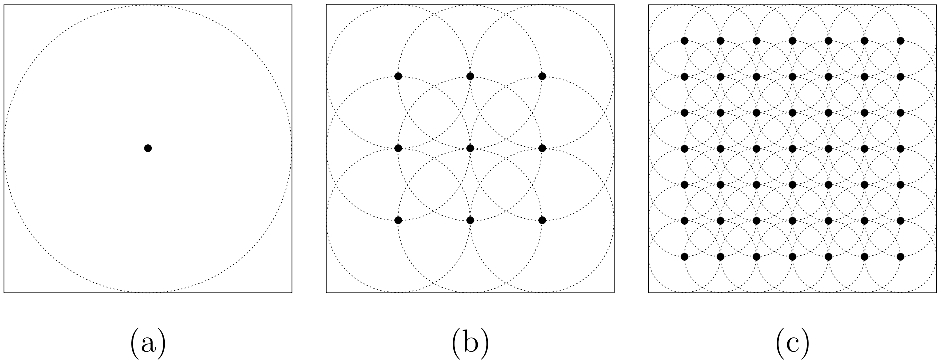

Fig. 3. Regions covered by Gabor filters in: (a) level 1 (top level); (b) level 2; (c) level 3 (bottom level). Areas taken care of by Gabor filters are gradually smaller as moving from top level to bottom level, in order to extract more localized information in the lower level.

ber of candidates increase exponentially with the number of Gabor filter sets, because the recognition rate depends on all Gabor filters from the top to the bottom level. The second drawback is that the selected Gabor filters depend on the used classifier. Therefore, we introduce another method that selects an optimal set of Gabor filters with a single frequency, such that FLD measure is maximized ; the FLD measure is a measure to how certain information is efficient for discrimination. This selection is carried out for every Gabor filter set in (5), independently.

Suppose that a set of n labelled pattern images, $\mathbf{I}$, consists of $I_1, \ldots, I_n$ and n_p labelled pattern images of subset $\mathbf{I}_p$ ($\subset \mathbf{I}$) are labelled c_p, where $p = 0, 1, \ldots, c-1$ for c-class problem ; the handwritten numeral recognition is a 10-class problem. Let $\mathbf{z}_{ji}^{ls}$ be a set of responses of G_{ji}^{ls} ($\in G_j^{ls}$) with $\mathbf{I}$, which is defined as follows

$$\mathbf{z}_{ji}^{ls} = \{\, z \,:\, z = \sum_x \sum_y I_k(x,y)\, G_{ji}^{ls}(x,y)\,,\; k = 1, \ldots, n\}. \tag{8}$$

Let $\mathbf{z}_{jip}^{ls}$ denote a subset of $\mathbf{z}_{ji}^{ls}$ associated with $\mathbf{I}_p$, then the within-class scatter is defined as

$$S_W(\,\mathbf{z}_{ji}^{ls}\,) = \sum_{p=0}^{c-1} \sum_{z \in \mathbf{z}_{jip}^{ls}} (z - m_p)^2\,, \tag{9}$$

where $m_p(= \frac{1}{n_p} \sum_{z \,\in\, \mathbf{z}_{jip}^{ls}} z)$ denotes a class mean. The between-class scatter is defined as

$$S_B(\,\mathbf{z}_{ji}^{ls}\,) = \sum_{p=0}^{c-1} n_p (m_p - m)^2\,, \tag{10}$$

where $m(= \frac{1}{n} \sum_{z \,\in\, \mathbf{z}_{ji}^{ls}} z)$ denotes the total mean. The FLD measure is defined by the ratio of between-class scatter and within-class scatter and for the set of responses $\mathbf{z}_{ji}^{ls}$, it becomes

$$f_{ji}^{ls} = f(\mathbf{z}_{ji}^{ls}) = \frac{S_B(\,\mathbf{z}_{ji}^{ls}\,)}{S_W(\,\mathbf{z}_{ji}^{ls}\,)}\,. \tag{11}$$

Given a orientation θ_j and a center p^{ls}, we search for an optimal frequency (among K possible frequencies), which produces the largest FLD measure. In other words, an optimal Gabor filter G_{j*}^{ls} ($\in G_j^{ls}$) has the frequency ω_{i^*} ($\in \Omega$), which is determined such as

$$i^* = \arg\max_i f_{ji}^{ls}\,. \tag{12}$$

Afterwards, we refer to a Gabor feature of a pattern image, I, as the response with an optimal Gabor filter, i.e.,

$$a_j^{ls} = \sum_x \sum_y I(x,y)\, G_{j*}^{ls}(x,y)\,. \tag{13}$$

For a sample point p^{ls}, we obtain a D-dimensional Gabor feature vector

$$\mathbf{a}^{ls} = [\,a_1^{ls}, \ldots, a_D^{ls}\,]^T\,, \tag{14}$$

where D is the number of orientations and T denotes the transpose. Finally, we define the hierarchial Gabor features (HGFs) of a pattern image as the collection of all Gabor feature vectors from the top to the bottom level and $\mathbf{a} = \{\mathbf{a}^{ls}\}$ denotes the HGFs.

3 A Hierarchical Bayesian Network Classifier

3.1 Bayesian network

A Bayesian network consists of a set of variables, $\mathbf{V} = \{A_1, \ldots, A_N\}$, and a set of directed edge, $\mathbf{E}$, between variables, which form a directed acyclic graph (DAG), $G = (\mathbf{V}, \mathbf{E})$, where a joint distribution of variables is represented by the product of conditional distributions of each variable given its parents (Pearl, 1988; Jordan, 1998). Each node, $A_i \in \mathbf{V}$, represents a random variable and a directed edge from A_i to A_j, $(A_i, A_j) \in \mathbf{E}$, represents the conditional dependency between A_i and A_j. In a Bayesian network, each variable is independent of its non-descendants, given a value of its parents in G. This independence encoded in G reduces the number of parameters which is required to characterize a joint distribution, so that posterior distribution given evidence can be efficiently done.

In a Bayesian network over $\mathbf{V} = \{A_1, \ldots, A_N\}$, the joint distribution $P(\mathbf{V})$ is the product of all conditional distributions specified in the Bayesian network, i.e.,

$$P(A_1, \ldots, A_N) = \prod_{i=1}^{N} P(A_i \,|\, \mathbf{Pa}_i)\,, \tag{15}$$

where $P(A_i \,|\, \mathbf{Pa}_i)$ is the conditional distribution of A_i, given $\mathbf{Pa}_i$ which denotes the parent set of A_i. A conditional distribution for each variable has a parametric form that can be learned by the maximum likelihood estimation. Please refer to (Pearl, 1988) for more details on Bayesian networks.

3.2 Bayesian Network Classifier

The Bayesian network classifier (Friedman et al., 1997) is a Bayesian network that distinguishes a class node from the nodes of feature variables (feature nodes) The Bayesian network classifier requires the class node to be a parent of all feature nodes. This ensures that all feature nodes are taken into account in the learned Bayesian network when one computes the posterior probability of the class node, which is a main term determining the classification. For example, the simplest Bayesian network classifier is the naive Bayesian classifier (NBC) which assumes the statistical independence between feature variables. Fig. 4 shows the pictorial description for NBC where C denotes the class node and A_1,. . . ,A_N denote the feature nodes.

In NBC, each feature variable is independent of the remainder of the feature variables, given a value of the class variable. It follows from (15) that the joint distribution determined by NBC is decomposed as

$$P(C, A_1, \ldots, A_N) = \prod_{i=1}^{N} P(A_i|C)P(C)\,. \tag{16}$$

Denote by $\mathbf{a} = \{a_1, \ldots, a_N\}$ the instances of feature variables. For classification, one computes the posterior distribution over the class variable, which is of the form

$$P(C|A_1 = a_1, \ldots, A_N = a_N) = \frac{\prod_{i=1}^{N} P(A_i = a_i|C)P(C)}{\sum_{C'} \prod_{i=1}^{N} P(A_i = a_i|C')P(C')}\,. \tag{17}$$

Then, $\mathbf{a}$ is assigned to a class through MAP

$$c^* = \arg\max_{C} P(C|A_1 = a_1, \ldots, A_N = a_N)\,. \tag{18}$$

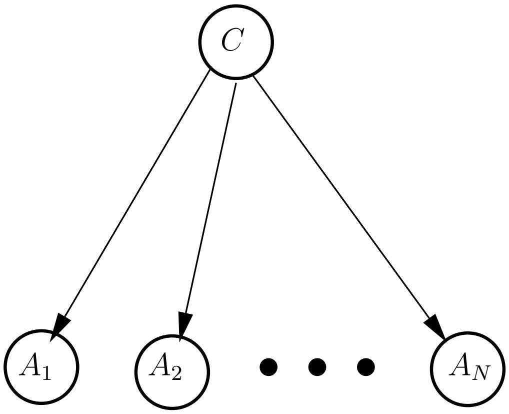

Fig. 4. The naive Bayesian classifier consists of a class node C and a set of feature nodes, $\{A_1, \ldots, A_N\}$, assuming that feature nodes are statistically independent, given C.

3.3 Hierarchical Bayesian Network Classifier

A critical limitation of NBC results from the assumption that feature nodes are statistically independent given a value of the class node, which is unrealistic in practical applications. In order to overcome this limitation, Friedman *et al.* (Friedman et al., 1997) extended NBC, incorporating with a tree structure, to improve the classification performance. In this paper we construct a hierarchical Bayesian network classifier (HBNC) which encodes the dependence between nodes in different levels, induced by HGFs.

Let A^{ls} be a node associated with the sample point p^{ls}, the value of which is assigned by Gabor feature vector $\mathbf{a}^{ls}$ in (14). We first define the structure of HBNC with only feature nodes excluding the class node. For $\mathbf{V} = \{A^{ls}\}$, suppose that the DAG $G = (\mathbf{V}, \mathbf{E})$ defines the structure of HBNC with excluding the class node, i.e., $C \notin \mathbf{V}$. Let $\mathbf{A}^l = \{A^{l1}, \ldots, A^{lN_l}\}$ denote a set of feature nodes in the level l. We also define $\mathbf{B}^{ls}$ as a set of nodes which are associated with 9 sub-sample points of A^{ls}, i.e.,

$$\mathbf{B}^{ls} = \{B_1^{ls}, \ldots, B_9^{ls}\}, \tag{19}$$

where $B_i^{ls} \in \mathbf{B}^{ls}$ is a node associated with the sub-sample point of p^{ls} and $\mathbf{B}^{ls} \subset \mathbf{A}^{l+1}$. Then, the structure of HBNC is composed of the following substructure

$$G^{ls} = \Big(\{A^{ls}\} \cup \mathbf{B}^{ls}, \mathbf{E}^{ls} = \{(A^{ls}, B_1^{ls}), \ldots, (A^{ls}, B_9^{ls})\} \Big), \tag{20}$$

for all l, s with $l = 1, \ldots, L-1$. Therefore, a set of directed edges in G is given by

$$\mathbf{E} = \bigcup_{l,s} \mathbf{E}^{ls}, \tag{21}$$

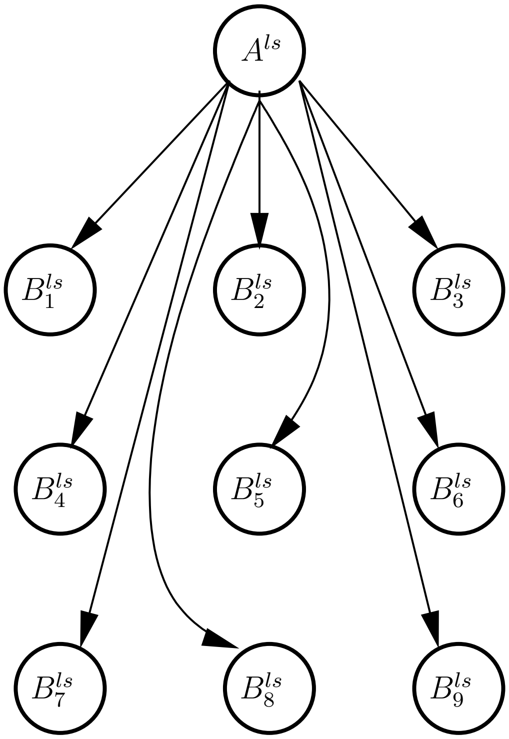

Fig. 5. The sub-structure of HBNC is shown. The nodes of $B_1^{ls}, \ldots, B_9^{ls}$, which are associated with 9 sub-sample points of p^{ls}, are influenced by its parent node A^{ls} that is associated with the sample point p^{ls}.

for all l, s with $l = 1, \ldots, L-1$. Fig. 5 and Fig. 6 show the sub-structure, G^{ls}, and the pictorial description of the HBNC excluding class node, G.

Note that A^{ls} directly influences only nodes in $\mathbf{B}^{ls}$ which represent more localized information than A^{ls}. See Fig. 5 for this local sub-structure of our HBNC. From the top to the bottom level, this limited direct dependence entirely encodes the hierarchical dependence between nodes in levels, implied by HGFs (see Fig. 6). In addition, nodes associated with the shared sub-sampling points have more parents than one so that they capture the correlations between their parents associated with neighbor sampling points located in the upper level.

Now we include the class node as a parent of all feature nodes in G , in order to complete the overall structure of our HBNC. We denote the overall DAG G_c (including the class node) by $G_c = (\mathbf{V}_c, \mathbf{E}_c)$ where $\mathbf{V}_c = \mathbf{V} \cup \{C\}$ and $\mathbf{E}_c = \mathbf{E} \cup \{(C, A^{ls})\}$ for all $A^{ls} \in \mathbf{V}$. Denote by $\mathbf{Pa}^{ls}$ a set of parent nodes of A^{ls} in G, then the set of parent nodes of A^{ls} in G_c becomes $\mathbf{Pa}^{ls} \cup \{C\}$.

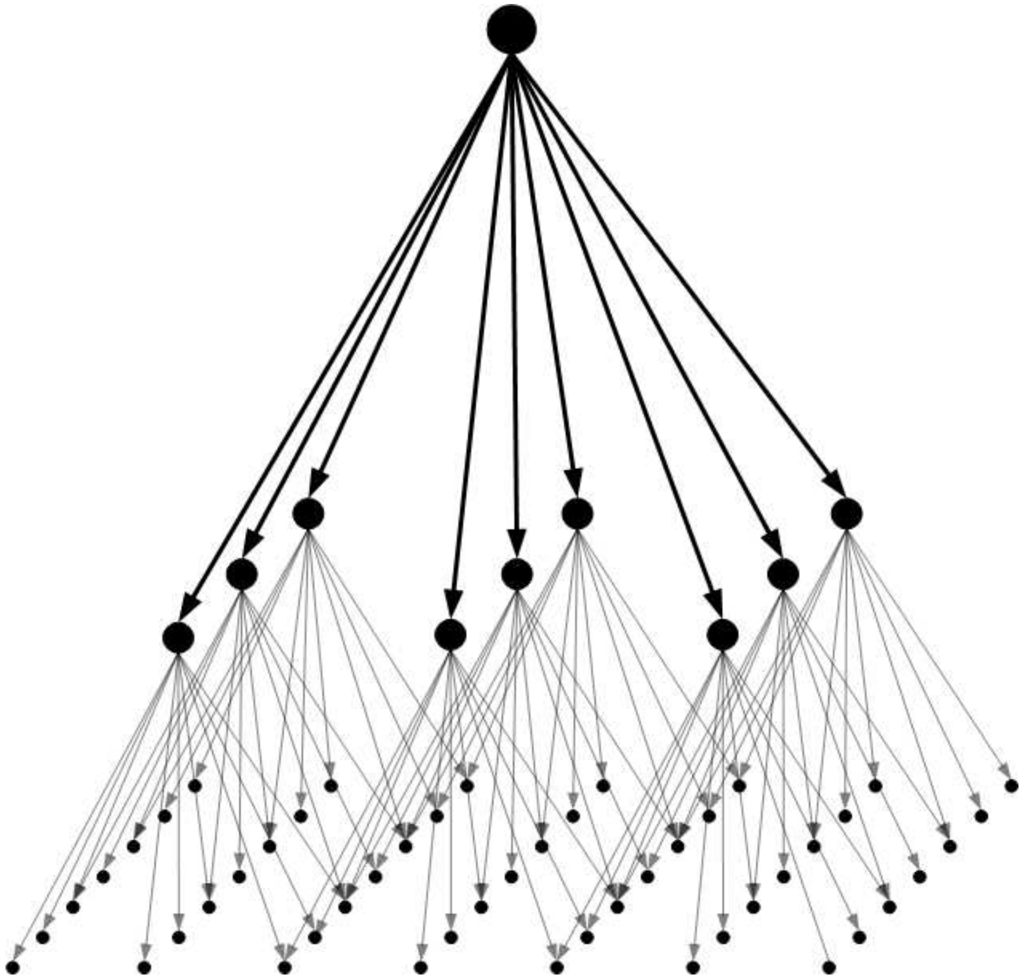

Fig. 6. The structure of our HBNC (excluding the class node C) is shown. Starting from a single point in the top level, the number of nodes are gradually increasing, following the 9-sub-sampling decomposition. Nodes associated with the shared sub-sampling points have more parents than one so that they capture correlations between their parents associated with neighbor sampling points located in the upper level.

Therefore, the joint distribution of all nodes in HBNC is decomposed as

$$P(C, \mathbf{V} = \{A^{ls}\}) = \prod_{l,s} P(A^{ls} \,|\, \mathbf{Pa}^{ls}, C) P(C)\,. \tag{22}$$

It follows from (22) that the complete definition of HBNC requires the conditional distributions to be specified. Apparently, C is a discrete random variable and A^{ls} are continuous multivariate random variables. We use the multinomial distribution for $P(C)$ and the conditional multivariate Gaussian density (Lauritzen and Wermuth, 1989) for $P(A^{ls} \,|\, \mathbf{Pa}^{ls}, C)$. These conditional distributions can be learned by maximum likelihood estimation, after HGFs are obtained from a set of labelled images.

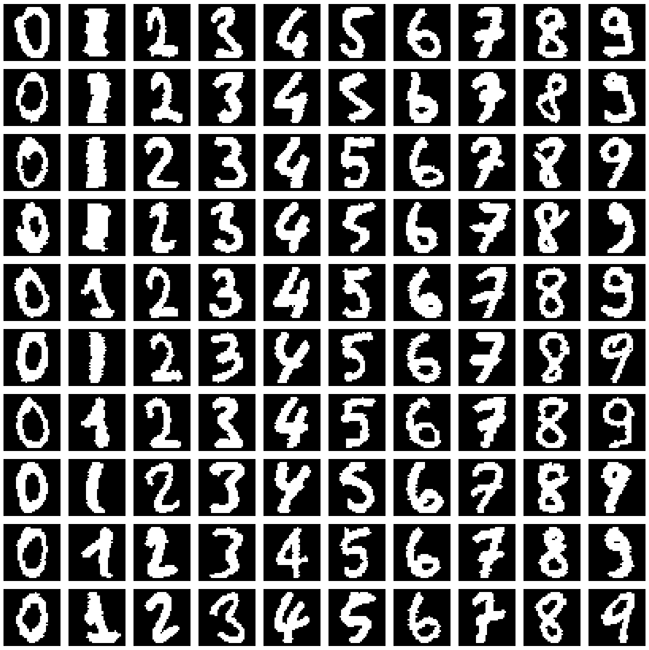

Fig. 7. 100 handwritten numeral images of UCI database.

Given HGFs $\mathbf{a} = \{\mathbf{a}^{ls}\}$ of a pattern image, the classification requires the computation of the posterior distribution of C given $\mathbf{a}$, which is of the form

$$P(C|\mathbf{V} = \mathbf{a}) = \frac{\prod_{ls} P(A^{ls} = \mathbf{a}^{ls}|\mathbf{Pa}^{ls} = \mathbf{pa}^{ls}, C)P(C)}{\sum_{C'} \prod_{ls} P(A^{ls} = \mathbf{a}^{ls}|\mathbf{Pa}^{ls} = \mathbf{pa}^{ls}, C')P(C')} . \tag{23}$$

In this paper, we compute this posterior distribution by the junction tree algorithm (Lauritzen and Spiegalhalter, 1988; Cowell et al., 1999) which is a well-known exact inference method for Bayesian networks. Finally, the classification is achieved by assigning the pattern to class label which produces maximum posterior probability.

4 Experiments

We used the 32 × 32 binary handwritten numeral image data from the UCI database (Blake and Merz, 1998). These numeral images were centered and normalized. We did not performed any other preprocessing, such as slant correction and smoothing. A few researchers used these data in their experiments (Kim and Bang, 2000). Fig. 7 shows some UCI numeral images, which we used. We randomly chose 25, 50, 300 images per class for training and 1,943 images for testing. Table 1 shows the configuration of these testing image data. Training and testing images did not overlap. For comparison, we tested our methods with the standard Gabor features, which did not represent the hierarchical information, and other well-known classifiers, such as the naive

Table 1
The number of testing numeral image data per class.

Class	0	1	2	3	4	5	6	7	8	9	Total
#	189	198	195	199	186	187	195	201	180	204	1934

Bayesian classifier (NBC), k-nearest neighbor classifier (KNN), and artificial neural network (NN).

4.1 Implementation

We used the imaginary part of the response in (4) as a Gabor feature and employed the fixed four orientations $\Theta = \{\ 0,\ \pi/4,\ \pi/2,\ 3\pi/4\ \}$. Based on the physiological findings mentioned in Sec. 2.1, we fixed the aspect ratio r to 2 and level size L to 3, such as that of Fig. 1. For each sampling point and orientation, we selected single optimal frequency among 10 frequencies $\Omega = \{$ 0.025, 0.05, 0.075, 0.1, 0.15, 0.2, 0.4, 0.6, 0.8, 1 $\}$ based on randomly chosen 300 training images per class by FLD measure in Sec 2.2. Therefore, we had 236 (=59 nodes × 4 orientations × 1 frequency) Gabor filters to extract HFGs, which consisted of 236 Gabor features. Fig. 8 shows some of those optimal Gabor filters with $\theta_j = \pi/4$.

All classifiers tested with same training and testing numeral images for the comparison. (1) For the proposed HBNC and the NBC, we learned the parameters of conditional distributions by maximum likelihood method with HGFs of training numeral images. Note that the NBC had exactly the same feature nodes as HBNC, but did not encode the dependencies among the HGFs. To classify HGFs, we inferred the posterior distribution of the class variable by the junction tree algorithm. (2) To learn the parameters of NN, we fixed the number of hidden units to 150, the learning rate to 0.01, the momentum rate to 0.5, and the number of learning iteration to 10,000. (3) KNN did not require parameter learning. In our experiments, when the number of the nearest neighbors, k, was one, KNN best performed among k=1, 3, 5. Therefore, we reported the result with only k=1.

4.2 Experimental Results

We performed experiments focused on two aspect. The first was to show the effectiveness of HGFs and the second was to verify that the HBNC improves

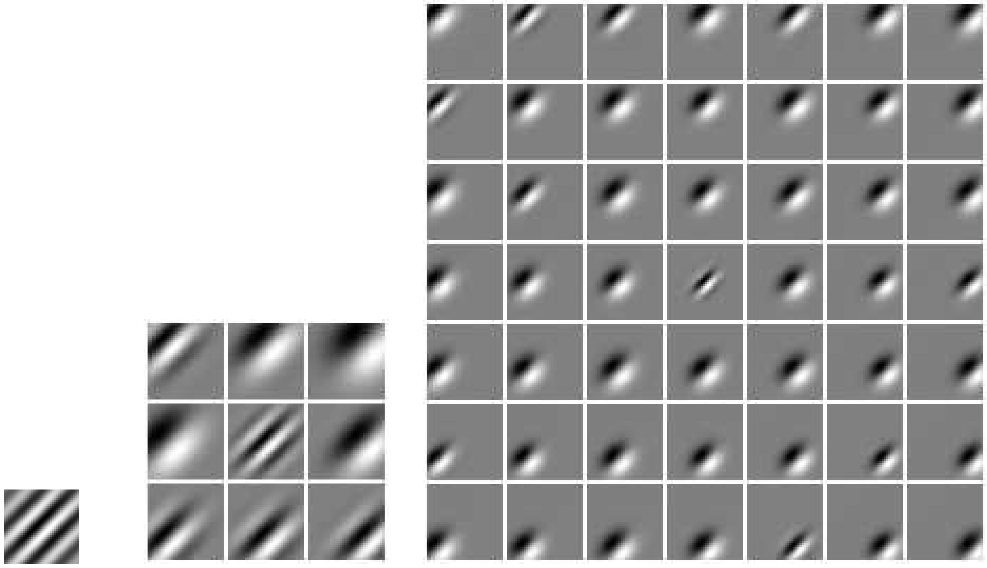
(a) level 1 (b) level 2 (c) level 3

Fig. 8. Optimal Gabor filters, which were selected using FLD measure from 300 labelled numeral images per class, are shown with orientation $\theta_j = \pi/4$ and aspect ratio $r = 2$. From the top to the bottom level, each optimal Gabor filter is centered at sample points (see Fig. 1) and has an optimal frequency in which FLD measure maximizes (see Sec. 2.2).

the classification performance of HGFs. For the first, We selected the Gabor features with only the bottom level in order to remove the hierarchical information. We refer to these Gabor features as the standard Gabor features (SGFs). The SGFs are similar features introduced in (Hamamoto et al., 1998). For the second, the HGFs of 1934 testing numeral image data were classified by the proposed HBNC and other well-known classifiers such as NBC, KNN and NN.

Fig. 9 and Table 2 show the average recognition performance from 30 trials of differently chosen training and testing numeral images. First, the recognition performance of the proposed HGFs outperformed that of SGFs within the same classification method. Moreover, the proposed HGFs significantly improved the recognition performance, when the number of training data were small. However, the improvement was less significant with NBC. The NBC showed the similar recognition performance on both HGFs and SGFs, because it did not encode the hierarchical dependencies imposed by HGFs. Second, the proposed HBNC, which overcomes the limitation of NBC, best classified HGFs among other classifiers in experiments. However, when the number of training numeral images per class was 25, the NBC outperformed the HBNC. This shows that 25 training numeral image data were insufficient to learn our HBNC, because conditional distributions of the HBNC had more parameters

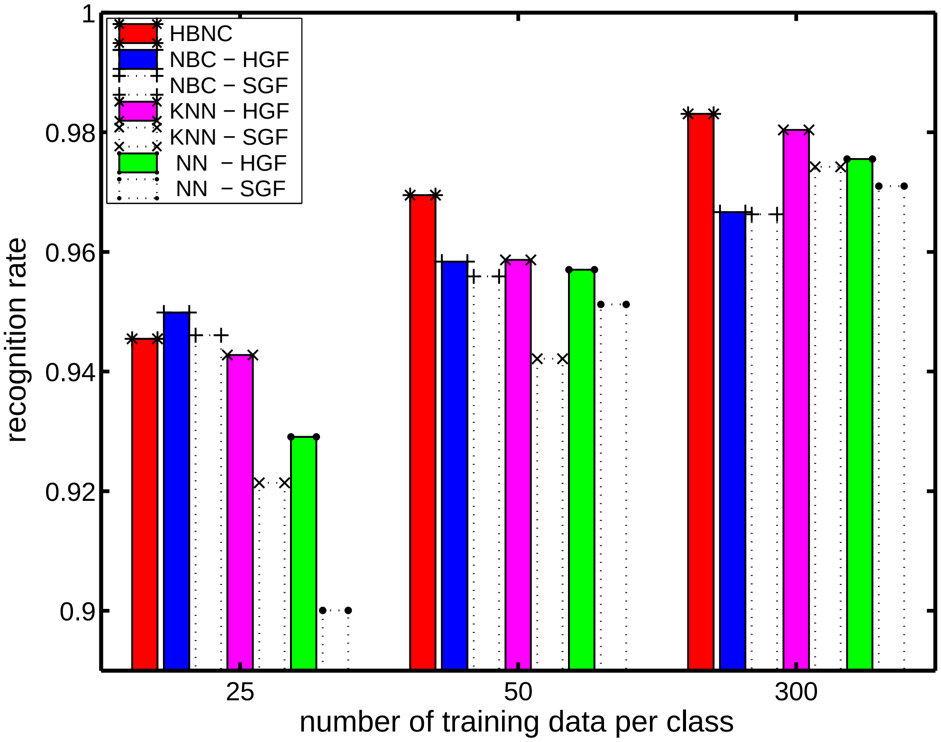

Fig. 9. The average recognition performance of 1934 testing numeral image data from 30 trials when the number of training numeral image data per class is 25, 50, and 300. HBNC : hierarchical Bayesian network classifier, NBC : naive Bayesian classifier, KNN : k-nearest neighbor classifier with k=1, NN : artificial neural network. HGF denotes the hierarchical Gabor features and SGF denotes the standard Gabor features.

Table 2
The average recognition performance of 1934 testing numeral image data from 30 trials when the number of training numeral image data per class is 300.

HBNC	NBC-HGF	NBC-SGF	KNN-HGF	KNN-SGF	NN-HGF	NN-SGF
0.983 ± 0.003	0.967 ± 0.004	0.966 ± 0.004	0.980 ± 0.003	0.974 ± 0.004	0.976 ± 0.003	0.971 ± 0.006

than NBC. Although KNN showed a remarkable performance in the case of 300 training numeral image data per class, it required larger computation time than other classifiers. This burden of computation with KNN generally comes from comparing each testing instance with every training instance. Therefore, we suggest that KNN should be considered only to compare it with other methods, but not in practice. Finally, Table 3 shows the average confusion tables of classification result on HGFs of 1934 testing numeral data, which are classified by HBNC and NBC, when the number of training numeral data per class is 300.

Table 3
Average confusion tables of classification result on HGFs of 1934 testing numeral image data by the proposed HBNC and the NBC from 30 trials, when the number of training numeral image data per class is 300.

Class	0	1	2	3	4	5	6	7	8	9	Rate	Total
0	187.5	0.0	0.0	0.0	0.2	0.9	0.2	0.0	0.1	0.1	0.99	189
1	0.0	194.0	0.4	0.3	0.1	0.6	0.0	0.5	1.5	1.0	0.98	198
2	0.0	0.2	192.8	0.1	0.4	0.0	0.0	0.0	1.1	0.5	0.99	195
3	0.0	0.0	0.1	196.1	0.0	1.0	0.0	0.2	0.7	1.0	0.99	199
4	0.1	0.0	0.0	0.0	181.8	0.3	0.7	0.4	1.1	1.7	0.98	186
5	0.0	0.0	0.0	0.7	0.3	183.7	0.0	0.2	0.5	1.8	0.98	187
6	0.4	0.9	0.0	0.0	0.6	0.1	191.7	0.0	1.5	0.0	0.98	195
7	0.0	0.0	0.0	0.6	0.0	0.0	0.0	199.3	0.5	0.7	0.99	201
8	0.0	1.7	0.0	0.2	0.0	0.0	0.0	0.0	177.7	0.4	0.99	180
9	0.0	0.0	0.0	3.7	1.3	0.4	0.0	0.5	1.5	196.7	0.96	204
	187.9	196.7	193.2	201.5	184.6	186.9	192.5	201.0	186.1	203.8	0.983	1934

(a) Proposed HBNC with HGFs

Class	0	1	2	3	4	5	6	7	8	9	Rate	Total
0	186.8	0.0	0.0	0.0	0.7	0.4	0.0	0.0	0.6	0.6	0.99	189
1	0.0	189.1	2.1	0.3	0.1	0.3	0.1	0.8	2.5	2.9	0.96	198
2	0.0	1.2	190.7	0.0	0.0	0.0	0.0	1.0	1.6	0.6	0.98	195
3	0.0	0.0	0.4	191.5	0.0	2.1	0.0	1.8	2.5	0.8	0.96	199
4	0.0	0.3	0.0	0.0	178.5	0.1	1.0	1.0	3.1	2.2	0.96	186
5	0.0	0.0	0.0	0.4	0.4	178.2	0.0	0.0	0.4	7.7	0.95	187
6	0.1	1.1	0.0	0.0	1.1	0.2	191.6	0.0	1.0	0.0	0.98	195
7	0.0	0.5	0.0	0.5	0.4	0.0	0.0	198.5	0.5	0.7	0.99	201
8	0.0	3.8	0.0	0.3	0.2	0.0	0.0	0.3	172.7	2.9	0.96	180
9	0.0	1.6	0.0	2.4	2.8	0.7	0.0	2.3	2.3	192.1	0.94	204
	186.9	197.5	193.1	195.3	183.9	181.8	192.7	205.4	187.1	210.5	0.967	1934

(b) NBC with HGFs

5 Conclusions

In this paper, we simultaneously considered feature extraction and classification within the hierarchical property for handwritten numeral recognition. For the feature extraction, we introduced Gabor filters to extract different levels of information. As a result, we obtained the hierarchical Gabor features. For the classification, we constructed the hierarchical Bayesian network classifier to encode dependencies implied by HGFs and to improve the classification performance.

For the handwritten numeral images of UCI database, we successfully demonstrated the useful behaviors of our proposed HGFs and HBNC. We showed through experiments that the HGFs are an effective representation of a pattern for the recognition. Based on this result, we suggest that the proposed HGFs should be considered, when Gabor filters are applied to extract features for the recognition. Also, we empirically verified that our HBNC better classify the HGFs in comparison with other well known classifiers. Therefore, we conclude that a Bayesian network classifier exploiting dependencies implied by features can improve the recognition performance, such as our HBNC.

Although we applied the proposed methods for the handwritten numeral data, we believe that the performance of our methods will be promising for other recognition problems.

Acknowledgments

This work was supported by ITEP Brain Neuroinformatics program, National Core Research Center for Systems Bio-Dynamics, and BK 21.

References

Blake, C. L., Merz, C. J., 1998. UCI repository of machine learning databases. Tech. rep., Dept. of Information and Computer Science, University of California, Irvine.

Cao, J., Ahmadi, M., Shridhar, M., 1997. A hierarchical neural network architecture for handwritten numeral recognition. Pattern Recognition 30, 289–294.

Chi, Z., Wu, J., Yan, H., 1995. Handwritten numeral recognition using self-organizing maps and fuzzy rules. Pattern Recognition 28, 59–66.

Choi, S., Cichocki, A., Park, H. M., Lee, S. Y., 2005. Blind source separation and independent component analysis: A review. Neural Information Processing - Letters and Review 6 (1), 1–57.

Cowell, R. G., Dawid, A. P., Lauritzen, S. L., Spiegelhalter, D. J., 1999. Probalilistic Networks and Expert Systems. Springer, New York.

Daugman, J. G., 1980. Two-dimensional spectral analysis of cortical receptive field profiles. Vision Researh 20, 847–856.

Daugman, J. G., 1985. Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters. J. Optical Soc. Amer. 2 (7), 1160–1169.

Duda, R. O., Hart, P., Stork, D. G., 2001. Pattern Classification (2nd edition). John Wiley and Sons, Inc.

Friedman, N., Geiger, D., Goldszmidt, M., 1997. Bayesian network classifier. Machine Learning 29, 131–163.

Gabor, D., 1946. Theory of communication. J. Inst. Electr. Engng. 93, 429–459.

Hamamoto, Y., Uchimura, S., Watanabe, M., Yasuda, T., Mitani, Y., Tomita, S., 1998. A Gabor filter-based method for recognizing handwritten numerals. Pattern Recognition 31 (4), 395–400.

Hyvärinen, A., Karhunen, J., Oja, E., 2001. Independent Component Analysis. John Wiley & Sons, Inc.

Joliffe, I. T., 1986. Pricipal Component Analysis. Springer, New York.

Jordan, M. I. (Ed.), 1998. Learning in Graphical Models. Kluwer Academic Publishers, Dordrecht, The Netherlands.

Kim, D., Bang, S. Y., 2000. A handwritten numeral character classification using tolerant rough set. IEEE Trans. Pattern Analysis and Machine Intelligence 22 (9), 923–937.

Lauritzen, S. L., Spiegalhalter, D. J., 1988. Local computations with probabilities on graphical structures and their application to expert systems (with discussion). Journal of the Royal Statistical Society, Series B 50, 157–224.

Lauritzen, S. L., Wermuth, N., 1989. Graphical models for associations between variables, some of which are qualitative and some quantitative. Annals of Statistics 17, 31–57.

Lee, T. S., Oct 1996. Image representation using 2D Gabor wavelets. IEEE Trans. Pattern Analysis and Machine Intelligence 18 (10), 959–971.

Li, S., Shawe-Taylor, J., 2005. Comparison and fusion of multiresolution features for texture classification. Pattern Recognition Letters 26, 633–638.

Liu, C.-L., Nakashima, K., Sako, H., Fujisawa, H., 2003. Handwritten digit recognition: benchmarking of state-of-the-art techniques. Pattern Recognition 36, 2271–2285.

Oja, E., 1988. Neural networks, pricipal components, and subspaces. Int. J. Neural Systems 1, 61–68.

Pearl, J., 1988. Probalilistic Reasoning in Intelligent Systems. Morgan Kaufmann, San Mateo, California.

Porat, M., Zeevi, Y. Y., Jul 1988. The generalized Gabor scheme of image representation in biological and machine vision. IEEE Trans. Pattern Analysis and Machine Intelligence 10 (4), 452–468.

Sastry, C. S., Pujari, A. K., Deekshatulu, B. L., Bhagvati, C., 2004. A wavelet based multiresolution algorithm for rotation invariant feature extraction. Pattern Recognition Letters 25, 1845–1855.

Shustorovich, A., 1994. A subspace projection approach to feature extraction: the two-dimensional Gabor transform for chararcher recognition. Neural Networks 7 (8), 1295–1301.

Yang, J., Jin, Z., Yang, J., Zhang, D., Frangi, A. F., 2004. Essence of kernel Fisher discriminant: KPCA plus LDA. Pattern Recognition 37, 2097–2100.

Zhang, B., Fu, M., Yan, H., 2001. A nonlinear neural network model of mixture of local principal component analysis: application to handwritten digits recognition. Pattern Recognition 34, 203–214.

Zhang, P., Bui, T. D., Suen, C. Y., 2004. Extraction of hybrid complex wavelet features for the verification of handwritten numerals. In: Proc. Int'l Workshop on Frontiers in Handwriting Recognition. Tokyo, Japan, pp. 347–350.