Hierarchical Bayesian Network for Handwritten Digit Recognition

JaeMo Sung and Sung-Yang Bang

Department of Computer Science and Engineering,
Pohang University of Science and Technology,
San 31, Hyoja-Dong, Nam-Gu, Pohang, 790-784, Korea
{emtidi, sybang}@postech.ac.kr

Abstract. This paper introduces a hierarchical Gabor features(HGFs) and hierarchical bayesian network(HBN) for handwritten digit recognition. The HGFs represent a different level of information which is structured such that the higher the level, the more global information they represent, and the lower the level, the more localized information they represent. The HGFs are extracted by the Gabor filters selected using a discriminant measure. The HBN is a statistical model to represent a joint probability which encodes hierarchical dependencies among the HGFs. We simulated our method about a handwritten digit data set for recognition and compared it with the naive bayesian classifier, the backpropagation neural network and the k-nearest neighbor classifier. The efficiency of our proposed method was shown in that our method outperformed all other methods in the experiments.

1 Introduction

We believe that human beings exploit structured information rather than non-structured information and use the relations among the structured information by some mechanism for recognition. We assume that this structured information is hierarchical and that the relations are limited by hierarchical dependencies. With above assumption, we propose a hierarchical Gabor features(HGFs) and hierarchical bayesian network(HBN) for a recognition mechanism.

2 Hierarchical Gabor Features Extraction

2.1 Gabor filter

The Gabor filter which is represented in the spatial-frequency domain is defined as

$$G(x, y, \omega_0, \sigma, r, \theta) = \frac{1}{\sqrt{\pi} r \sigma} \, e^{-\frac{1}{2}\left[\frac{(rR_1)^2 + {R_2}^2}{(r\sigma)^2}\right]} e^{i\omega_0 R_1} \,, \qquad (1)$$

where $R_1 = x \cos\theta + y \sin\theta$, $R_2 = -x \sin\theta + y \cos\theta$, ω_0 is the radial frequency in radians per unit length, θ is the orientation in radians and σ is the standard

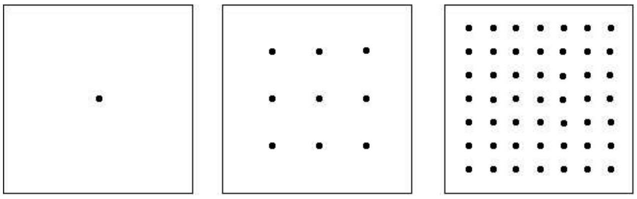


Fig. 1. Sampling points of each level, with level, 1, 2, 3 from left to right

deviation of the elliptical gaussian envelope along the x axes. The Gabor filter is centered at $(x = 0, y = 0)$ in the spatial domain. Also, the elliptical gaussian envelope of the Gabor filter has the aspect ratio $\sigma_y/\sigma_x = r$ and the plane wave's propagating direction along the short axis, where σ_x, σ_y are the standard deviations of elliptical gaussian envelope along the x, y axes[1].

2.2 Hierarchical Gabor Features Extraction

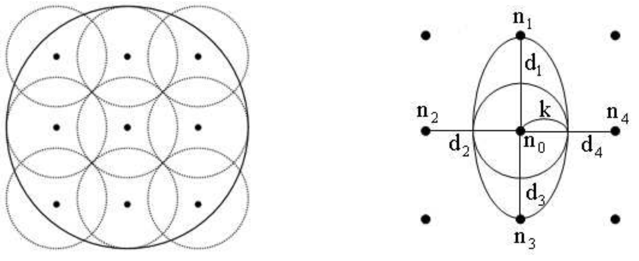
To structure features hierarchically, the features must be able to represent different level information such that the features in the higher level represent more global information and the features in the lower level represent more localized information. First, the Gabor filter banks whose Gabor filters can represent the global or the localized information are defined. Next, the optimal Gabor filters are selected from the Gabor filter banks using a discriminant measure, and the HGFs are then extracted from the optimal Gabor filters.

To define the Gabor filter banks, recursively from the highest level which has only one sampling point, a sampling point is decomposed into nine sub-sampling points in the lower level. This sub-sampling decomposition is shown in Fig.1. The position of a sampling point is the center of a Gabor filter in the spatial domain.

In order to extract information having the global property at a high level and the localized property at a low level from the Gabor filters(See the Fig.2(a)), the standard deviation σ^{ls} must be restricted according to level such that the contour's radius having half of max of the circular gaussian envelope becomes k. From the equation (1), σ^{ls} becomes

$$\sigma^{ls} = \frac{k}{\sqrt{ln2}}, \qquad (2)$$

where ls is the index for a sampling point s at level l, $l = 1, \ldots, N_L$ and $s = 1, \ldots, N_{lS}$. N_L is a level size and N_{lS} is the number of sampling points at level l. To extract the localized information which is not represented in the higher level, k is selected as a half mean of distances, $d1$, $d2$, $d3, d4$, where d_1, d_2, d_3, d_4 are distances from a sampling point to its four neighbor sampling points, n_1, n_2, n_3, n_4(See Fig.2(b)).



(a) (b)

Fig. 2. (a) A big circle shows a region covered by gaussian envelope of the upper sampling point. The smaller nine circles show regions covered by gaussian envelopes of the sub-sampling points (b) A Circle is a contour having half of the max of the circular gaussian envelope with $k = \text{mean}(d_1, d_2, d_3, d_4)/2$. In the case of ellipse, aspect ratio $r = 2$

After the standard deviation σ^{ls} and the aspect ratio r of the Gabor filter are determined, the Gabor filter bank $\mathbf{GB}_j^{ls}$ is defined as

$$\begin{aligned}
&\mathbf{GB}_j^{ls} = \{\, G_{j1}^{ls}, \ldots, G_{jN_\omega}^{ls} \}, \; G_{ji}^{ls}(\, x, y\,) = G(x^{\,ls} - x, y^{\,ls} - y, \omega_i, \sigma^{\,ls}, r, \theta_j), \quad (3)\\
&\omega_i \in \mathbf{\Omega} \quad \text{and} \quad \mathbf{\Omega} = \{\, \omega_1, \ldots, \omega_{N_\omega} \}, \quad i = 1, \ldots, N_\omega\,,\\
&\theta_j \in \mathbf{\Theta} \quad \text{and} \quad \mathbf{\Theta} = \{\, \theta_1, \ldots, \theta_{N_\theta} \}, \quad j = 1, \ldots, N_\theta\,,
\end{aligned}$$

where $\mathbf{\Omega}$ is a set of spatial frequencies, $\mathbf{\Theta}$ is a set of the orientations, and G_{ji}^{ls} is a Gabor filter centered at $(x^{\,ls}, y^{\,ls})$ from the equation (1). $(x^{\,ls}, y^{\,ls})$ is xy-coordinates of the sampling point in an image plane. Thus, for each sampling point and orientation, the Gabor filter bank $\mathbf{GB}_j^{ls}$ is a set of Gabor filters which have different frequencies in the $\mathbf{\Omega}$.

An optimal Gabor filter OG_j^{ls} is selected from $\mathbf{GB}_j^{ls}$ using a discriminant measure. The discriminant measure is a measure of how certain information is efficient for discrimination(See the Appendix). Using the discriminant measure is reasonable because our ultimate goal is classification. Let $(\, h_d, I_d)$ be a pre-classified training image, where $h_d \in \mathbf{C}$ and $I_d \in \mathbf{I}$. $\mathbf{C} = \{c_i\!: \; i = 1, \cdots, N_c, \, N_c :$ the number of classes$\}$ is a set of class hypotheses and $\mathbf{I}$ is a set of training images. An optimal Gabor filter OG_j^{ls} is selected such as

$$\begin{aligned}
&OG_j^{ls} = \underset{G_{ji}^{ls} \;\; \{i\}}{arg\; max}\, (DM_i)\,, \quad DM_i = Discriminant\ Measure\,(\mathcal{X}_i)\,, \qquad (4)\\
&\mathcal{X}_i = \{(h_1, g_1), \ldots, (h_{N_I}, g_{N_I})\}\,, \quad g_d = \sum_{\{x\}} \sum_{\{y\}} I_d(x, y)\, G_{ji}^{ls}(x, y)\,,\\
&G_{ji}^{ls} \in \mathbf{GB}_j^{ls}\,, \quad i = 1, \ldots, N_\omega,
\end{aligned}$$

where g_d is a Gabor filter response of an image and N_I is the number of training images. For each Gabor filter G_{ji}^{ls} in $\mathbf{GB}_j^{ls}$, the Gabor filter responses of all the training images are calculated. Next, the Gabor filter whose frequency gives the highest discriminant measure for the training data set is selected as the optimal Gabor filter OG_j^{ls}.

After obtaining every OG_j^{ls}, the Gabor feature of a sampling point about an image I is defined as

$$\mathbf{a}^{ls} = [a_1^{ls}\ a_2^{ls}\ \ldots\ a_{N_\theta}^{ls}]^T, \quad a_j^{ls} = \sum_{\{x\}} \sum_{\{y\}} I(\,x, y\,)\, OG_j^{ls}(\,x, y\,) \tag{5}$$

The Gabor feature $\mathbf{a}^{ls}$ of a sampling point s at level l becomes an N_θ-dimensional vector whose elements are responses of optimal Gabor filters on an image I for all orientations. Finally the HGFs $\mathbf{a}$ of an image I consists of Gabor features of all the sampling points.

$$\mathbf{a} = \{\,\mathbf{a}^1, \mathbf{a}^2, \ldots, \mathbf{a}^{N_L}\,\}, \quad \mathbf{a}^l = \{\,\mathbf{a}^{l1}, \mathbf{a}^{l2}, \ldots, \mathbf{a}^{lN_{lS}}\,\}, \tag{6}$$

where $\mathbf{a}^l$ is a set of Gabor features of level l.

3 Hierarchical bayesian network

3.1 Bayesian network

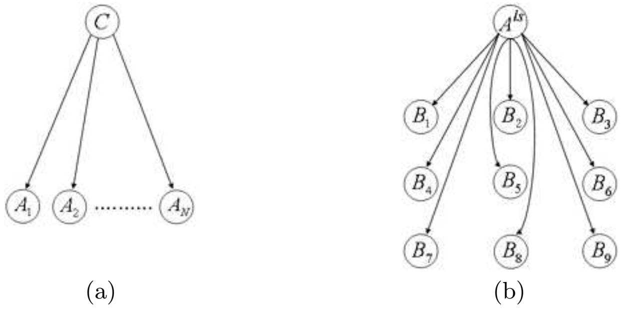
About a finite set of random variables, $\mathbf{U} = \{A_1, \ldots, A_n\}$, a bayesian network[2][3] is generally defined by $< DAG, \mathbf{CP} >$. The $DAG = (\mathbf{V}, \mathbf{E})$, that is, a directed acyclic graph defines the structure of a bayesian network. $\mathbf{V} = \{A_1, \ldots, A_n\}$ is a set of nodes and $\mathbf{E} = \{(A_i, A_j) : A_i, A_j \in \mathbf{V}, \text{where } i \neq j\}$ is a set of direct edges, where (A_i, A_j) denotes directed edge from A_i to A_j which implies that the node A_i affects the node A_j directly. There is a one-to-one correspondence between elements of $\mathbf{V}$ and $\mathbf{U}$. A directed edge set $\mathbf{E}$ represents directed dependencies between the random variables in $\mathbf{U}$. $\mathbf{CP}$ is a set of conditional probability distributions of nodes. The conditional probability distribution of a node A_i is defined by $P(A_i|\,\mathbf{\Pi}_{A_i})$ where $\mathbf{\Pi}_{A_i}$ is the parent node set of A_i in DAG. Also, the joint probability distribution $P(\mathbf{U})$ explained by a bayesian network can be factorized by conditional probability distributions in the $\mathbf{CP}$ and is followed as

$$P(A_1, \ldots, A_n) = \prod_{i=1}^{n} P(A_i|\,\mathbf{\Pi}_{A_i}) \tag{7}$$

For example, the structure of the naive bayesian classifier[4], which does not represent any dependencies among the feature nodes, is shown in Fig.3.(a) and the joint probability explained by the naive bayesian classifier can be factorized such as $P(A_1, \ldots, A_N, C) = \prod_{i=1}^{N} P(A_i|\,C) P(C)$.

3.2 Hierarchical bayesian network

The HBN is constructed to the hierarchical structure so that the Gabor features at a certain level affect the Gabor features at its lower level with the more local property.



(a) (b)

Fig. 3. (a) Structure of bayesian network for naive bayesian classifier (b) Sub-structure of HBN

The structure of HBN, excluding a class node, is defined as the $DAG_H =< \mathbf{V}_H, \mathbf{E}_H >$. Let A^{ls} be a node in $\mathbf{V}_H$ or a random variable in $\mathbf{U}_H$. There is a one-to-one correspondence between a sampling point s at level l in section.2.2 and a node A^{ls}, that is, a random variable A^{ls} has a value as the Gabor feature $\mathbf{a}^{ls}$. Thus, the node set $\mathbf{V}_H$, for the HGFs, becomes

$$\mathbf{V}_H = \mathbf{A}^1 \cup \cdots \cup \mathbf{A}^{N_L} \,, \quad \mathbf{A}^l = \{A^{l1}, \ldots, A^{lN_{lS}}\} \,, \tag{8}$$

where $\mathbf{A}^l$ is a set of nodes at level l. A node set $\mathbf{\Phi}^{\,ls}$, nodes for nine subsampling points of A^{ls}, is defined as

$$\mathbf{\Phi}^{\,ls} = \{B_1^{ls}, \ldots, B_9^{ls}\}, \quad B_i^{ls} \in \mathbf{A}^{l+1} \,, \tag{9}$$

where $l = 1, \ldots, N_L - 1$. Thus, a directed edge set $\mathbf{E}_H$ is defined as

$$\begin{aligned} &\mathbf{E}_H = \mathbf{E}^1 \cup \cdots \cup \mathbf{E}^{N_L - 1} \,, \\ &\mathbf{E}^l = \mathbf{E}^{\,l1} \cup \ldots \cup \mathbf{E}^{\,lN_{lS}} \,, \\ &\mathbf{E}^{\,ls} \;= \{(A^{ls}, B_1^{ls}), \ldots, (A^{ls}, B_9^{ls})\} \,, \end{aligned} \tag{10}$$

where (A^{ls}, B_i^{ls}) is a directed edge from A^{ls} to B_i^{ls}, $B_i^{ls} \in \mathbf{\Phi}^{\,ls}$ and level $l = 1, \ldots, N_L - 1$.

In the hierarchical structure DAG_H of HBN, the node A^{ls} affects the nodes in $\mathbf{\Phi}^{\,ls}$ corresponding to its nine sub-sampling points at level $l+1$ (See Fig.3(b)). Thus, directed dependencies from a node to nodes in the lower level are limited to the nodes of nine sub-sampling points.

For classification, the hierarchical structure DAG_H must be modified to $DAG_H{}'$ for including the class node C. $DAG_H{}' =< \mathbf{V}_H{}', \mathbf{E}_H{}' >$ is defined as

$$\begin{aligned} \mathbf{V}_H{}' \;&= \mathbf{U}_H{}' = \mathbf{V}_H \cup \{C\} \,, \\ \mathbf{E}_H{}' \;&= \mathbf{E}_H \cup \mathbf{E}_c \,, \quad \mathbf{E}_c = \{(C, A^{ls}) : \text{ for all } A^{ls} \in \mathbf{V}_H\} \,, \end{aligned} \tag{11}$$

where $\mathbf{E}_c$ is a set of directed edges from the class node C to all nodes in the set $\mathbf{V}_H$. All nodes excepting node C in $DAG_H{}'$ have node C as its parent.

Table 1. The number of testing data per class

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

For the complete definition of HBN with hierarchical structure ${DAG_H}'$, a set of conditional probability distributions, denoted by $\mathbf{CP}$, must be defined. HBN has mixed types of continuous and discrete variables. The variables in $\mathbf{U}_H$ for HGFs are continuous and only the class variable C is discrete. Thus, for each continuous Gabor feature variable A^{ls}, the conditional probability distribution $P(A^{ls}|{\mathbf{\Pi}_{ls}}')$ is defined as a conditional multivariate gaussian[3], where ${\mathbf{\Pi}^{ls}}'$ is a set of parents of A^{ls} in ${DAG_H}'$. Also, for the discrete class variable C which does not have any parents, the conditional probability distribution $P(C)$ is defined as a multinomial distribution[3]. The joint probability distribution of ${\mathbf{U}_H}'$ can be factorized by the conditional probability distributions such as equation (7).

With the HBN defined as $< {DAG_H}', \mathbf{CP} >$, an inference can be made by a belief propagation algorithm in [2][3]. As the interesting variable is the class variable C for classification, inference is performed for $P(C|\mathbf{U}_H = \mathbf{a})$, where $\mathbf{a}$ is an instance of the HGFs of an image from (6). Afterwards, the instance $\mathbf{a}$ is assigned to a class label maximizing $P(C|\mathbf{U}_H = \mathbf{a})$ for classification.

4 Experimental Results

Our HBN was simulated about binary handwritten numerical data set for recognition[5]. This numerical data set was obtained from the UCI(University of California, Irvine) databases[6].

The experiments were conducted with the following conditions for comparison with other methods. The training data set consisted of randomly chosen 500 samples(50 per class) and the testing data set consisted of the remaining 1,943 samples. The number of testing data set per class is shown in Table.1. The training and testing data set were not overlapped. For extracting the HGFs, the imaginary part of Gabor filter was used. The parameters of Gabor filter banks were set up such as $\mathbf{\Omega}$=$\{0.025, 0.05, 0.075, 0.1, 0.15, 0.2, 0.4, 0.6, 0.8, 1\}$ for frequencies, $\mathbf{\Theta}$=$\{0, \frac{1}{4}\pi, \frac{1}{2}\pi, \frac{3}{4}\pi\}$ for orientations, aspect ratio $r = 2$, and the level size $N_L = 3$.

Experiment 1 : Our proposed HBN was simulated. From the training handwritten numerical character images, the HGFs were extracted such as in section 2.2. After constructing the hierarchical structure of HBN, the parameters of the conditional probability distributions of the HBN were learned by the maximum likelihood(ML) method from the HGFs of training images[2][3].

Experiment 2 : The naive bayesian classifier[4](See Fig.3(a)), which had exactly the same nodes of the HBN and the HGFs in the experiment 1, was simulated.

Table 2. Recognition results with 90% confidence interval. HBN : hierarchical bayesian network, NBC : naive bayesian classifier, KNN : k-nearest neighbor classifier, NN : backpropagation neural network

Experiment 1	Experiment 2	Experiment 3	Experiment 4
HBN	NBC	KNN with $k = 1$	NN
0.9675 ± 0.0031	0.9567 ± 0.0031	0.9565 ± 0.0040	0.9451 ± 0.0048

Experiment 3 : For the inputs of the k-nearest neighbor classifier[7][8], the HGFs in the experiment 1 were modified to a 236(=59 × 4) dimensional feature vector, where 59 was the number of the Gabor features in the HGFs and 4 was the dimension of a Gabor feature. In these experiments the k-nearest neighbor classifiers with $k = 1, 3, 5$ were simulated. In this experiment, the case of the $k = 1$ showed the best recognition result.

Experiment 4 : The number of input nodes of backpropagation neural network[7][8] were set up to 236(=59 × 4) to accept the same HGFs in experiment 1. Also, the parameters of the backpropagation neural network were set up such as 150 hidden units, learning rate $\eta = 0.01$, momentum rate $\alpha = 0.5$, number of learning iteration = 10,000.

The results of the experiments are shown in Table.2. From the results, it is reliable that the HGFs are efficient for recognition in spite of relatively small training data set. That the hierarchical dependencies within the HBN for the HGFs more improve the recognition is explained from that the HBN outperformed over all other methods which do not represent any hierarchical dependencies.

5 Conclusion

In this paper we have proposed a HGFs and HBN for a recognition mechanism. To represent the hierarchical property with the HGFs, we decomposed a sampling point into nine sub-sampling points and adjusted covered regions of Gabor filters with levels. And the optimal Gabor filters were selected using the discriminant measure. To represent dependencies within the HGFs, we constructed a bayesian network structure to be hierarchical by analogy of the HGFs extraction method.

Our proposed method was applied to the problem of handwritten digit recognition and compared it with other methods, such as the naive classifier, k-nearest neighbor classifier, and backpropagation neural network. The results confirmed the useful behavior of our method in which the HGFs are well structured information and the hierarchical dependencies in the HBN improve recognition.

Although we only applied this approach to the problem of the handwritten digit recognition, we believe our method can be extended to a general recognition system.

References

1. Tai Sing Lee.: Image Representation Using 2D Gabor Wavelets. IEEE trans. PAMI, vol. 18. no. 10. (1996) 959-971
2. Pearl, J. : Probabilistic Inference in Intelligent Systems. Morgan Kaufmann, San Mateo, California, (1988)
3. Robert G. Cowell, A. Philip Dawid, Steffen L. Lauritzen, David J, Spiegelhalter.: Probabilistic Networks and Expert Systems. Springer (1999)
4. N. Friedman, D. Geiger, M Goldszmidt.: Bayesian network classifiers. Mach. Learn 29 (1997) 131-163
5. Daijin Kim, Sung-Yang Bang.: A Handwritten Numeral Character Classification Using Tolerant Rough Set. IEEE trans. PAMI, vol. 22. no. 9. (2000) 923-937
6. C. L. Blake and C. J. Merz.: UCI Repository of Machine Learning Databases. Dept. of Information and Computer Science, Uive. of California, Irvine, (1998)
7. Chistopher M. Bishop.: Neural Networks for Pattern Recognition. Oxford University Press (1995)
8. Richard O. Duda, Peter E. Hart, David G. Strok.: Pattern Classification. John Wily and Sons. (2001)

Appendix : Discriminant Measure

A discriminant measure is a measure to how certain information is efficient for discrimination. The discriminant measure is defined by a within-class scatter matrix and a between-class scatter matrix. If within the one class, scatter of information are smaller and among the classes, the scatter of information are larger[8], this discriminant measure gives a higher output.

For the c-class problem, suppose that a set of n d-dimensional instances, $\mathcal{X}$, have its elements such as $\mathbf{x}_1, \ldots, \mathbf{x}_n$, n_i in the subset $\mathcal{X}_i$ labeled c_i. Thus within-class scatter matrix $\mathbf{S}_W$ is defined by

$$\mathbf{S}_i = \sum_{\mathbf{x}\in\mathcal{X}_i} (\mathbf{x}-\mathbf{m}_i)(\mathbf{x}-\mathbf{m}_i)^T, \qquad \mathbf{m}_i = \frac{1}{n_i}\sum_{\mathbf{x}\in\mathcal{X}_i} \mathbf{x} \qquad \mathbf{S}_W = \sum_{i=1}^{c} \mathbf{S}_i,$$

where T is matrix transpose. After defining a total mean vector $\mathbf{m}$, between-class scatter matrix $\mathbf{S}_B$ is defined as

$$\mathbf{m} = \frac{1}{n}\sum_{\mathbf{x}\in\mathcal{X}} \mathbf{x} = \frac{1}{n}\sum_{i=1}^{c} n_i \mathbf{m}_i, \qquad \mathbf{S}_B = \sum_{i=1}^{c} n_i(\mathbf{m}_i-\mathbf{m})(\mathbf{m}_i-\mathbf{m})^T$$

A simple scalar measure of scatter is the determinant of the scatter matrix. From this scatter measure, *Discriminant Measure* is

$$\textit{Discriminant Measure} = \frac{|\mathbf{S}_B|}{|\mathbf{S}_W|}, \text{ where } |\cdot| \text{ denotes determinant.}$$

In our approach, an instance $\mathbf{x}$ is a scalar.