A Hybrid SVM-HMM Method for Handwritten Numeral Recognition

Eui Chan Kim*, and Sang-Woo Kim**

* Department of Electronic and Electrical Engineering, POSTECH, Pohang, Korea
(Tel : +82-54-279-5018; E-mail: cruise75@postech.ac.kr)
**Department of Electronic and Electrical Engineering, POSTECH, Pohang, Korea
(Tel : +82-54-279-2237; E-mail: swkim@postech.ac.kr)

Abstract: The field of handwriting recognition has been researched for many years. A hybrid classifier has been proven to be able to increase the recognition rate compared with a single classifier. In this paper, we combine support vector machine (SVM) and hidden Markov model (HMM) for offline handwritten numeral recognition. To improve the performance, we extract features adapted for each classifier and propose the modified SVM decision structure. The experimental results show that the proposed method can achieve improved recognition rate for handwritten numeral recognition.

Keywords: Handwritten numeral recognition, Support vector machine, Hidden Markov model, Principal component analysis, Modified SVM decision structure

1. INTRODUCTION

The support vector machine (SVM) is a kind of learning machine, whose fundamental is statistics learning theory. It has been used as a powerful recognizer due to a strong discrimination capability. We combine the SVM with the hidden Markov model (HMM) which provides a good probabilistic representation of sequential inputs. In addition, we propose a modified SVM decision structure. This improves the conventional multiclass SVM strategies for handwritten numeral recognition.

Section 2 shows the system architecture for handwritten numeral recognition. In section 3, we review the background of SVM, HMM and principal component analysis (PCA). Section 4 details the recognition procedure including modified SVM decision structure and recognition results. Finally, we present the conclusion in section 5.

2. THE SYSTEM ARCHITECTURE

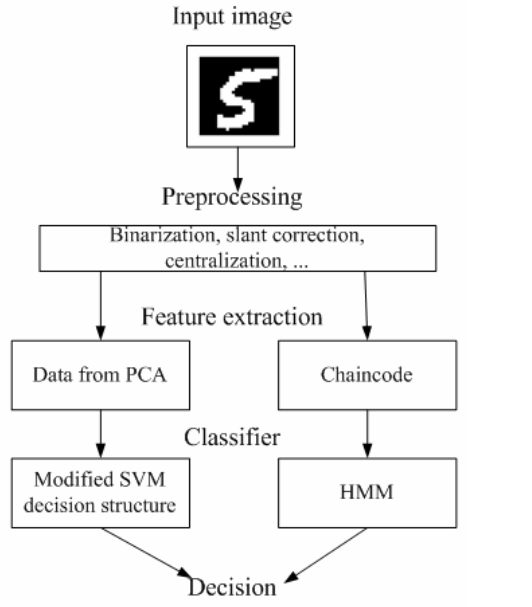

Fig. 1 The system architecture

Figure 1 shows the handwritten digit recognition system. From the digit images with resolution of 28×28 pixels, we preprocess images. We extract appropriate features for each classifier, then combine the each result.

3. BACKGROUND

3.1 SVM

The basic idea of the SVM utilized in pattern recognition is to construct a hyperplane as the decision plane, which separates the positive and negative patterns with the largest margin. The positive and negative samples are denoted by $\{x_i, y_i\}$, $i = 1,2,...,l$, $y_i \in \{-1,1\}$, $x_i \in R^d$. Suppose they are completely separated by a *d*-dimension hyperplane described by

$$< w \cdot x > + b = 0 \tag{1}$$

The classification is done by

$$\begin{aligned} &< w \cdot x_i > + b \geq +1, \quad \text{for } y_i = +1 \\ &< w \cdot x_i > + b \leq -1, \quad \text{for } y_i = -1 \end{aligned} \tag{2}$$

The SVM is to find the hyperplane, which has the largest margin $2/\|w\|$. The primal problem is

$$\begin{aligned} &\min_{w,b} \ \Phi(w) = \frac{1}{2}\|w\|^2 \\ &\text{subject to } y_i(< w \cdot x_i > + b) \geq 1, \quad i = 1,2,...,l \end{aligned} \tag{3}$$

Generally, for linearly non-separable case, the non-negative slack variables $\xi = (\xi_1, \xi_2, ..., \xi_l)$ are introduced and the linear constrained conditions change into

$$y_i(< w \cdot x_i > + b) \geq 1 - \xi_i, \quad i = 1,2,...,l \tag{4}$$

The primal problem of non-separable case becomes

$$\begin{aligned} &\min_{w,b,\xi} \ \Phi(w) = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{l}\xi_i \\ &\text{subject to } y_i(< w \cdot x_i > + b) \geq 1 - \xi_i, \quad i = 1,2,...,l \end{aligned} \tag{5}$$

where C is a positive constant.

Introducing Lagrange multiplier the dual problem of linearly non-separable one is gotten :

$$\max_{\alpha} W(\alpha) = \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{l} y_i y_j \alpha_i \alpha_j < x_i \cdot x_j > \quad \text{subject to } \sum_{i=1}^{l} y_i \alpha_i = 0, \ 0 \le \alpha_i \le C, \ i = 1,2,...,l \tag{6}$$

Those samples with $\alpha_i > 0$ are defined as Support Vector (SV) that is used to determine w^* and the separating hyperplane.

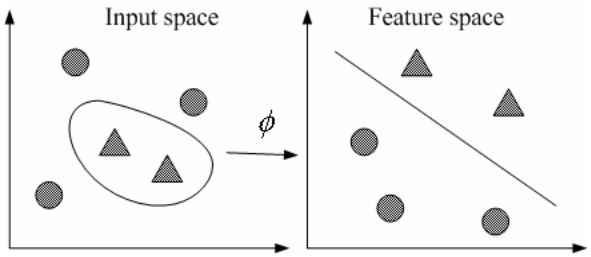

Fig. 2 The idea of Support vector machine

Figure 2 shows the idea of SVM, which is to map the data into some other dot product space (called the feature space) F via a nonlinear map $\phi : R^d \to F$ and perform the above linear algorithm in F.

Substitute input variable x with feature vector $\phi(x)$:

$$d(x) = < w^* \cdot \phi(x) > + b^* = \sum_{i=1}^{l} y_i \alpha_i^* < \phi(x_i) \cdot \phi(x) > + b^* \tag{7}$$

If there exist Kernel Function $K(x,y)$, which satisfies

$$K(x,y) = < \phi(x) \cdot \phi(y) > \tag{8}$$

then

$$d(x) = \sum_{i=1}^{l} y_i \alpha_i^* K(x_i, x) > + b^* \tag{9}$$

Table 1 Typical kernel function in SVM

Feature of the SVM	Kernel function
Linear function	$K(x,y) = < x \cdot y >$
Polynomial kernel	$K(x,y) = < x \cdot y + 1 >^d$
Gaussian kernel	$K(x,y) = \exp(-\frac{\|x-y\|^2}{2\sigma^2})$
Multi-layer neural net	$K(x,y) = \tanh(\gamma < x \cdot y > -\theta)$

3.2 HMM

The HMM is a statistical model of observation sequences. These are supposed to be produced by a system that changes state at regular steps. The set of possible states has finite size N and the transition from any state i to any other state j is governed by a stochastic process. The probability of being in state s_t at step t cam be expressed as

$$p(s_t) = p(s_t \mid s_{t-1}, s_{t-2}, ..., s_{t-d}) \tag{10}$$

but, in general, the assumption d=1 is made so that the state at step t depends only on the state at step t-1. The model is then called a first order HMM. A second important assumption is that the transition probability does not depend on t and is then stationary. The evolution dynamic of the system is entirely represented by the matrix $A = \{a_{ij}\} = \{p(s_t = j \mid s_{t-1} = i)\}$.

To start the process, an initial state probability π_i (probability of being in state i at first step) must also defined. The vector $\pi = \{\pi_1, \pi_2, ..., \pi_N\}$ represents the initial state probability distribution. The sequence of the states is assumed to be not observable (thus the name hidden), but gives rise to a second stochastic process characterized by the probabilities $b_j(o)$ to emit observation o while being in state j. The b_j are probability density functions in the space of the observations and are continuous or discrete depending on the nature of the observations themselves. The set $B = \{b_1(o), b_2(o), ..., b_N(o)\}$ is the bridge between the hidden states and the observations.

A Hidden Markov Model λ is represented by the set $\{A, B, \pi\}$. The calculation of the parameters in A, B and π (the training), is performed with the Baum-Welch Algorithm, a particular form of the EM algorithm [6].

Once a model λ is trained, the probability $P(O \mid \lambda)$ of an observation sequence O being produced by the model λ is calculated usually with the Viterbi algorithm, which gives an acceptable approximation of the probability with a reduced computational effort [6].

3.3 PCA

Principal component analysis (PCA) is a well-established technique for dimension reduction. It replaces the original variables of a data set with a smaller number of uncorrelated variables called the principal components. Given N images, convert each image to a column vector of length D so we have a set of N D-dimensional data points :

$$x_1, x_2, x_3, ..., x_N \tag{11}$$

Then, the mean and covariance matrix can be calculated :

$$m = \frac{1}{N} \sum_{i=1}^{N} x_i \tag{12}$$

$$\sum = \frac{1}{N} \sum_{i=1}^{N} (x_i - m)(x_i - m)^T \tag{13}$$

Defining a projection matrix A_d composed of the d eigenvectors of covariance matrix with highest eigenvalues, the d-dimensional representation of an original is given by the projection :

$$y_i = A_d^T (x_i - m) \quad (14)$$

It is used as inputs for recognition. And an image is reconstructed from its eigen-space representation by :

$$x = A_d y + m \quad (15)$$

4. HANDWRITTEN NUMERAL RECOGNITION

4.1 Modified SVM decision structure

We focus on the SVM classifier and the HMM is used as a complement for decision. Especially, we modify the decision structure of the multiclass SVM.

The SVM is a binary classifier. Various approaches have been developed in order to deal with multiclass classification problems : *K* one-against-all, *K(K-1)/2* one-against-one and trees of SVM [2][3].

Figure 3 shows the modified SVM decision structure. We need 55 SVMs (*K+K(K-1)/2*=55, where *K*=10) in the training phase. Instead of one-against-all SVM, we propose one-against-necessary other SVM. This is based on excluding unnecessary data in 'all' term of one-against-all. For example, we just need not all, but '3' & '4' in '0'-SVM : we call '0' as main-term and '3', '4' as sub-term.

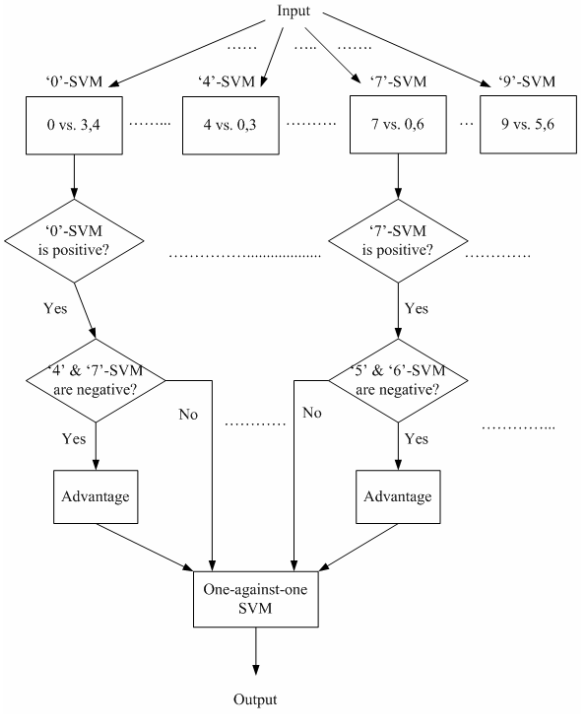

Fig. 3 Modified SVM decision structure

The test procedure is as follows :

1) 10 one-against-necessary other SVMs are used first. If the output of 'N'-SVM is positive, go to the stage 2).
2) If the output of SVM in which 'N' has been used as sub-term is negative, 'N' has an advantage, +1 vote. For example, in the case of '0', if outputs of '4'-SVM and '7'-SVM are all negative, '0' has an advantage.
3) We use one-against-one SVMs for remaining numerals with advantageous votes.

The recognition rate in the first stage can be increased as compared with the tree SVM and the number of SVMs used in the test phase is less than one-against-one SVM structure.

4.2 The recognition results

The database for our experiments is from the MNIST database and slab ID numbers of POSCO which is Korean steel manufacturing company : 10,000 images for training and 5,000 images for testing. We preprocess images using the binarization, slant correction and centralizing. Then, we extract the appropriate features (the dimension reduced data from PCA and the chaincode) for each classifier, SVM and HMM. We reduce the dimension from 784 to 100 and get the chaincode from a contour feature [8].

We have used a SVM classifier with Gaussian kernel. 'Max win' voting strategy has been used in one-against-one SVM. The tree SVM has been built using k-means algorithm : (0,1,4,7) vs. (2,3,5,6,8,9).

In decision, we apply the following rule :

1) If the max output of the SVM is just one, the HMM is ignored in decision.
2) If the max output of the SVM is two at least, the decision is dependent on the output of the HMM.

The results show that systems with modified SVM decision structure provide better recognition rates than conventional multiclass SVMs. However, a hybrid method has a drawback for test speed.

Table 2 Recognition results

	Recognition rate	Test time (sec/numeral)
One-against-all	97.12	0.13
One-against-one	97.58	0.43
Tree SVM	97.08	0.11
Modified SVM	97.78	0.21
Modified SVM + HMM	97.84	0.46

5. CONCLUSION

In the paper, a hybrid SVM-HMM method is introduced for handwritten numeral recognition. We use adapted features for each classifier and combine each output in the last. To improve the performance, we specially propose the modified SVM decision structure. It seems to combine the conventional multiclass SVMs and uses only necessary data in one-against-all SVM stage. This improves the accuracy of tree SVM and the test speed of one-against-one SVM. A hybrid SVM-HMM method has the highest recognition rate in the experiment, but the test speed is slow. We need to consider the feature extraction for the HMM.

REFERENCES

[1] Nello Cristianini and John Shawe-Taylor, *An Introduction to Support Vector Machines*, Cambridge University Press, 2000.

[2] Zhao Bin, Liu Yong and Xia Shao-Wei, "Support vector machine and its application in handwritten numeral," *Proceedings of 15th International Conference on Pattern Recognition*, Vol. 2, pp. 720-723, 2000.

[3] Chih-Wei Hsu and Chih-Jen, "A comparison of methods for multiclass support vector machines," *IEEE Trans. on Neural Networks*, Vol. 13, No. 2, pp. 415-425, 2002.

[4] A. Bellili, M. Gilloux and P. Gallinari, "An hybrid MLP-SVM handwritten digit recognizer," *Proceedings of Sixth International Conference on Document Analysis and Recognition*, pp. 28-32, 2001.

[5] R. O. Duda, P. E. Hart and D. G. Stork, *Pattern Classification, 2nd Edition*, John Wiley & Sons, Inc., 2001.

[6] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of the IEEE*, Vol. 77, No. 2, pp. 257-286, 1989.

[7] Alessandro Vinciarelli, "A survey on off-line cursive word recognition" *Pattern Recognition*, Vol. 35, No. 7, pp. 1433-1446, 2002.

[8] C. M. Travieso, C. R. Morales, I. G.. Alonso, and M. A. Ferrer, "Handwritten digits parameterisation for HMM based recognition," *Seventh International Conference on Image Processing and Its Applications*, Vol. 2, 1999.

[9] R. C. Gonzalez and R. E. Woods, *Digital Image Processing, 2nd Edition,* Prentice-Hall, Inc., 2002.

[10] M. A. Turk and A. P. Pentland, "Face recognition using eigenfaces," *Proceedings CVPR of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 586-591, 1991.