Two Novel Learning-Based Criteria and Methods Based on Multiple Classifiers for Rejecting Poor Handwritten Digits

Weina Wang

A Thesis
in
The Department
of
Computer Science and Software Engineering

Presented in Partial Fulfillment of the Requirements
for the Degree of Master of Computer Science at
Concordia University
Montreal, Quebec, Canada

April 2013

CONCORDIA UNIVERSITY
School of Graduate Studies

This is to certify that the thesis prepared

By: Weina Wang

Entitled: Two Novel Learning-Based Criteria and Methods Based on Multiple Classifiers for Rejecting Poor Handwritten Digits

and submitted in partial fulfillment of the requirements for the degree of

Master of Computer Science

complies with the regulations of the University and meets the accepted standards with respect to originality and quality.

Signed by the final Examining Committee:

_____ Chair
Dr. Y. Yan

_____ Examiner
Dr. B. Fung

_____ Examiner
Dr. L. Lam

_____ Supervisor
Dr. C. Y. Suen

Approved by _____
Chair of Department or Graduate Program Director

_____ 20__ _____
Dr. Drew, Robin A. L., Dean
Faculty of Engineering and Computer Science

Abstract

Two Novel Learning-Based Criteria and Methods Based on Multiple Classifiers for Rejecting Poor Handwritten Digits

Weina Wang

In pattern recognition, the reliability and the recognition accuracy of a classification system are of same importance, because even a small percentage of errors could cause a huge loss in real-life handwritten numeral recognition systems, like cheque-reading at financial institutions.

Aiming at improving the reliability of recognition systems, this thesis presents two novel learning-based rejection criteria for single classifiers including SVM-based measurement (SVMM) and Area Under the Curve measurement (AUCM).

Voting based combination methods of multiple classifier system (MCS) are also proposed for rejecting poor handwritten digits. Different rejection criteria (FRM, FTRM and SVMM) are individually combined with MCSs as weight parameters in voting. This method is then evaluated on three renowned databases including MNIST, CENPARMI and USPS. Experimental results indicate that these combinations improve the rejection performances consistently. To further improve the performance of the MCS based rejection method, specialist information has been integrated into the combination process by introducing a new confidence weight parameter. The best result on MNIST is obtained by the simpler one of the two proposed methods of deriving this parameter, which reaches 100% reliability with a rejection rate of only 4.09%, the best value in this field.

Acknowledgements

I would like to thank my supervisor, Dr. Ching Y. Suen, for his patient guidance, encouragement and advice throughout my time as his student. I feel extremely fortunate to have a supervisor who cared so much about my work and responded so promptly to all my questions. This thesis could not be accomplished without his valuable suggestions and enthusiasm.

I must also express my gratitude to Dr. Louisa Lam who gave me a lot of creative suggestions and comments on my research work.

In particular, I would like to thank my boyfriend, Mr. Zi Ye, for his endless support and encouragement. He has always been there cheering me up and stood by me through the good times and bad.

Completing this work would have been more difficult without the support and friendship provided by other Centre for Pattern Recognition and Machine Intelligence (CENPARMI) members: Dr. Andreas Fischer, Fariba Haghbin, Adnan Abueid, Muna G. Al-Khayat, Mahdi Biparva and Ms. Guilin Guo, and two exchange Ph.D students: Boyuan Feng and Xuyao Zhang,

Special thanks to our Research Manager, Mr. Nicola Nobile, for his excellent technical support and to our Secretary, Ms. Marleah Blom, who cheer up all CENPARMI students by organizing activities. I also want to thank Ms. Phoebe Chan for her considerable efforts in editing and proofreading my thesis.

Finally, I am eternally grateful to my wonderful parents and friends who gladly and selflessly sacrificed their time and effort during my research endeavor.

Table of Contents

List of Figures

List of Tables

Chapter 1: Introduction

An overview of the research topic, purpose, challenge, previous works and the outline of this thesis will be presented in this chapter. Section 1.1 will provide a brief description of the research topic. Then, the following Sections 1.2 and 1.3 will explain the purpose of this topic and the major challenges respectively. Section 1.4 will review some of the previous works that has been completed in this field. An overall description of our new method will be depicted in Section 1.5 and finally Section 1.6 will provide the outline of this thesis.

1.1 Research Topic

Pattern recognition contains many branches including character recognition, object recognition, voice recognition, face recognition and etc, among which, handwritten recognition has been studied extensively for the last several decades. To achieve the goal of creating a machine that could recognize human's handwriting with as few errors as possible, tremendous efforts have been made, making handwriting recognition important and intriguing to researchers. Two main types, online and offline are known in the field of handwriting character recognition. Considering online recognition utilizes real time information that is not available to the offline one, discrepancies are shown between performances. As a result, the offline handwriting recognition requires continuous improvement which explains why more research is needed in the field. The main goal of this thesis is to further improve the performance

of offline handwriting recognition system, especially on unconstrained numeral tasks, allowing the system's reconfiguration in enhancing its accuracy and reliability.

1.2 Motivation

Handwritten numeral recognition is playing a significant role in solving handwriting recognition problems, as it is helpful in a variety of specific applications such as cheque processing at the financial institutes, ZIP codes reading in the postal system and numbers extracting from forms. A lot of this work that was used to be conducted by human beings can now be performed by automatic systems with high accuracy rates with the help of handwriting recognition technology. Actually, some handwritten recognition systems have already been developed and used in real-world applications [1, 2].

However, as in most of the other pattern recognition systems, errors still persist in any handwriting recognition systems for the reason that it is the machine, instead of human, who is conducting the recognition job. Misclassifications can be caused by a lot of unpredictable reasons such as confusing nature of some pairs of samples, the width of the tip of the pen, different people's writing styles, cursive writing, low quality of scanning instruments, etc. Hence, some handwritten characters cannot be classified correctly even by human beings [3]. Although a recognition system learns from a large amount of training data inputs, it is requested to classify totally unknown data in the testing set. That is why a perfect recognition rate is still difficult to attain. Therefore, our goal is to enable automation of handwriting recognition systems through the improvements of recognition rate along with the reliability so that the

systems will be eventually adopted by institutions.

1.3 Challenge

In pattern recognition, the recognition rate is always an important factor in evaluating the classifier's performance. Plenty of classifiers or multiple classifier systems have achieved high recognition rates based on different datasets like MNIST digit database [4], CENPARMI digit database [5], USPS handwritten digit database [6], NIST character database [7], and so forth in the past decades. Although some models have reached error rates of less than 1% on the benchmark MNIST dataset and CENPARMI numeral dataset [8, 9], 100% recognition accuracy is still unattainable. Therefore, disparity continues to exist between researches in the lab and usages in practical applications. In real world applications, a small percentage of errors in recognition could still cause an enormous loss at financial institutions. Even if they may be discovered later without any fiscal loss, much resources would be spent through labor and time loss. So, it is necessary to build systems that focus on the reliability, as illustrated through formulas, to prevent this scenario from occurring.

$$Recognition\ rate = \frac{Number\ of\ correct\ samples}{Total\ number\ of\ testing\ samples} \tag{1}$$

$$Rejection\ rate = \frac{Number\ of\ rejected\ samples}{Total\ number\ of\ testing\ samples} \tag{2}$$

$$Reliability = \frac{Number\ of\ correct\ samples\ among\ non\text{rejected}\ \ ones}{Total\ number\ of\ testing\ samples - \text{number of rejected samples}} \tag{3}$$

In order to improve a classifier's reliability, some confusing patterns must be rejected before entering the testing loop in order to prevent errors. That is why some

useful rejection criteria are produced to determine and filter out the confusing samples. The main challenge is to design rejection criteria that can keep high reliabilities with as few samples rejected as possible.

1.4 Previous Works

Handwriting recognition has been intensively investigated by researchers for several decades and many of them have made extraordinary achievements in improving recognition accuracy and reliability. In this section, recent studies of offline handwriting numeral recognition and some benchmark rejection criteria will be introduced.

During the research history of offline handwritten isolated digits recognition, various classic statistical classifiers have been applied to solve the problem, such as K-Nearest Neighbors (KNN), Fisher discriminant analysis [10], Modified Quadratic Discriminant Function (MQDF) [11] and so forth. In addition, many improved machine learning classifiers are widely adopted in this field, including Multi-Layer Perceptrons (MLP) [12], Radial Basis Function networks (RBFs) [13], Polynomial Classifier (PC) [14, 15] and so on. Among these classifiers, Support Vector Machine (SVM) is the most popular one, not only because of its simpler model when compared to many others; but also its outstanding recognition ability in various branches of pattern recognition such as face recognition [16], text recognition [17], speech recognition [18], and handwriting recognition [8, 19]. The introduction of the deep learning idea by LeCun et al [20] makes the research of handwriting recognition step into a new era.

Most of the studies focus on increasing the recognition rate by choosing more recognition-sensitive features and by designing more effective classification models. For feature extraction, many approaches have been introduced [21] and among them, directional feature has been proven to be one of the most effective features in handwriting recognition [22]. Liu et al [9] pre-processed images with normalization and blurring, and extracted different types of features for recognition afterwards. With a SVM based on 8 direction gradient features, an error rate of only 0.85% was obtained on CENPARMI numeral dataset. They also evaluated the proposed pre-process method on NIST numeral dataset which yielded a recognition rate of 99.47% with the same features based on discriminative learning quadratic discriminant function (DLQDF) [23]. LeCun, one of the fore-runners of deep learning algorithm, achieved a recognition rate of 99.05% with the proposed LeNet5 [20] Convolutional Neural Network (CNN) model and 99.30% with the boosted LeNet4 CNN model on MNIST numeral database [4]. Simard et al proposed elastic distortion algorithm to expand datasets and gained an error rate of 0.40% with simple CNN model [24]; Lauer et al introduced a novel TFE-SVM classifier which used LeNet5 CNN model in trainable feature extracting and performed the recognition tasks with a SVM. It outperformed either of the single models. By adopting the training set expanding method used by Simard et al in 2003, it achieved error rates of 0.56% and 0.54% with elastic and affine distortion respectively based on MNIST digit dataset [25].

Later, researchers shifted their focuses from single classifiers to Multiple

Classifier System (MCS) which consists of several different classifiers in order to improve the individuals' performances. MCS is supposed to perform better than single ones for the reason that different classifiers are sensitive to different features or samples and the ensemble system can combine the decisions of several classifiers and make a final decision. Lam et al [26] implemented Bayesian combination algorithm and a weighted majority voting method to combine 7 different classifiers. The combination system was then evaluated on handwritten numerals and proven that combination of classifiers can improve the performance of single ones. Meanwhile, Suen et al [27] applied different combination methods to different types of outputs which produced higher recognition rates. Yet, the better results were accompanied with higher costs. Recently, some researchers have yielded state-of-the-art performances in handwritten numeral recognition based on differently designed MCSs. Recognition rates of 99.77% on the MNIST numeral dataset and 99.23% on NIST SD19 [7] digits dataset are achieved with an MCS consisting of 35 CNN classifiers by Ciresan et al [28]. They built the 35 committees by normalizing the width of all characters and randomly initializing CNN models. Wu et al obtained the same recognition rate of 99.77% on MNIST digits based on a cascade-based MCS with 5 CNNs trained on different training sets as well as different operations of spatial pooling [29]. Niu et al produced the best recognition rate so far: 99.81% on MNIST numeral dataset, with a hybrid classifier consisting of a CNN model for feature extracting and a SVM model for classification [30].

As the classifier's reliability became increasingly important, this research area

attracted plenty of researchers who sought to produce reliable handwritten recognition systems for practical applications. As a result, some useful rejection criteria have been created. He and Suen [31] proposed a Linear Discriminant Analysis Measurement (LDAM) rejection criterion based on Linear Discriminant Function (LDF) method [10] and tested its performance on different handwriting numeral datasets. The results proved that it surpassed other classic rejection criteria including the First Rank Measurement [19] and First Two Rank Measurement FTRM [5] in performance. They also introduced another two rejection criteria including Differential Measurement (DM) and Probability Measurement (PM), and a hybrid system consisted of a SVM, a MQDF, a CNN and the combination of the three. The hybrid system achieved recognition rates ranging from 95.54% to 99.11% with a reliability of 99.54% to 99.11% [32]. A cascade-based MCS was proposed and applied for the purpose of handwritten digits recognition and rejection by Zhang [33]. The results of 99.96% reliability with minimal rejection and 99.59% recognition rate without rejection indicated that this method could enhance the performances in both recognition rate and reliability.

Based on this literature review, we design two novel learning-based rejection criteria for single classifiers, as well as attempting to conduct rejection with multiple classifiers which will be discussed in the next section.

1.5 Proposed Methods

In this thesis, our work is mostly focused on handwritten numerals. Considering that current recognition systems is unable to achieve 100% recognition rate and that

mistakes may cause extensive damage in the long run, a classifier's reliability, defined in Eq.(1, 2, 3), is as important as its recognition accuracy. Again, some confusing patterns that are error-prone must be thrown out before making the final decision in order to prevent errors. Some helpful rejection criteria are therefore produced to determine and filter out the confusing samples. In the previous studies, the criteria are designed based on some heuristic ideas while the rejection processes are performed in or after the testing stage. The measurement-level outputs [32] are extracted to solve a two-class recognition problem, one of which stands for rejection and the other for non-rejection. These methods perform rejection by setting thresholds and comparing with the confidence values of a sample according to different criteria.

Considering a classifier learns to recognize specific types of samples from the training set, it is assumed that the quality of the training process affects the testing result in a large scale. In other words, the testing results are based on whether useful and recognition-sensitive information has been extracted from the training data; thus, the training phase is critical to the whole pattern recognition procedure. From this, it can be assumed that training data is as significant for pattern rejection as for recognition and we attempt to extend the rejection process from heuristic design to learning-based procedure. Compared to the traditional rejection criteria, the use of learning-based method on the training set to predict the rejection on testing samples is more straight-forward and can make use of much more information extracted from the data.

Based on the idea to extend rejection criteria designing into training process, two

novel rejection criteria are proposed, including Support Vector Machine based Measurement (SVMM) and Area Under the Curve Measurement (AUCM). SVMM uses the SVM classifier as a basic model and locates an optimal boundary between confusing and clear samples based on the training data. AUCM uses a model based on the ROC curve representing the relationship between the number of rejected samples and the reliability. It searches for the best combination of measurement-level outputs to maximize the area under the curve for rejection based on training set. Both of them are tested on the benchmark MNIST database with a CNN model to verify their effectiveness.

Besides these two learning-based rejection criteria for single classifier, a rejection method based on MCS has also been introduced. In the past several decades, MCS has contributed a lot to recognition and has achieved many outstanding results; however, it is seldom used in rejection. MCS is so effective in recognition that it is assumed to be useful in rejection as well. Therefore, we propose a weighted voting method to combine decisions from single classifiers in a MCS for rejection which will eventually be evaluated through MNIST, CENPARMI and USPS.

1.6 Thesis Outline

The main content of this thesis can be summarized in two phases: (a) learning-based rejection criteria for single classifier; and (b) voting-based rejection method with multiple classifiers. From here, the rest of the thesis will be organized as the following:

Chapter 2 will introduce the basic rejection, recognition and distortion algorithms

as well as database information used for our research. To be more specific, some background knowledge and traditional pattern rejection methods will be presented. Then, theoretical background of CNN classifier will be explained along with its two structures that have achieved high recognition rates. We will also provide the basic information about the databases that are used. At last, we will briefly study the elastic distortion algorithm that is applied in the phase of dataset re-sampling within MCS construction.

Chapter 3 will introduce two novel learning-based rejection criteria: SVMM and AUCM. Main designing ideas and architectures of these two criteria will be provided while comparisons with other traditional criteria based on MNIST handwritten digits database will be presented afterwards.

Chapter 4 will discuss the architecture and algorithm of a new rejection method with MCS. It is implemented by using voting methods to combine decisions from various single classifiers. To construct the MCS committees, two simple ways including dataset re-sampling and structure modification have been chosen. The performance of this rejection method will be tested on MNIST, CENPARMI and USPS.

Chapter 5 is a continuation of Chapter 4. In order to further improve the MCS based rejection method's efficiency, we will add specialist information of single models in various categories into the combination process. A new confidence weight parameter will be introduced with the purpose of representing the specialist capability of single classifiers. On MNIST database, the new weight parameter will be adopted

into the process of combination to evaluate its effectiveness.

Chapter 6 will draw conclusions and will illustrate the main contributions of this thesis. Also, future research directions will be presented in a brief synopsis.

Chapter 2: Theoretical Background

The concepts behind basic algorithms and rejection criteria in pattern recognition will be introduced in this chapter. Section 2.1 will look at the background knowledge of pattern rejection along with three classic criteria including First Rank Measurement (FRM), First Two Rank Measurement (FTRM), and Linear Discriminant Analysis Measurement (LDAM) [31]. Then, CNN classifier and two structures of it which have achieved high recognition rates will be discussed in Section 2.2. Section 2.3 will look at the three databases that are used for evaluation: MNIST, CENPARMI and USPS handwritten digit databases. In addition, randomly selected samples and previous extraordinary results will be displayed respectively. Section 2.4 will look at an elastic deformation algorithm that forms the basis for MCS construction in later chapters.

2.1 Rejection Criteria

Pattern rejection can be viewed as a two-class recognition problem, taking the output values of a classifier as features to recognize a pattern as a confusing one to reject or a clear one to accept. Generally, for a regular classifier, the output is always a vector consisting of confidence values or probabilities of possible classes. Given a pattern x, suppose the output vector of the classification is (c is the number of possible classes):

$$\{f_1, f_2, \dots, f_c\}, \qquad f_i \geq 0, \qquad i = 1,2, \dots, c \tag{4}$$

Then, this pattern is classified according to $x \in \text{class } \arg\max_{1 \leq i \leq c} f_i$. In case that the outputs are negative, normalization can be used to guarantee that all the values are

positive (e.g. $f_i = f_i - f_{min},\ f_{min} = \min_{1 \le i \le c} f_i$).

In the field of rejection, some traditional rejection criteria have been studied before and have produced high recognition rates as well as high reliabilities. In this section, some useful criteria are presented. The first rank confidence value (FR) and the second rank confidence value (SR) can be described as:

$$FR = \max_{1 \le i \le c} f_i, \quad SR = \max_{1 \le i \le c, f_i \ne FR} f_i \tag{5}$$

They are the most meaningful ones among all the confidence values. FR is expected to be much larger than all the other output values for a clear sample. Besides, the gap between FR and SR is also viewed as a practical factor to reflect the sample's quality. That is why First Rank Measurement (FRM) and First Two Rank Measurement (FTRM) have been proposed for rejection [31].

(1) FRM

FRM is one of the most important criteria since it takes only FR of the output vector into account. It rejects samples by setting a threshold T_1 to FR and accepts those satisfying $FR \ge T_1$.

(2) FTRM

FTRM is another important factor for rejection. Unlike FRM, it emphasizes on the gap between FR and SR. FTRM sets a threshold T_2 to the gap and accepts only the samples satisfying $FR - SR \ge T_2$.

(3) LDAM

He et al [31] propose a novel LDA measurement (LDAM), which relies on the principle of Fisher Linear Discriminant Function. The authors apply the

principle of LDA on outputs for the rejection option as a one dimensional application which shifts the Fisher criterion to:

$$J(w) = \frac{S_B}{S_W} = \frac{(\mu_1 - \mu_2)}{\Sigma_{12}} \quad (6)$$

where μ_1 and μ_2 are the centers of two classes and Σ_{12} is within-class scatter respectively. Then, they define two classes for rejecting and accepting samples: $G^{(1)} = \{\hat{f}_1\}$ and $G^{(2)} = \{\hat{f}_2, \dots, \hat{f}_c\}$, in order to maximize the separation between FR and all the other confidence values. (Here $\hat{f}_i$ are confidence values in a descending order). Thus, in LDA, $J(w)$ can be defined by:

$$J(w) = \frac{\{\sum_{i=2}^{c}(\hat{f}_1 - \hat{f}_i)\}^2}{(c-1)^2\Sigma_{12}} \quad (7)$$

where $\mu_1 = \hat{f}_1$, $\mu_2 = \frac{1}{c-1}\sum_{i=2}^{c}\hat{f}_i$, $\Sigma_1 = 0$, $\Sigma_2 = \frac{1}{c-1}\sum_{i=2}^{c}(\hat{f}_i - \mu_2)^2$ and $\Sigma_{12} = \frac{1}{2}\Sigma_2$.

A threshold T_3 is set and samples are accepted if they satisfy $J(w) \geq T_3$. The criterion has been proven to produce a better performance than FRM and FTRM based on eight-direction gradient features with SVM classifier for handwritten character recognition [31].

These three above-mentioned rejection criteria have been proven to be useful in pattern rejection [19, 31, 32]; hence, they are used for comparison with our proposed criteria in order to verify the effectiveness of the new ones.

2.2 Convolutional Neural Network (CNN)

The CNN classifier [4] is a special type of multi-layer neural network which adopts deep learning algorithm for parameter adjustment. It differs from the standard

neural network because of the function that allows automatic extraction of topological properties from the raw image. Therefore, it can work as both a feature extractor and a classifier. The feature extractor part retrieves topological features from raw images through multiple times of convolutional filtering calculation and down sampling. There are different numbers of feature maps which store the extracted features in the convolution layers. Each feature map has its own convolution coefficients and bias which are shared by all the units in this map. Each unit in the feature maps is calculated through the area at a specific spot of its previous layer, which is also known as receptive field, while performing a convolution operation with the coefficients plus the bias. Each convolution layer is followed by a sub-sampling layer including exactly the same number of feature maps to reduce their spatial resolution. The classifier part is just like traditional neural networks.

A widely used typical CNN classifier known as LeNet5 [4] is displayed in Figure 1. It takes an image of 32 by 32 pixels as an input and contains three convolution layers (C1, C3 and C5), 2 sub-sampling layers (S2 and S4) and two fully connected layers (F6 and output). C1, C3 and C5 are composed of 6, 16 and 120 feature maps respectively which are used for storing features. The sizes of feature maps in these convolution layers are 28 by 28 for C1, 14 by 14 for C3, and 1 by 1 single neuron for C5. Considering all the local receptive fields have the size of 5 by 5 pixels, all the feature maps have the size of their inputs minus 4 in both horizontal and vertical directions (2 pixels loss at each border) after convolution calculation. Sub-sampling layers are used to reduce the spatial resolution of the feature maps in convolution

layers, so they are put just after each convolution layer. Each unit of a sub-sampling layer relates to a 2*2 receptive field of its previous convolution layer. It is computed by averaging these 4 input units. That is the reason why feature maps in sub-sampling layers have the sizes of half of their inputs, as presented in Figure 1: S2 14 by 14 and S4 5 by 5. C5 and the last two layers are fully connected just like the standard neural network. The last layer has ten units for the 10 classes (0-9) in digit recognition. The neuron with the maximum value in this layer generates the final decision.

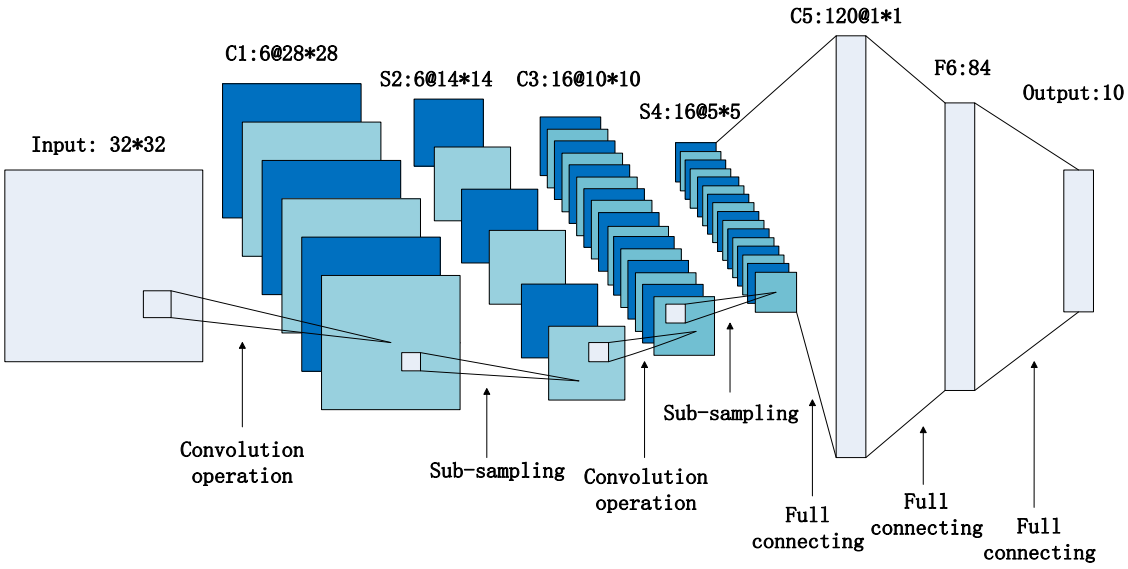

Figure 1. Structure of CNN model LeNet5 [4]

A simplified CNN architecture [30, 34] has achieved similar recognition results as LeNet5. In our research, we applied this simpler CNN model [30], which is presented in Figure 2, as a basic CNN model for our experiments. This CNN model compresses the architecture of LeNet5 to 5 layers including 1 input layer, 2 feature map layers, each conducting both convolutional filtering and down sampling tasks, and a hidden layer fully connected with the last output layer. The input is a 29 by 29 matrix with the normalized pattern centered inside. Then, the two feature map layers,

containing 25 and 50 feature maps respectively, retrieve the features by performing convolution and down sampling calculation with the receptive fields in their previous layers. After that, a hidden layer with 100 single units to store features is fully connected to the output layer. In the output layer, the final recognition decision is provided.

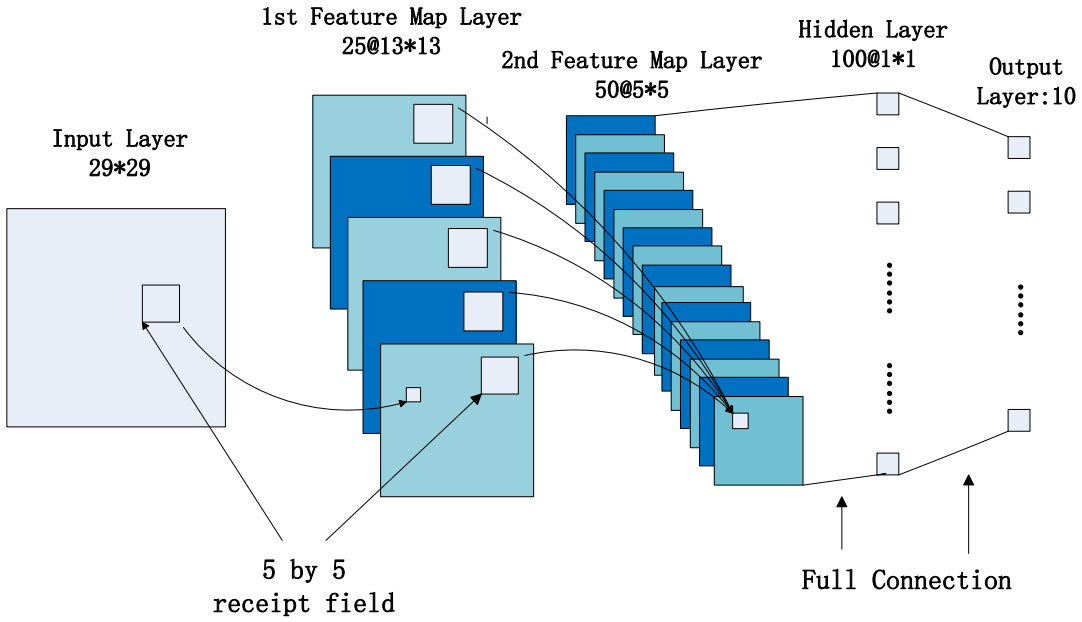

Figure 2. Structure of a simplified CNN model [30]

2.3 Description of Databases

Three famous handwritten digit datasets, including MNIST, CENPARMI and USPS, have been used for the experiments and they will be described briefly in the following section.

(1) MNIST database [4, 35]

MNIST database is a subset of well-known NIST database [7]. The training set contains 60000 binary images of handwritten digits. 30000 of them are constructed from NIST's Special Database 3 (SD-3) and the other 30000 are

from Special Database 1 (SD-1). The testing set contains 10000 patterns, 5000 from SD-3 and 5000 from SD-1. All the patterns in the training set are developed by approximately 250 writers; and, the testing sets were developed by different writers. All the samples are normalized to fix-size (20 by 20 pixels) images and centered in 28 by 28 pixels planes. MNIST is a benchmark database for handwritten digit recognition and has been widely used to evaluate classifiers' performances for over a decade [35]. Figure 3 displays some randomly selected images from the training set of MNIST, and some state-of-the-art recognition and rejection results on the MNIST isolated numerals database are listed in Table 1.

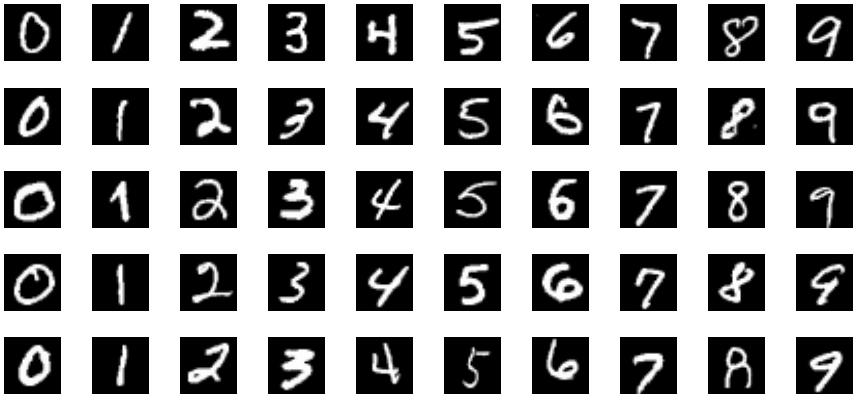

Figure 3. Image samples from MNIST handwritten digit database

Table 1. Selected testing results on MNIST database

Method	Distortion	Error (%)	Reject (%)
Boosted LeNet4 [20]	Affine, scaling, squeezing	0.70	0.0
KNN [36]	Non-linear deformation	0.52	0.0
TFE-SVM [25]	Affine	0.54	0.0
CNN [24]	Elastic	0.40	0.0
CNNs[37]	Elastic	0.39	0.0
MCDNN [28]	Width normalization	0.23	0.0
Cascaded CNNs [29]	Elastic, scaling , rotating	0.23	0.0
Hybrid CNN-SVM [30]	Elastic, scaling, rotating	0.19	0.0
Hybrid CNN-SVM [30]	Elastic, scaling, rotating	0.00	5.6

(2) CENPARMI database [5]

The CENPARMI handwritten digit database was assembled from U.S. ZIP code database of CENPARMI lab based at Concordia University. It contains approximately 17000 run-length coded binarized digits with an estimated number of 3400 writers. Samples are all unconstrained handwritten numeral images collected from dead letter envelopes, known as undeliverable mail, by the U.S. Postal Service which are then scanned in 166 PPI. In the CENPARMI database, there are 4000 images (equal number for each class of 0-9) used for training and 2000 (equal number for each class of 0-9) used for testing. All the images are of different sizes. Figure 4 displays some samples from the training set and Table 2 provides some sources of high accuracy on this database.

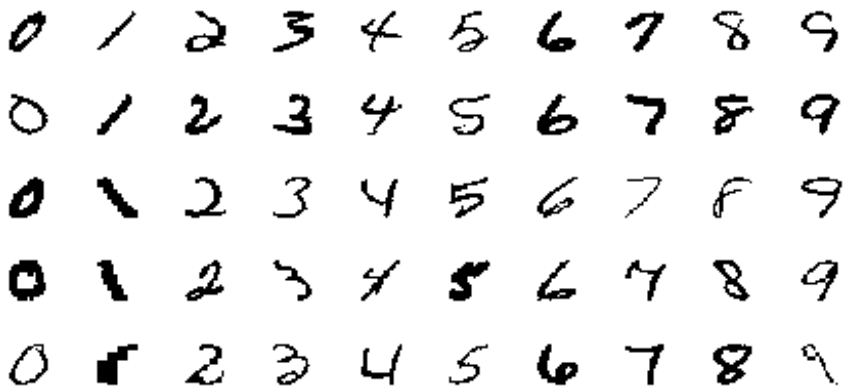

Figure 4. Image samples from CENPARMI handwritten digit database

Table 2. Selected testing results on CENPARMI database

Method	Error (%)	Reject (%)
4-expert system[5]	0.0	6.95
multiple- expert system [38]	1.15	0.0
LQDF [8]	0.95	0.0
SVC-rbf [8]	0.95	0.0
8-direction, SVC-rbf [9]	0.85	0.0
SVM, LDAM [31]	0.33	8.75

(3) USPS database [6, 39]

USPS digits data was gathered as part of a project sponsored by the United States Postal Service. Digital images found in this database included approximately 500 city names, 5000 state names, 10000 ZIP Codes, and 50000 alphanumeric characters. They were scanned from mails in a working post office at 300 PPI in 8-bit grayscale [6]. This database was traditionally used in a splitting of 7291 samples for training and 2007 samples for testing (Version 1, referred to as V1). However, these two sets were actually collected in slightly different ways and samples in the testing set were much harder to classify than the ones in the training set. Hence, it was not very suitable for demonstrating learning algorithms. From there, all the samples from both training and testing sets were gathered and reshuffled to divide anew into training and test sets, containing 4649 samples each (Version 2, referred to as V2). All the 9298 digits images of USPS handwritten digit data have a fixed size of 16 by 16 pixels. Randomly selected samples from this database are displayed in Figure 5 while Table 3 lists selected previous recognition results.

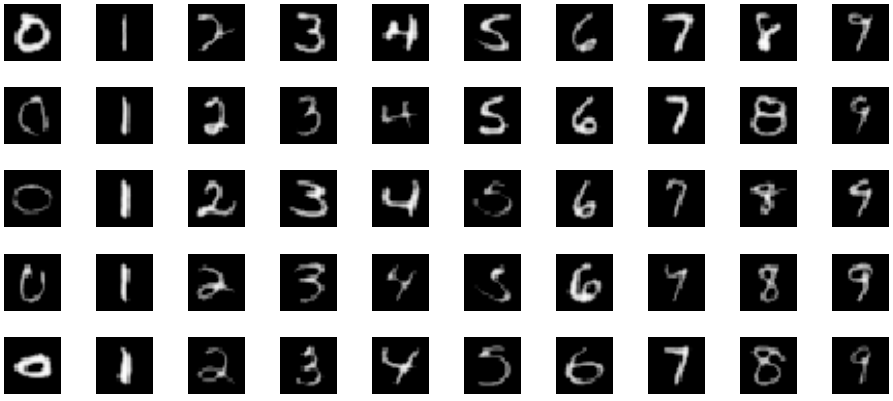

Figure 5. Image samples from USPS handwritten digit database

Table 3. Selected testing results on USPS database

Method	No. version	Error (%)	Reject (%)
Tangent Distance, 1-NN [40]*	1	2.5	0.0
Boosted Neural Network [41]*	1	2.6	0.0
LeNet1[42]	1	4.2	0.0
RVM [43]	1	5.1	0.0
GMD, VTS, TD [44]	1	2.7	0.0
KD, VTS, TD, Bagging [44]	1	2.2	0.0
SVM-rbf, e-grc3 [45]	1	2.39	0.0
SVM-rbf, e-grc3 [45]	2	1.33	0.0

*: training set extended with 2400 machine-printed digits

2.4 Distortion Methods

The CNN classifier is very powerful at classifying visual patterns, as it continues to yield state-of-the-art performances on visual analysis tasks. In order to further improve its recognition performance, especially in the cases where the numbers of training samples are small or the distributions have some transformation-invariant attributes, producing new samples to expand the datasets through transformation methods is feasible [24, 25]. Brief descriptions of two types of transformation methods are presented as follows:

(1) Affine distortion [25]

Affine distortion is a simple way to expand the dataset. It applies affine displacement fields to images in order to conduct the procedures of transformation including rotation, scaling and skewing. It is implemented by computing each pixel (x,y) displacement fields, $\Delta x(x,y)$ and $\Delta y(x,y)$, to locate the target position. The general form of affine distortion is:

$$\begin{pmatrix}\mathrm{x}\\ \mathrm{y}\end{pmatrix} = \mathrm{A}\begin{pmatrix}\mathrm{x}\\ \mathrm{y}\end{pmatrix} + \mathrm{B} + \begin{pmatrix}\mathrm{x}\\ \mathrm{y}\end{pmatrix} \tag{8}$$

where A is a 2*2 matrix and B is a column vector, storing parameters for transformation. For instance, $A = \begin{pmatrix} \alpha & 0 \\ 0 & \alpha \end{pmatrix}$ and $B = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$ are for scaling.

(2) Elastic distortion [24]

Elastic distortion is another transformation method introduced by Simard et al [24]. Within this transformation method, random displacement fields are first generated as shown in Eq. (9):

$$\Delta x(x, y) = \text{rand}(-1, +1), \Delta y(x, y) = \text{rand}(-1, +1) \quad (9)$$

where $\text{rand}(-1, +1)$ is a random number between -1 and $+1$, generated by a uniform distribution. Then, Δx and Δy are convolved with a Gaussian standard deviation σ which stands for the elastic coefficient. A small σ means more elastic distortion while a large σ makes deformation approach affine. After that, the field values are normalized and multiplied by a scaling factor α, controlling the intensity of deformation. Finally, the displacement fields are applied to each pixel of the image.

This elastic distortion method is adopted for dataset expansion for the process of MCS generation with dataset re-sampling method. Randomly selected samples from the training set are distorted with this method to generate new samples in order to form a new training set.

Chapter 3: Learning-based Rejection Criteria

Our main goal in this chapter is to improve the reliability of the single classifier by detecting error-prone samples and eliminating them from the testing process. To accomplish this, we have designed two novel rejection criteria, named SVM-based Measurement (SVMM) and Area Under the Curve Measurement (AUCM). The main difference between these two and other traditional rejection criteria is that they are learning-based criteria which extend the rejection process from heuristic design to learning procedure with training data. To evaluate the effectiveness of rejection, we can draw a Receiver Operating Characteristics (ROC) graph [46] in the coordinate system whose x-axis is the number of rejected samples and y-axis is reliability. A good rejection criterion can achieve a higher reliability with fewer samples rejected, so the curve is expected to be as close to the top left corner as possible. While the ROC curve will be introduced in Section 3.1, detailed designing ideas and architectures of SVMM and AUCM will be discussed in Sections 3.2 and 3.3 respectively. Both of these two novel rejection criteria will be compared with their traditional counterparts such as FRM, FTRM and LDAM through experiments on the MNIST database with the chosen CNN model.

3.1 Introduction of ROC analysis

A receiver operating characteristics (ROC) graph is used for visualizing, organizing and selecting classifiers based on their performance. It has a long history of usage in a variety of categories such as signal detection, visualizing and analyzing

diagnostic systems, medical decision making, etc. It is first adopted in the field of machine learning by Spackman in 1989 to evaluate and compare different algorithms [46].

ROC graphs are two-dimensional, depicting relative tradeoffs between benefits and costs. In the case of pattern rejection, there is always a tradeoff between two factors: the number of rejected samples and the reliability of the system. That is because reliability increases whenever confusing samples are rejected at early stages. In previous research of pattern rejection, reliability is always considered individually to evaluate a criterion's effectiveness. However, it is insufficient to evaluate rejection performance based on this factor exclusively since it cannot be determined which method is superior in rejection if their reliabilities are based on different numbers of rejected samples. A system with low reliability based on few rejected samples may achieve very high reliability once the rejection rate increases. As a result, these two factors are supposed to be considered simultaneously to evaluate the performances of rejection systems and that is why we introduce the ROC curve for pattern rejection.

The ROC space is a two-dimensional coordinate system whose x-axis and y-axis represent the number of rejected samples and reliability, respectively. For a rejection criterion, there is always an output value and by setting thresholds for this value, it is decided whether a sample should be rejected or accepted. If different thresholds are set and rejection procedures are conducted accordingly, we can obtain a pair consisting of the number of rejected samples and corresponding reliability for each threshold. These pairs can be presented in the created ROC space as single points and

a smooth curve joining all of them (referred to as a ROC curve) represents the performance of the rejection criterion. A good rejection criterion can achieve a higher reliability with fewer samples rejected. So, we expect a good ROC curve to be as close to the top left corner as possible. This ROC curve will be used as a tool to evaluate all the proposed rejection criteria throughout the thesis.

3.2 SVM-based Measurement (SVMM)

3.2.1 Architecture of SVMM

Pattern rejection can be viewed as a two-class recognition problem, taking the classifier's output values as features in order to recognize a pattern for rejection or acceptance. The traditional rejection criteria discussed in Section 2.2 have been designed based on some heuristic ideas. In this section, we propose a novel SVMM to extend the rejection process into a learning-based method.

Specifically, rejection can be viewed as a two-class recognition problem, one stands for rejected samples and the other for accepted ones. In SVMM, the classifier selected is SVM and the input is the output vector (always confidence values for possible classes) of a certain classifier. For a classifier, the output of a sample is a vector of confidence values $\{f_1, f_2, \dots, f_c\},\ f_i \geq 0,\ i = 1,2, \dots, c$, as mentioned before. Then, these values are used as features and sorted into a descending order:

$$\{\hat{f}_1, \hat{f}_2, \dots, \hat{f}_c\},\ \hat{f}_1 \geq \hat{f}_2 \geq \cdots \geq \hat{f}_c \tag{10}$$

The correctly and incorrectly classified samples are labeled differently: correctly classified samples are labeled with "1" while incorrectly classified ones are labeled

with "-1". This information is then used to train an SVM classifier. Linear SVM is selected for training in order to locate the rejection boundary. Therefore, the decision boundary is a linear function combining all the components of the output vector, represented in Eq. (11), where $\{w_i\}_{i=0}^{c}$ are the coefficients of SVM:

$$T = \sum_{i=1}^{c} w_i \hat{f}_i + w_0 \tag{11}$$

The reason for choosing a linear kernel for SVM rather than a nonlinear one, such as RBF kernel, is based on the following reasons:

(1) A linear kernel works very fast in training and testing, and an optimal linear separating boundary is a good way to avoid over-fitting.

(2) A linear boundary is more meaningful physically and function Eq.(11) includes some special cases in it. For instance, FRM can be viewed as a linear boundary with $w_1 = 1$ and $w_2 = w_3 = \cdots = w_c = w_0 = 0$; while FTRM can be viewed as: $w_1 = 1$, $w_2 = -1$ and $w_3 = w_4 = \cdots = w_c = w_0 = 0$.

Note that in the training process of SVMM, the number of samples in class "1" is always much larger than that of class "-1", because the baseline accuracy of the classifier is high. In this case, the problem is an unbalanced classification problem. To solve this problem, we use different weighting functions for different classes in the "libsvm" software [47]. In the testing process, the same features are extracted and sorted in descending order, and a sample is rejected if the calculated T in Eq. (11) for it is smaller than a pre-defined threshold. Figure 6 is a flow chart depicting the whole rejection process:

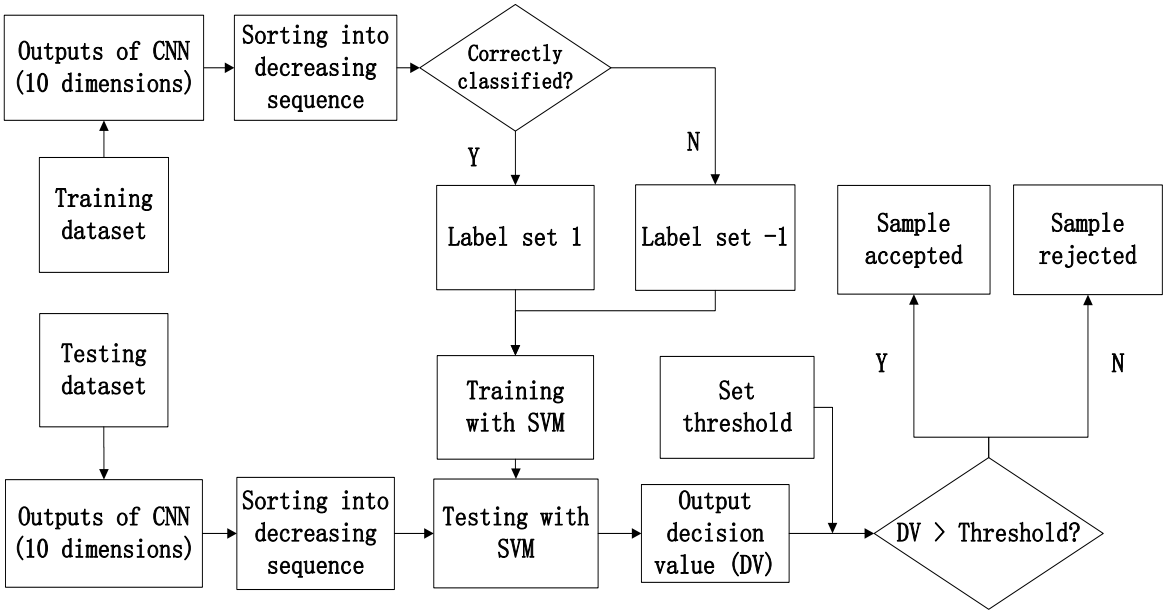

Figure 6. Flow chart of SVM-based Measurement (SVMM)

With this new criterion, the linear rejection boundary is found by training an SVM with the training set. The main difference between SVMM and other criteria, like FRM, FTRM and LDAM, is that SVMM extends the rejection process from heuristic design to a learning-based procedure. Using the learning-based method with training set to predict the rejection decision on testing samples is more straight-forward and allows researchers to use more information from the data.

3.2.2 Experiment with SVMM

In the selected CNN model presented in Section 2.2 (referred to as "M0"), the output of each sample is a 10-dimensional vector consisting of confidence values for the 10 possible classes. FRM, FTRM and LDAM are used respectively as rejection criteria with this basic model. Thresholds are searched incrementally. As in CNN model, the outputs are confidence values instead of probabilities, the most appropriate starting point, step and ending point for thresholds searching vary according to

different rejection criteria. The search steps for them are all 0.1 at regular intervals and 0.01 at the sections where the number of rejected samples changes sharply.

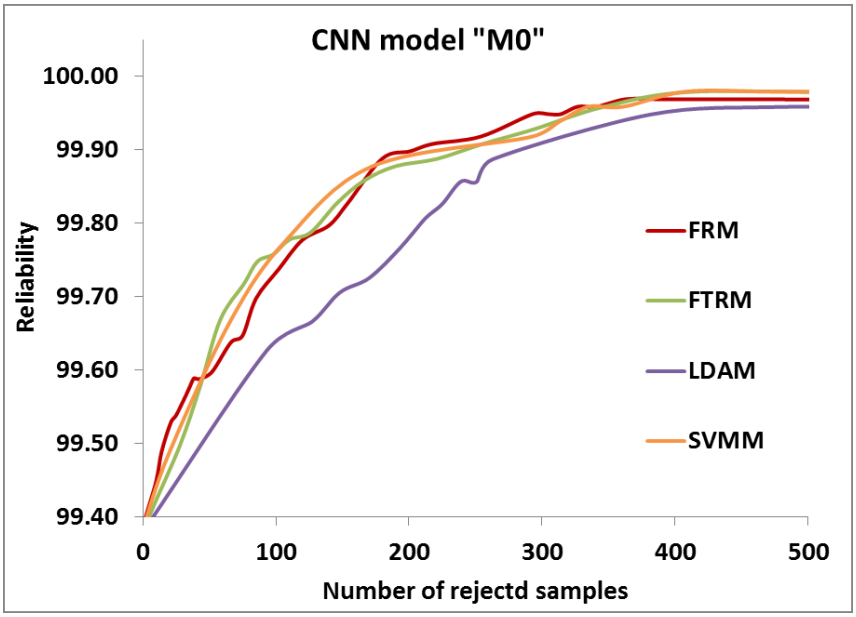

Figure 7. ROC curves of SVMM and other rejection criteria with classifier "M0"

For the newly proposed SVMM, "libsvm" tools are applied and the same CNN model "M0" is used as a feature extractor. Out of 60000, there are 216 samples labeled “-1” while the rest are labeled “1” for the training process. Since the training set is relatively unbalanced with the number of samples in class "1" at almost 300 times that of class "-1", the weight parameter is set to "400" for class "-1". A linear kernel is selected in order to find a linear decision boundary in the feature space. Normalization is conducted on the decision value with SVM of each sample with the purpose of making the threshold-setting procedure more convenient. Then, different thresholds are set for rejection. Since the output is a normalized value, the starting and ending points for threshold searching are 0 and 1 respectively while the search steps are 0.1 at regular places and 0.01 at the sections where the number of rejected samples fluctuates sharply. All the results are shown by the ROC curves presenting the

relationship between the number of rejected samples and reliability in Figure 7.

3.2.3 Comparison with other Rejection Criteria

Although LDAM is proven to have a better performance than FRM and FTRM in He et al's research [31] based on eight-direction gradient features with an SVM classifier; the results demonstrate that LDAM is the least useful one in our experimentations with the CNN model. As for FRM and FTRM from He's work, their performances varied and yet, they are very similar when applied to the CNN model "M0". Therefore, it can be concluded that these pre-defined criteria vary in performance with different classifier models or types of features. In Figure 8, some randomly selected samples from the training set are displayed in a 2-dimensional coordinate system based on their first two rank confidence values (FR and SR).

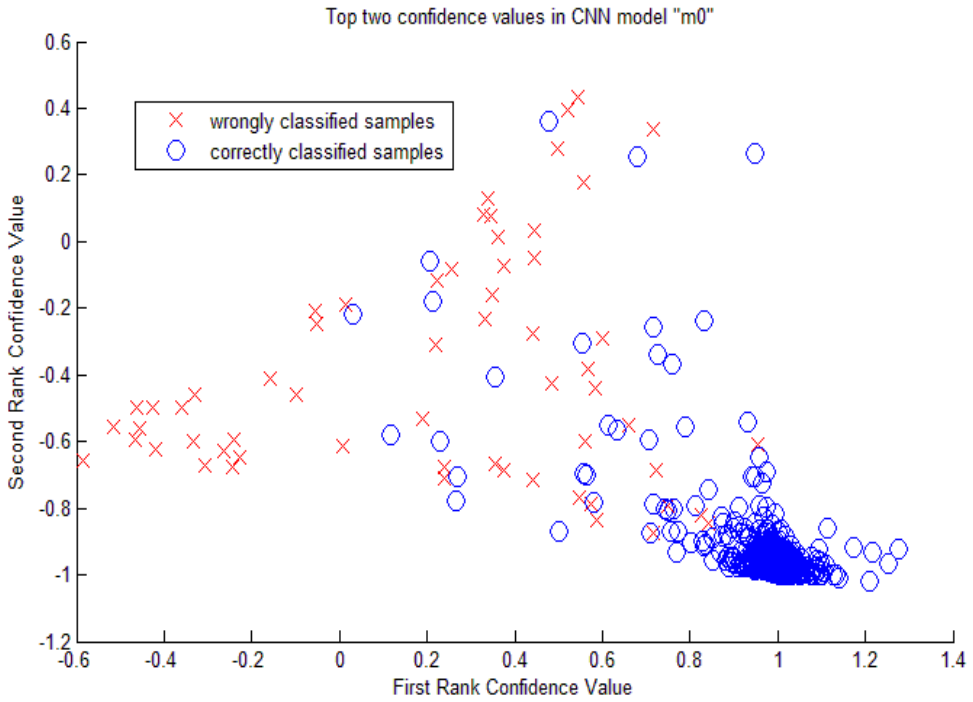

Figure 8. Samples in FR-SR feature space

We can see that FR and SR of correctly classified samples are extremely close to

1 and -1 respectively. As a result, a line with slope "1" standing for FTRM in the coordinate system is an optimal boundary to separate wrongly and correctly classified samples. That is why FTRM is an effective criterion in this model. Another effective criterion, FRM, can also be viewed as a way of finding a boundary parallel to the y-axis, which is less effective than FTRM through observations. However, it is noticed that although these two criteria can be useful, many correctly classified samples will also be rejected by them no matter where the boundary is.

It is also shown in Figure 7 that SVMM works as effectively as FTRM in rejection and their performances are too close to determine which one is better. The same work has been completed with two other CNN models whose structures are similar to "M0". These CNN models produced only small changes in the number of maps in feature map layers. The results of these two models are displayed in Figure 9. It is apparent that SVMM and FTRM are always the relatively best ones among all of the rejection criteria. The reason behind the performances can be traced back to the training process of CNN model where the expected values in the decision layer are set to be "1" for the true class and "-1" for the other classes. Hence, FTRM is already a distinctively effective criterion to determine the quality of a sample as analyzed with Figure 8. When we use the SVMM, which uses all the values of the output vector, FR and SR contribute much more than the other eight confidence values since the others are slightly different from SR. Therefore, the rejection boundary of SVMM is very close to that of FTRM, explaining why these criteria display similar performances. In addition, despite the presence of a weight parameter for the class of rejection in

SVM training, the unbalanced data remains a critical factor affecting SVMM's overall performance.

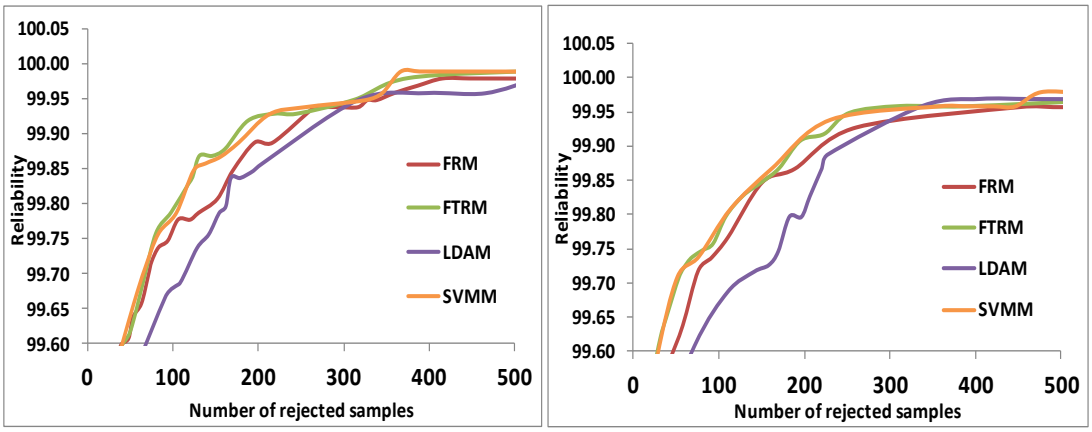

Figure 9. ROC curves of different rejection criteria with other CNN models

3.3 Area Under the Curve Measurement (AUCM)

3.3.1 Algorithm of AUCM

AUC, the name given to the novel rejection criterion, is the abbreviation of the expression: "area under the curve". It is mentioned previously in Section 3.1 that in order to evaluate the effectiveness of a rejection criterion, we can draw a ROC curve in the coordinate system whose x-axis represents the number of rejected samples whereas y-axis represents the reliability. A good rejection criterion can achieve a higher reliability with fewer samples rejected, hence we expect a good curve to be as close to the top left corner as possible. In other words, we expect a good rejection criterion to make the area under this ROC curve to be as large as possible. To accomplish this goal, we attempt to determine a linear combination of FR and SR as a rejection criterion based on all training samples, because FR and SR are the most meaningful ones among all these confidence values.

Firstly, we create a linear combination of FR and SR, as followed in Eq. (12):

$$T = w_1 * FR + w_2 * SR \quad (12)$$

where $FR = \text{First Rank Confidence Value}$, $SR = \text{Second Rank Confidence Value}$, and w_1 and w_2 are two parameters that will be derived from the training data to maximize the area under the ROC curve. Specifically, we simply fix the value of w_1 at "1" and search different values for w_2. For each w_2, there is a combination where T is the outcome. Pairs of number of rejected samples and reliability are calculated individually based on different thresholds of T and displayed as single points in the ROC space. Then, a ROC curve is formed by connecting all the single points smoothly.

In order to compute the area under this curve, we approximate it by the sum of hundreds of small trapezoids as shown in Figure 10. The segmentation of the small trapezoids is based on the thresholds. For each combination of T, rejection decision values of all the training samples can be calculated in order to find out the maximum value and the minimum value. Subsequently, the space between these two is divided equally into 200 parts, each of which is set as a threshold incrementally and used to generate a small trapezoid. The two parallel sides are the reliability values with the current threshold and its previous one. The height is the absolute difference between the number of rejected samples based on the current threshold and its previous one. Then, the area of the trapezoid for each threshold is calculated and the area under the curve can be computed accordingly by summing all of the small trapezoids. The areas under the curve of different Ts are compared to find out the maximum one in order to

obtain the best w_2.

In the testing process, with the optimal w_2, the responding combination T is adopted as rejection criterion for the testing samples. A sample is rejected if its rejection value of T is smaller than a pre-defined threshold.

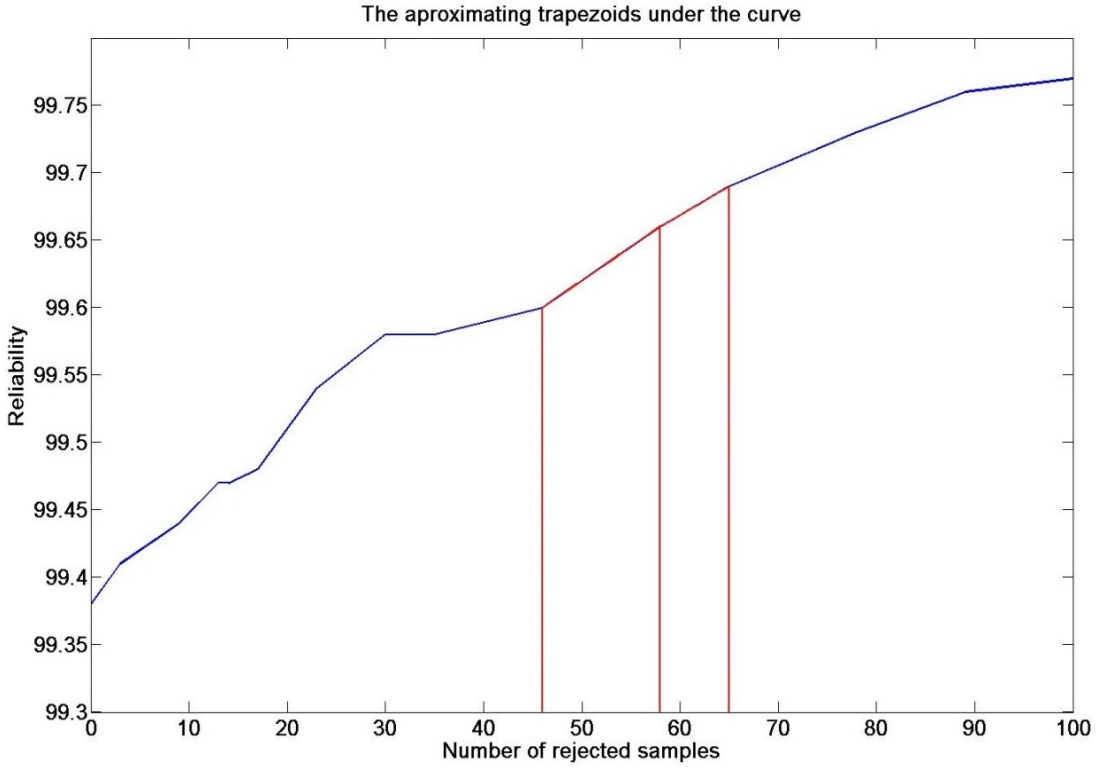

Figure 10. The approximating trapezoids under the curve

3.3.2 Experiment with AUCM

This AUCM rejection criterion is also evaluated with CNN model "M0" on MNIST database. As mentioned in Section 3.3.1, w_1 is fixed at "1". The w_2 is searched from "-5.0" with an incremental step of "0.05" until "5.0". For each pair, the area under the curve is calculated through the approximation of the sum of hundreds of small trapezoids under it, as discussed in Section 3.3.1. The optimal w_2 searched out to maximize the area under the ROC curve is "-1.75" in our experiment. So, the rejection criterion T is determined to be:

$$T = FR - 1.75SR \tag{13}$$

A sample is rejected if its rejection decision value T in Eq. (13) is smaller than a threshold.

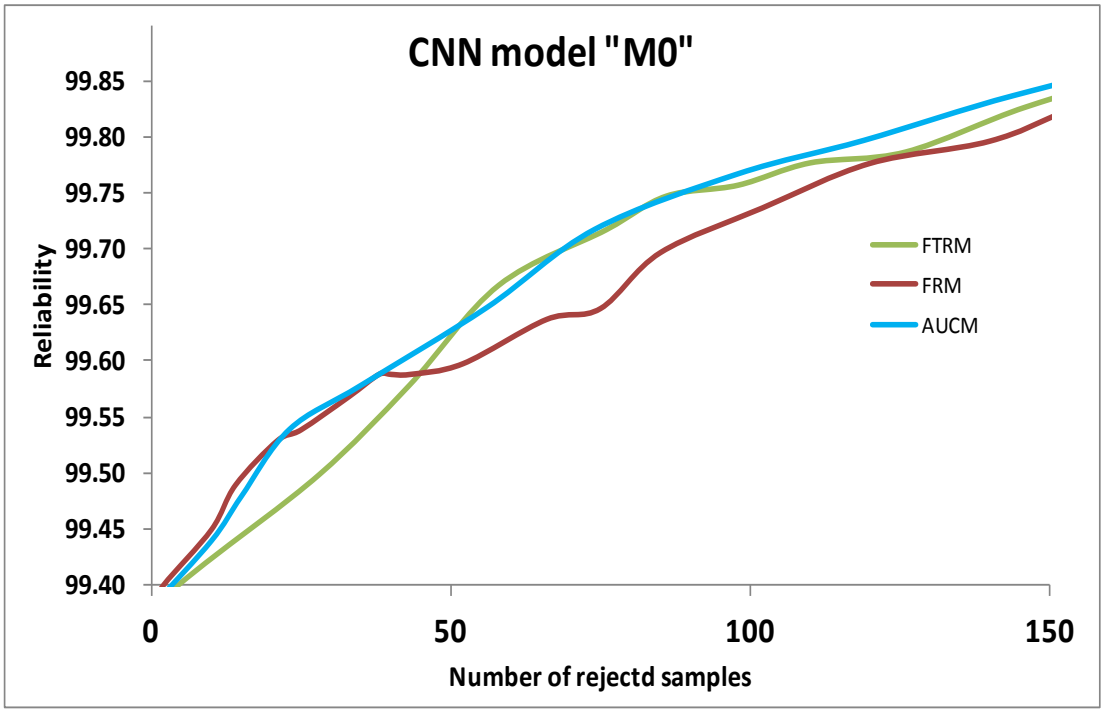

Figure 11. ROC curves of AUCM and other rejection criteria with classifier "M0"

In the testing process, because the range of the output values varies according to the different combination Ts, the starting and ending points for threshold setting remain unstable. So, we first look for the maximum and minimum values for a specific T to determine the starting and ending points. In this experiment, the starting point is -0.4 and the ending point is 2.7. The search steps for the threshold are still 0.1 at regular places and 0.01 at the sections where the number of rejected samples fluctuates sharply. The rejection result of AUCM is shown in Figure.11 with those of FRM and FTRM, as ROC curves illustrating the relationships between the number of rejected samples and reliability.

3.3.3 Comparison with other Rejection Criteria

It is clearly indicated from Figure 11 that the AUCM achieves a higher performance than the other two criteria, because its ROC curve is closer to the left-top corner and remains higher than those of FTRM and FRM in almost its entire path. It means that with AUCM, we can always obtain a higher system reliability when compared with FTRM and FRM based on the same number of patterns rejected. It proves the effectiveness of this new rejection criterion.

The advantage of AUCM can be explained from its designing idea of finding the optimal combination of FR and SR as a rejection criterion from the training data. This information is then interpreted as finding the best combination to maximize the area under the ROC curve, which is applied to evaluate the rejection criterion's performance. In the parameters searching process, the coefficient of FR (w_1) is fixed at "1" and that of SR (w_2) varies between "-5" and "5", which includes the FRM and FTRM as special cases of the combinations ($w_1 = 1, w_2 = 0$ for FRM and $w_1 = 1, w_2 = -1$ for FTRM). Therefore, the combination with the optimal w_1 and w_2 pair works more effectively than FRM and FTRM based on the training set, since its area under the curve is the largest among all including those of FRM and FTRM. Generally, it is assumed that the training and testing sets are closely related; hence demonstrating that the optimal combination on the training set is supposed to achieve a better performance on testing set. Later, this assumption is proven by the experimental result that the optimal combination of FR and SR on the training set works more effectively than other criteria on the testing set as well.

Chapter 4: Rejection with MCS

In this chapter, the Multiple Classifier System (MCS) will be studied for the purpose of pattern rejection which is implemented by using voting methods to combine decisions from different single classifiers. The CNN classifier "M0" [30] will still be used as a basic model. To construct the committees of a MCS, two simple methods including dataset re-sampling (DR) and structure modification (SM) are chosen. The details on how the MCSs are constructed will be described in Section 4.1. Section 4.2 will provide the proposed voting-based combination methods' algorithms. Both hard voting and soft voting will be considered. In Section 4.3, the new MCS based rejection method will be evaluated on three notable handwritten digit databases: MNIST, CENPARMI and USPS. All the experimental results and analyses will also be displayed in this section.

4.1 Construction of MCS

Re-sampling the dataset (with Bagging [48], Boosting [49] and so forth) and changing the classifier (in structure or type [50]) are two main ways to produce committees of MCSs. Many researchers have used these methods to produce a group of classifiers and applied certain combination methods for recognition. On the other hand, CNN classifier, especially MCS based on CNN, works extremely effectively in handwritten character recognition [28, 29, 30]. Therefore, the CNN model "M0" is selected as the core classifier and both of the two methods, dataset re-sampling (DR)

and structure modification (SM), are adopted to build MCSs according to our strategy.

As seen in Chapter 2, our CNN model "M0" has 2 feature map layers and 1 hidden layer with 25, 50 and 100 feature maps respectively to store the features after convolutional filtering and they are named as F1, F2 and F3. Two types of modifications have been explored: one is by adding or subtracting the numbers of feature maps in each of the three feature map layers; the other is using "Bagging" method, such as dataset re-sampling, to randomly select samples from the training sets to train the same CNN model numerous times.

(1) MNIST

For the MNIST database, SM method is initially applied to build committees. We alter the model's structure slightly by increasing and decreasing the number in each feature map layer. Specifically, there are six modified structures ("M1" to "M6"), as shown in Table 4 below. In order to diversify recognition results, we change the number of feature maps in one of these layers and keep the rest intact every time. In M1 and M2, we merely change F1, in M3 and M4, we change F2, and in M5 and M6, F3. Then, all of the models are trained to the 500th epoch until the recognition rates of the training set remain stable. All the error rates, generated from the testing loops, are listed in Table 4.

After that, the model structure is fixed at "M0" and DR is adopted to generate the committees. It is noted that 30000 samples, which represent half of the samples in the training set, are randomly extracted for each committee. The elastic distortion algorithm [24] is then implemented to produce 30000 new

samples with parameters $\sigma = 10$ and $\alpha = 1$. Some samples as well as their distorted counterparts are presented in Figure 12. Afterwards, these two groups of samples are merged to form the new training set with 60000 samples. This procedure is repeated five times to create five distinct training datasets while the "M0" is trained on them respectively to build a MCS with five committees ("G1" to "G5"). The information of each re-sampled training set is listed in Table 5 along with their recognition error rates based on the MNIST testing dataset at the 300th epoch of training when the recognition rates achieve stability.

Table 4. Information about SM in MCS on MNIST database

	M0	M1	M2	M3	M4	M5	M6
F1	25	25	25	25	25	10	40
F2	50	50	50	30	80	50	50
F3	100	80	120	100	100	100	100
Training Error Rate (%)	0.36	0.34	0.31	0.34	0.26	0.34	0.29
Testing Error Rate (%)	0.62	0.63	0.61	0.60	0.58	0.63	0.61

Table 5. Information about DR in MCS on MNIST database

	G1	G2	G3	G4	G5
0	2938	2936	2945	2940	3009
1	3467	3412	3399	3339	3420
2	3008	2936	3026	2953	2939
3	2959	3083	3055	3105	3028
4	2895	2866	2850	2996	2803
5	2672	2745	2700	2676	2788
6	2990	2982	2946	2996	3031
7	3144	3076	3165	3060	3137
8	2906	2965	2992	2987	2954
9	3021	2999	2922	2948	2891
Training Error Rate (%) of re-sampled dataset	0.79	1.10	1.11	1.43	1.10
Training Error Rate (%) of original dataset	0.68	0.76	0.72	0.83	0.74
Testing Error Rate (%)	0.60	0.73	0.75	0.79	0.71

Figure 12. Samples from MNIST database and their distorted counterparts

(2) CENPARMI

For the CENPARMI database, we start by increasing the numbers of feature maps in each feature map layer (F1, F2 and F2) of the "M0" while training all the models to the 150th epoch when the error rates remain stable, as shown in Table 6, to construct the MCS.

Table 6. Information about SM in MCS on CENPARMI database

	M0 (basis)	M1	M2	M3
F1	25	50	50	70
F2	50	75	90	75
F3	100	120	100	100
Training Error Rate (%)	0.50	0.38	0.38	0.43
Testing Error Rate (%)	2.45	2.45	2.25	2.45

Table 7. Information about DR training sets on CENPARMI database

	G1	G2	G3	G4
0	474	450	458	402
1	462	408	482	440
2	416	358	408	380
3	350	404	340	390
4	332	394	372	430
5	394	382	410	426
6	380	424	392	370
7	370	424	426	412
8	400	396	386	350
9	422	360	326	400
Training Error Rate (%) of re-sampled dataset	1.65	1.52	1.27	1.77
Training Error Rate (%) of original dataset	1.42	1.78	1.9	1.4
Testing Error Rate (%)	2.80	3.65	3.5	3.45

DR method is then used to build the MCS. During this phase, model structure is fixed as the basic one. Different training sets are formed by randomly selecting 2000 training samples and distorting them with elastic distortion algorithm [24] using the same parameters as in MNIST ($\sigma = 10$ and $\alpha = 1$). The process is repeated four times to obtain four different training sets (G1-G4) with 4000 samples each, as seen in Table 7 alongside with recognition results on the testing set after 150 epochs when the error rates get stable.

(3) USPS

For the USPS database, there are two versions including one with 7291 training samples and 2007 testing samples (referred to as V1) and the other version with 4649 samples for each of the two sets (referred to as V2).

Firstly, we increase the amount of feature maps in each feature map layer (F1, F2 and F3) of the "M0" to build “M1” to “M3”. All the models are trained for 300 epochs until the recognition rates on training set achieve stability. This work is completed for both of the two versions of USPS database and the results are shown in Table 8.

Table 8. Information about SM in MCS on USPS database

	M0	M1	M2	M3
F1	25	40	25	25
F2	50	50	80	50
F3	100	100	100	120
Training Error Rate (%) on V1	2.15	2.13	2.08	2.07
Testing Error Rate (%) on V1	3.84	4.04	3.89	3.99
Training Error Rate (%) on V2	2.41	2.54	2.58	2.47
Testing Error Rate (%) on V2	4.45	4.47	4.58	4.39

Table 9 (a). Information about DR training sets on USPS database (V1)

	G1	G2	G3	G4
0	600	621	606	617
1	487	460	505	494
2	378	373	371	351
3	324	299	327	332
4	347	326	355	342
5	276	295	267	269
6	321	311	310	311
7	311	311	303	321
8	291	302	273	284
9	310	347	328	324
Training Error Rate of re-sampled dataset on V1(%)	5.47	4.75	4.66	4.65
Training Error Rate of re-sampled dataset onV1 (%)	3.74	3.32	3.51	3.51
Testing Error Rate on V1 (%)	4.63	4.93	4.83	4.98

Table 9 (b). Information about DR training sets on USPS database (V2)

	G1	G2	G3	G4
0	365	380	393	398
1	301	295	275	290
2	240	220	254	248
3	194	231	219	195
4	229	187	197	209
5	193	164	164	175
6	215	219	220	197
7	183	207	212	192
8	196	175	180	201
9	209	247	211	220
Training Error Rate of re-sampled dataset on V2(%)	5.27	5.16	5.16	5.87
Training Error Rate of re-sampled dataset onV2 (%)	4.26	4.56	4.43	4.52
Testing Error Rate on V2 (%)	5.7	5.61	5.96	6

Secondly, DR method is adopted to build committees on USPS data. Also, the model structure is fixed at "M0". For V1, 3645 training samples, which consist of approximately half of the training set, are selected randomly and

distorted with the same elastic distortion algorithm and parameters. Then, all the selected samples and the distorted ones are mixed to form the new training set with 7290 samples. For V2, 2325 training samples, which contain about half of the training set, are selected and distorted to generate the new training set containing 4650 samples. The same job has been performed four times to obtain four different training sets for each of the two versions. The basic CNN model is trained for 300 epochs on specific datasets until training recognition rates become constant. All the information is provided in Tables 9 (a) and (b).

4.2 Pattern Rejection with MCS based on Voting

Despite MCS' effectiveness and contribution to the recognition field, it is seldom associated with pattern rejection, another important branch in pattern recognition. Therefore, there will be an attempt to adopt MCS for a rejection problem. Voting has always been seen as a good choice for the purpose of combining multiple classifiers due to its simplicity and efficiency. While hard voting is the simplest voting method which assigns equal weight to all votes, soft voting assigns a weight to each classifier according to the classifier's performance and will possibly produce more accurate and reliable results [26, 27].

4.2.1 Hard Voting for Rejection

In this section, hard voting is considered as a combination method for MCS rejection and the algorithm is followed in Section 4.2.1.1.

4.2.1.1 Algorithm of Hard Voting for MCS Rejection

Suppose there are N different classifiers in the MCS, denoted as $g_1, g_2, \ldots, g_N$. For a pattern, each of them would give a prediction of the label, denoted as $y_1, y_2, \ldots, y_N$. Once a class is predicted, it obtains one vote and the outcome would be a voting value $V_j (j = 1,2, \ldots, c)$ for each of the possible classes. Then, thresholds T_{com} are set for the top voting value $V_{max} (V_{max} = \max_{1 \le j \le c} V_j)$ and only the samples satisfying $V_{max} \geq T_{com}$ are accepted while the others are rejected. The threshold T_{com} can be set to an integer satisfying $N/2 \leq T_{com} \leq N$ to make sure at least half of all classifiers vote for the same class.

4.2.1.2 Experiment with Hard Voting for Rejection

The experimentation with hard voting combination method for rejection is conducted with the MCS constructed by SM method on MNIST dataset. Again, the information about MCS is available in Section 4.1, Table 4.

Seven classifiers are chosen for this MCS. Thresholds for rejection are set to several integers (7, 6, 5 and 4) to ensure that at least half of all models provide the same prediction to accept a sample. The recognition and rejection information with different thresholds is listed in Table 10.

Table 10. MCS rejection based on hard voting method

Threshold	No. non-rejected samples	No. rejected samples	No. correct samples	Rejection rate (%)	Reliability (%)
7 (All)	9882	118	9868	1.18	99.86
6	9882	118	9868	1.18	99.86
5	9928	72	9907	0.72	99.79
4	9962	38	9930	0.38	99.68

In hard voting, having a range limitation makes the rejection process inflexible since the thresholds can only be set to limited values. So, once the maximum value, which equals the number of classifiers in the MCS, is reached, the reliability cannot be improved any more. Also, this method cannot yield an ROC curve in the ROC space. The highest reliability is 99.86% with 118 samples rejected when the threshold is set to "7". In order to solve this problem, soft voting method will be used.

4.2.2 Soft Voting for Rejection

The soft voting process is quite similar to hard voting, except unequal weights are considered for different classifiers. Compared to the simple majority voting method, a weighted soft voting can produce more accurate and reliable results [26].

4.2.2.1 Algorithm of Soft Voting for MCS Rejection

In order to improve the rejection performance of hard voting combination method in MCS, an attempt of soft voting method is performed. For the weights part, all the rejection criteria mentioned in Chapter 3 can be selected for the reason that they reflect the rejection performances of single classifiers. A certain type of rejection criterion is assigned to each model as weight in the voting procedure, and the class label with the highest voting value provides the final decision for each sample.

As mentioned in Section 4.2.1, suppose there are N different classifiers in the MCS, denoted as $g_1, g_2, \dots, g_N$. In this case, for a random pattern, each classifier g_i $(i = 1,2, \dots, N)$ would provide a prediction of the label y_i as well as an output vector $\{f_1^i, f_2^i, \dots, f_c^i\}$. Then, for each classifier, the selected rejection

measurement (FRM, FTRM, SVMM and so forth) can be calculated based on the output vector, denoted as $t_i(i = 1,2, \dots, N)$. After that, soft voting is performed and a voting value $V_j(j = 1,2, \dots, c)$ is calculated for each of the classes as Eq. (14).

$$V_j = \sum_{i=1}^{N} t_i I(y_i, j), \quad I(y_i, j) = \begin{cases} 1 & \text{if } y_i = j \\ 0 & \text{else} \end{cases} \tag{14}$$

Among V_j, a maximum voting value $V_{max} = \max_{1 \le j \le c} V_j$ can be found and a threshold T_{com} is searched and determined. A pattern is rejected if V_{max} is smaller than a threshold. As the voting values are sums of all models, the thresholds T_{com} can be any real number between 0 and N. The whole procedure of MCS based pattern rejection is shown in Figure 13.

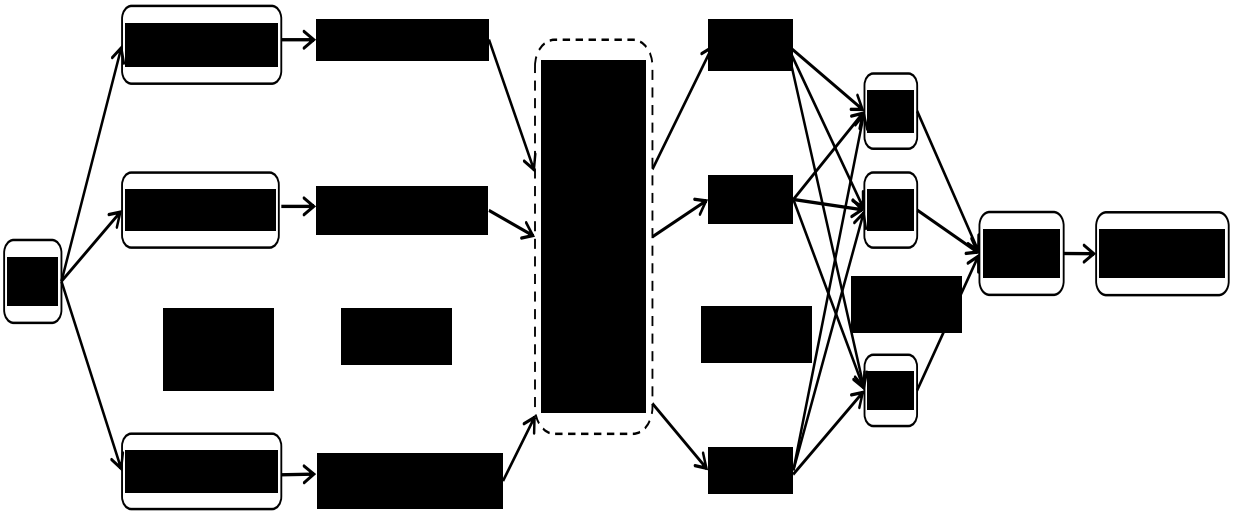

Figure 13. Flow chart of voting based combination of MCS for pattern rejection

4.2.2.2 Experiment with Soft Voting for Rejection

The experimental procedures of applying soft voting combination to the MCSs for rejection will be described in this section. Experiments are conducted on the three handwritten databases mentioned in Chapter 2. The MCSs have been constructed in different ways on each database and trained sufficiently as displayed in Tables 4~9.

For MNIST, three rejection criteria, including FRM, FTRM and SVMM, are chosen as weight parameters for combination. For both CENPARMI and USPS databases, one pre-designed criterion: FTRM and one learning-based criterion: SVMM are selected as weight parameters. Details will be presented below:

(1) MNIST

The proposed soft voting based combination method has been applied to MCSs constructed by both of SM and DR on MNIST database, as described in Section 4.1 (Tables 4 and 5). Firstly, the experiment is conducted with MCS built by SM. Two pre-designed rejection criteria: FRM, FTRM and one learning-based criterion: SVM are adopted as weights for combination. Since these criteria have different value ranges, different starting points, searching steps and ending points are chosen specifically for them. For SVMM, the starting and ending points are 0 and 1 respectively, because the decision values of SVMM are normalized. For FRM, the starting and ending points are -0.5 and 1 respectively, since the decision values of it is the first rank confidence value given by the CNN classifier, which can be a negative number; then for FTRM, the starting and ending points are 0 and 2. The searching steps for all of them are 0.1 at regular places and 0.01 at the segments where the number of rejected samples increases sharply based on different criteria. The results with different criteria are shown in Figures 14 (a~c) as ROC curves representing the relationship between the number of rejected samples and reliability.

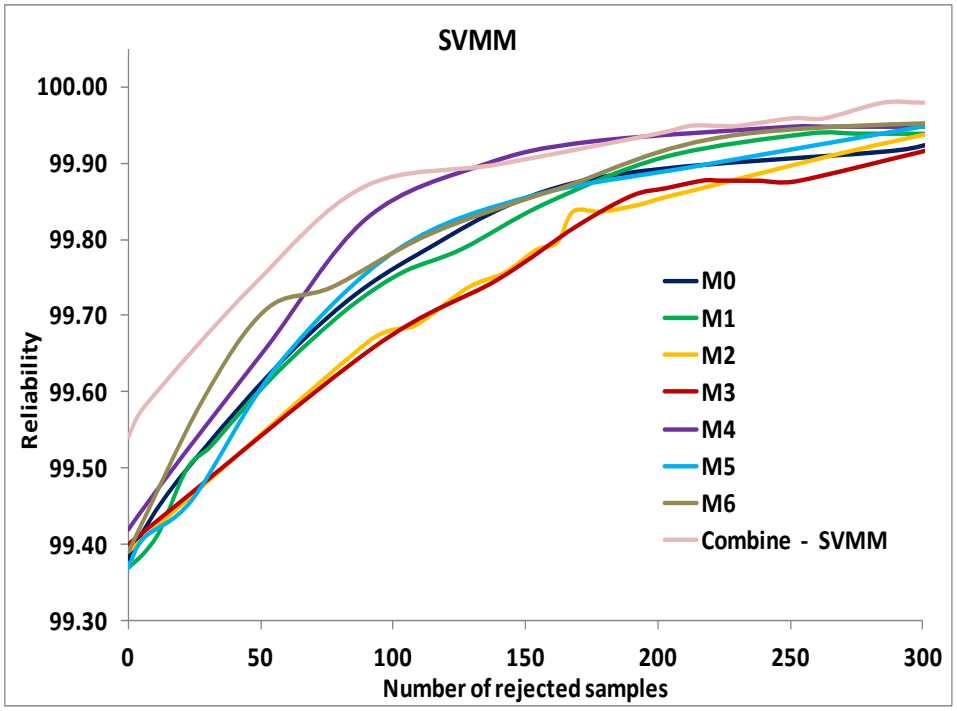

Figure 14 (a) ROC curves of MCS (SM) and single models with SVMM on MNIST database

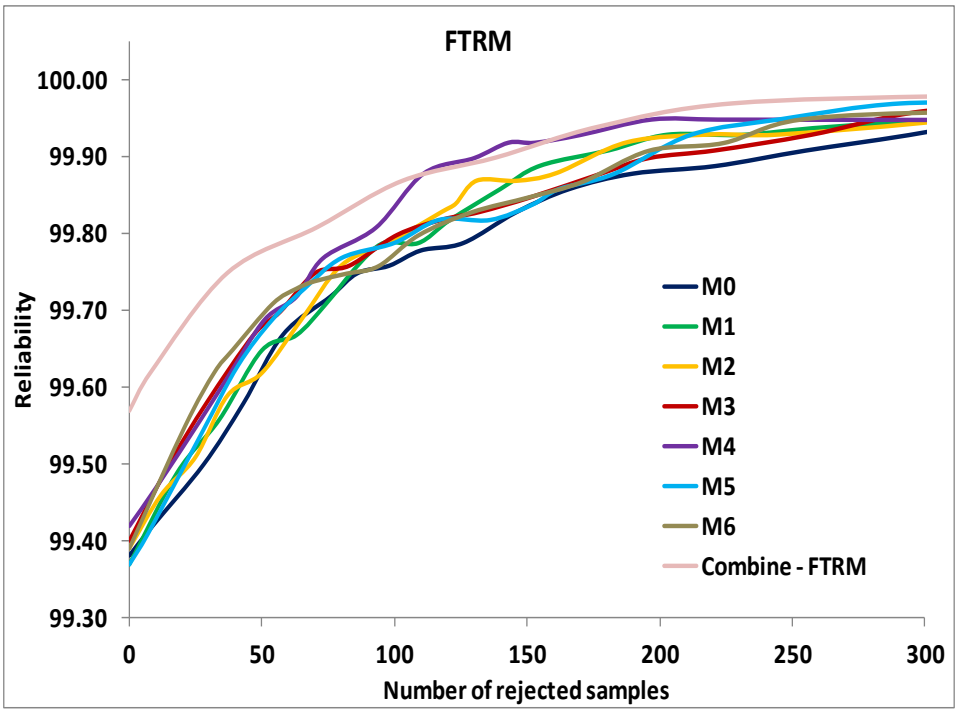

Figure 14 (b) ROC curves of MCS (SM) and single models with FTRM on MNIST database

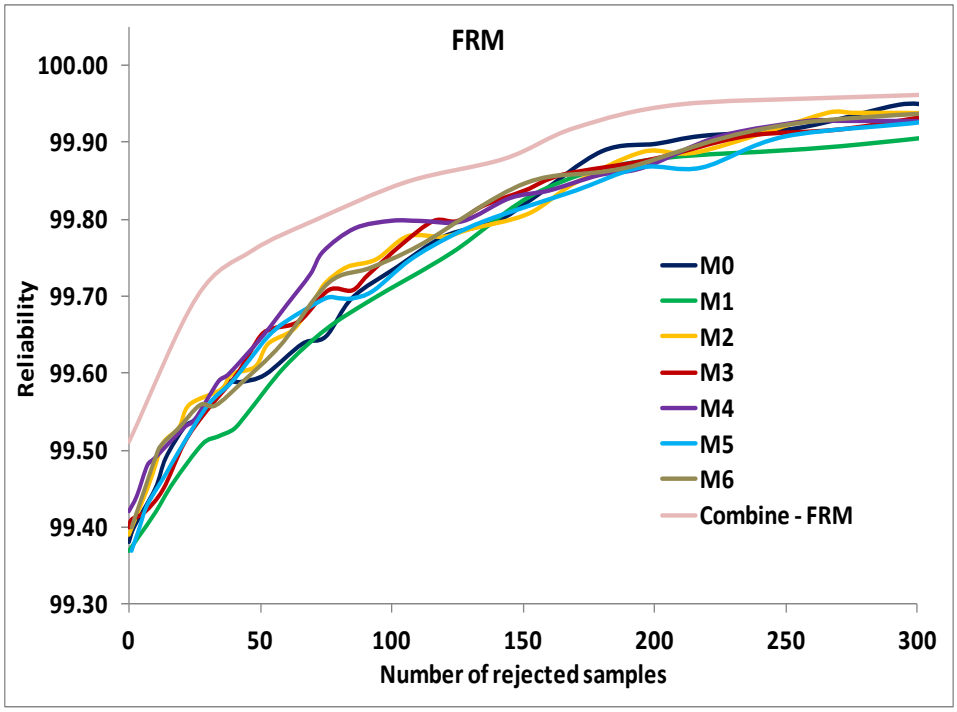

Figure 14 (c) ROC curves of MCS (SM) and single models with FRM on MNIST database

Figures 14 (a~c) demonstrate that the rejection performances are consistently improved for all rejection criteria (FTM, FTRM and SVMM) with the combination of seven CNN models. In addition, by applying combination to the single classifiers, the recognition performance without rejection (0 point of x-axis) is also enhanced for all three criteria. The error rates are decreased for about 25%, from about 0.6% to 0.45%, for both FTRM and SVMM.

The same experiment with MCS built by DR is performed. All of the results with different criteria are shown in Figures 15 (a~c).

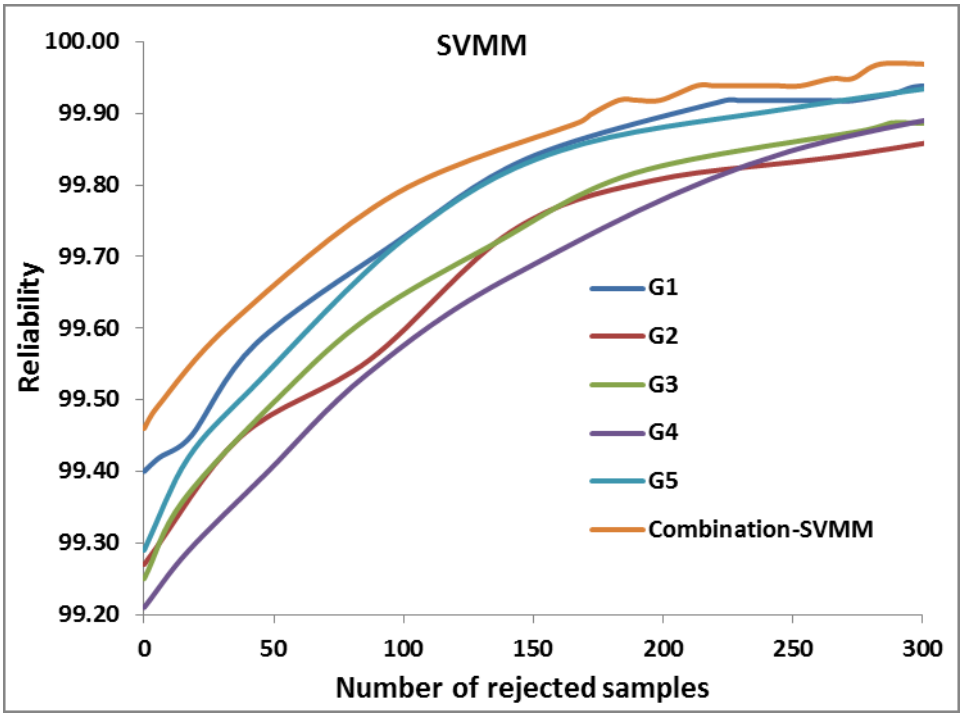

Figure 15 (a) ROC curves of MCS (DR) and single models with SVMM on MNIST database

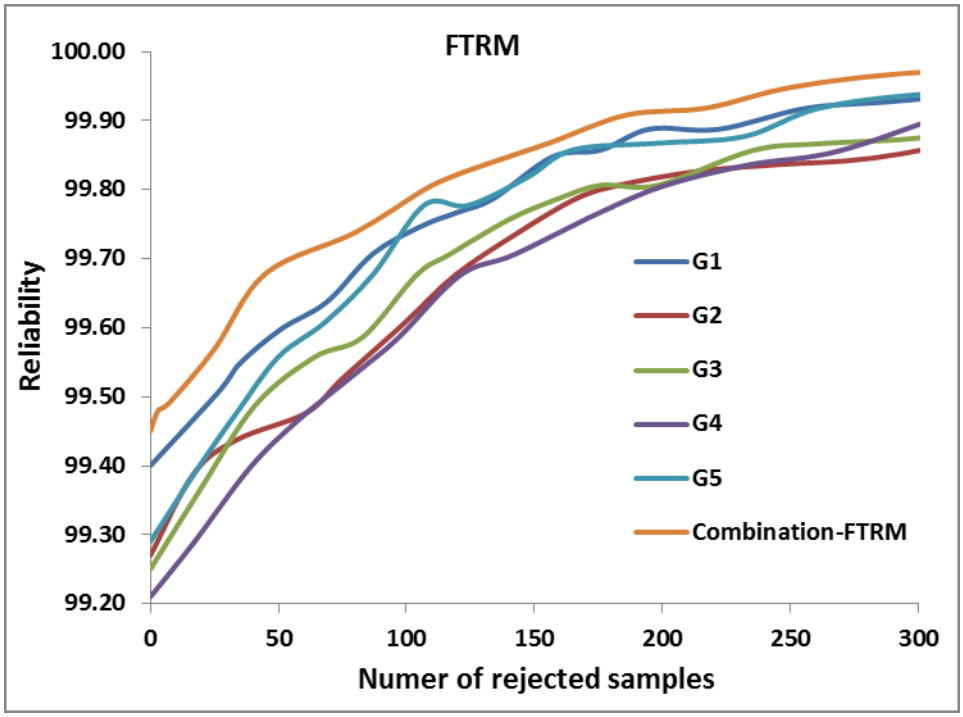

Figure 15 (b) ROC curves of MCS (DR) and single models with FTRM on MNIST database

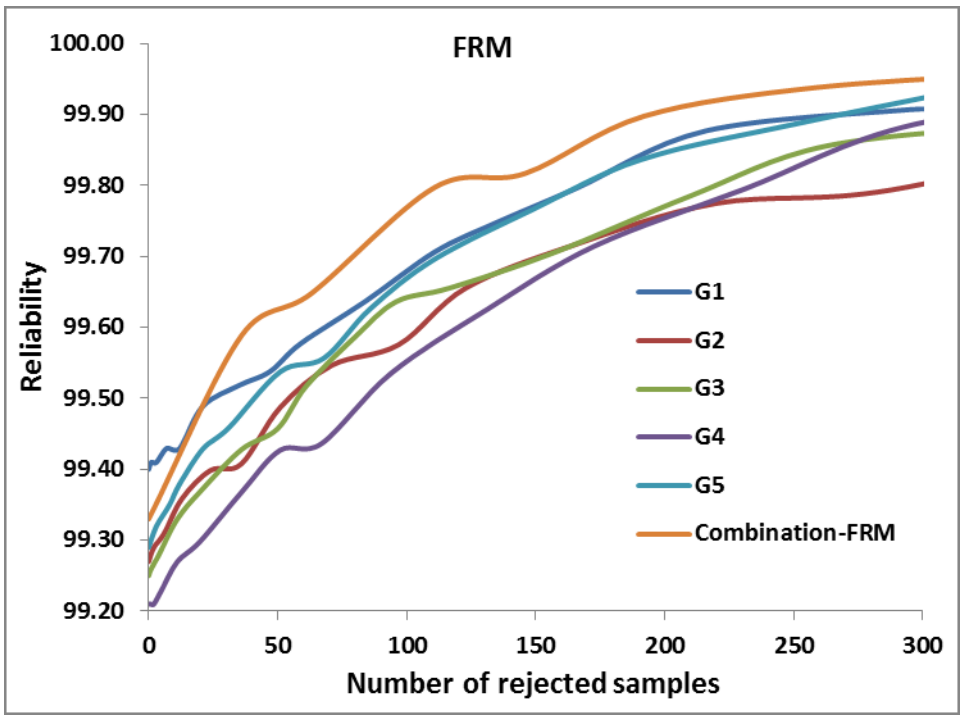

Figure 15 (c) ROC curves of MCS (DR) and single models with FRM on MNIST database

Through the ROC curves, it is proven again that by adopting the combination method, higher rejection and recognition (0 point of x-axis) performances can be obtained by MCS. With five single models whose recognition rates are almost all below 99.30% (only one is 99.40%), the combination systems with FTRM and SVMM as weight parameters achieve 99.46% and 99.45% recognition rates (0 point of x-axis) separately. With rejection rates of about 4.7%, both of them reach 100% reliabilities. On the other hand, although almost all the recognition rates of single models in MCSs built by DR are less than those by SM, the combination ROC curves (with different weight parameters) of DR method rise faster than those of SM method. In spite of lower starting points, MCSs built by DR achieve 100% reliability with fewer samples rejected than those by SM. For MCSs built by DR, the combination

systems with both FTRM and SVMM achieve 100% reliability when 4.7% of the samples have been rejected, while in MCSs built by SM, 100% reliability can only be reached with the rejection rate of at least 5.7% (in the case of SVMM), as shown in Table 11. It is demonstrated that the MCSs with construction method DR work more efficiently than those with SM. Analyses indicate that the reason for this is that building MCS with DR makes errors between different classifiers in the system much more diverse. As a result, combining the decisions of individual classifiers can make the clear samples distinct from confusing ones, since they are prone to gain consistent decisions from different classifiers, producing much larger combination output values. Therefore, it is easier to separate the confusing samples by setting relatively large thresholds on the output values.

Table 11. Combination results of different MCSs designed by different methods with different types of weight parameters on MNIST

	SVMM-SM	SVMM-DR	FTRM-SM	FTRM-DR
Error rate of combination (%)	0.46%	0.54%	0.43%	0.55%
Rejection rate at 100% reliability (%)	5.70%	4.75%	5.95%	4.74%

(2) CENPARMI

In this experiment, we apply the soft voting combination method for MCS rejection on CENPARMI handwritten numeral database. The MCSs are also constructed by both SM and DR methods, as presented previously in Tables 6 and 7 of Section 4.1. Both FTRM and SVMM are chosen as weight parameters

for soft voting combination, since they come from different criterion categories (heuristic- and learning-based respectively). Thresholds are searched from 0 with an incremental step of 0.05 until suitable reliability values are reached. The results are shown as ROC curves displaying the relationship between number of rejected samples and reliability, as presented in the following four figures.

Figures 16 (a) and (b) display the result ROC curves of single models and MCSs built by both of SM and DR with FTRM selected as weight parameter.

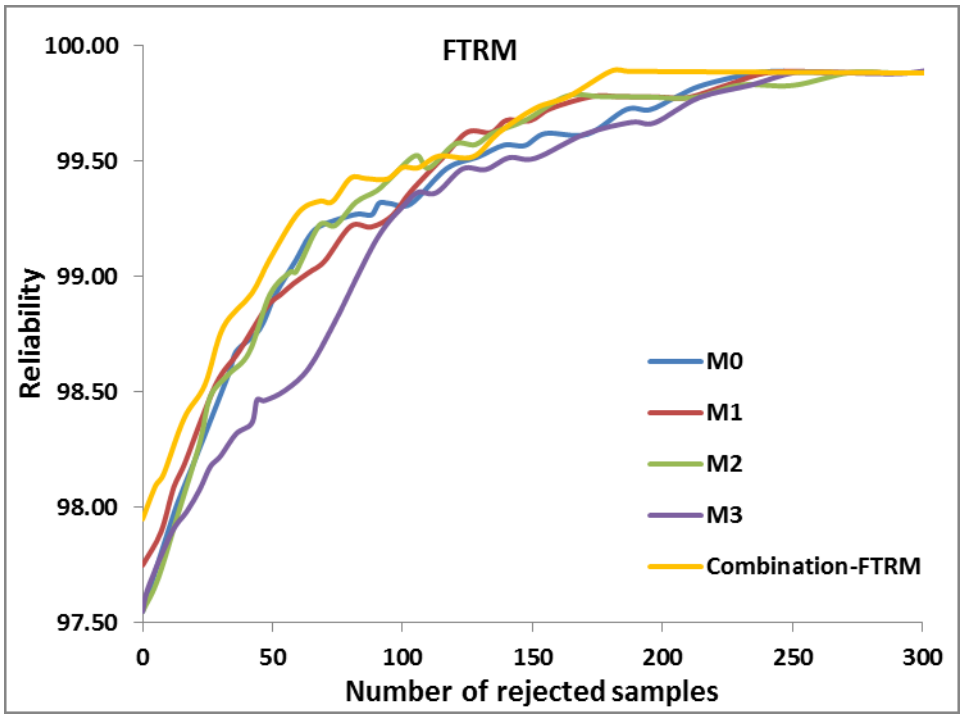

Figure 16 (a) ROC curves of MCS (SM) and single models with FTRM on CENPARMI database

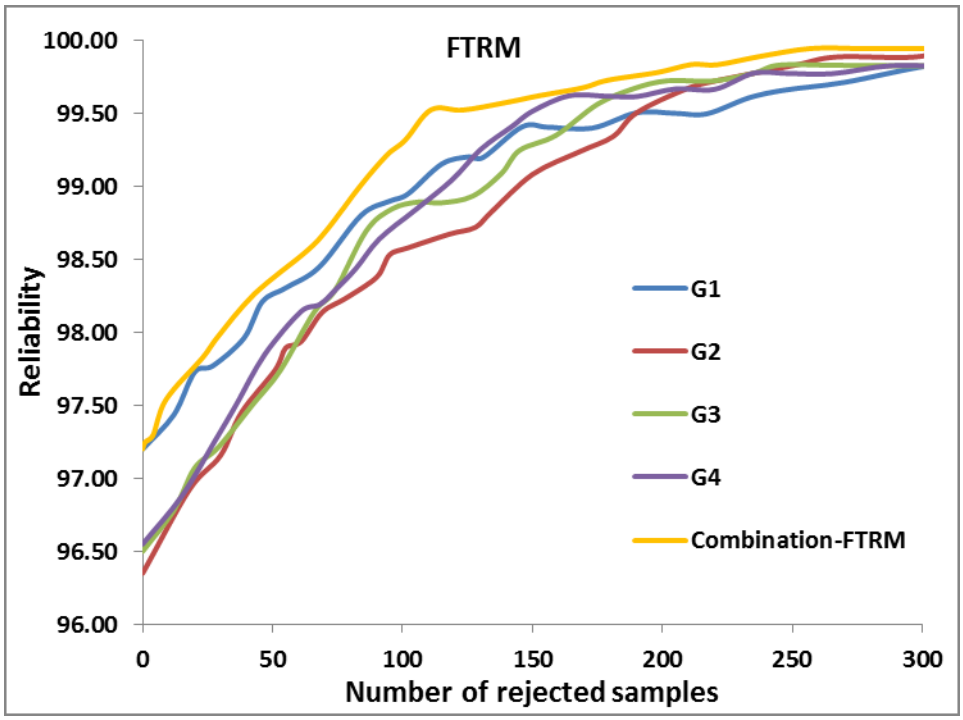

Figure 16 (b) ROC curves of MCS (DR) and single models with FTRM on CENPARMI database

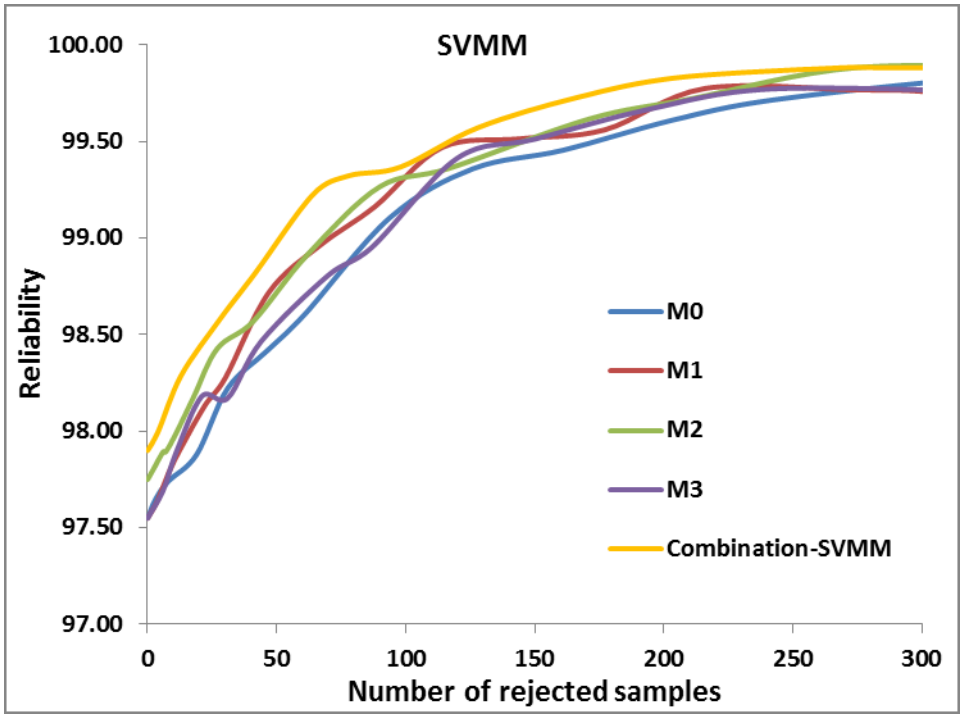

Figure 17 (a) ROC curves of MCS (SM) and single models with SVMM on CENPARMI database

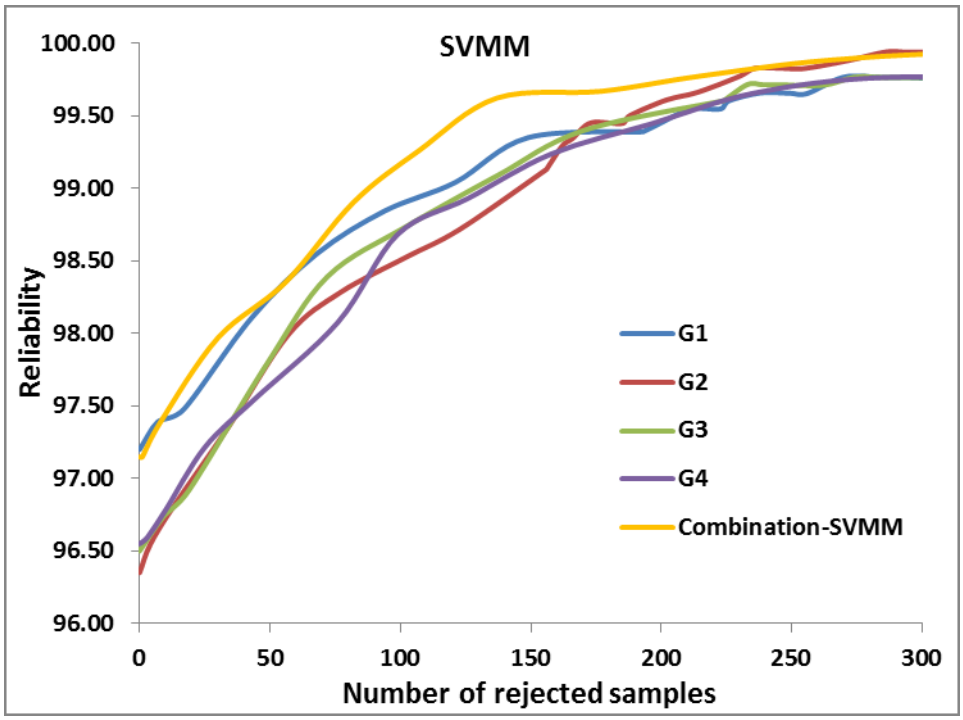

Figure 17 (b) ROC curves of MCS (DR) and single models with SVMM on CENPARMI database

Figures 17 (a) and (b) present the result ROC curves of single models and MCSs built by both of SM and DR. But in this case, SVMM is chosen as weight parameter.

From these figures, it is proven again that soft voting combination method with MCS can improve the rejection performance of the system no matter which method is adopted to construct the MCS or which criterion is selected as weight parameter.

Furthermore, Figure 17 (a) shows that although MCS does not necessarily improve the recognition rate (0 point of x-axis), , it can still improve the rejection performance of the whole system through the proposed combination method. Table 12 below lists some information from these two figures along with He's research result [31], in which, it is claimed that by using LDAM, a reliability of

99.67% is achieved with 175 samples rejected. With our voting based combination methods, the MCS built by SM (Com-SM) obtains a higher reliability of 99.78% with 11 fewer samples rejected. The other MCS built by DR (Com-DR) achieves the same reliability 99.67% as LDAM with 6 fewer samples rejected and 99.73% with 179 samples rejected. MCSs constructed by both of the two methods obtain better rejection results than state-of-the-art rejection method on the same database.

Comparing these two different construction methods of MCS (SM and DR), it is clear that the system with DR performs better than that with SM, since to reach the high reliability of 99.94%, DR should reject 257 samples while SM should reject 393 samples, even if the original recognition rate (0 point of x-axis) of DR is lower than that of SM (refer to Tables 6 and 7). This also demonstrates that MCS built by DR makes errors between different classifiers in the system much more diverse; thus, the rejection performance is enhanced by combining the classifiers' decisions with the proposed voting based method.

Table 12. Rejection performances of different rejection methods on CENPARMI

Number of rejected samples	Reliability	Method
175	99.67%	[6]
164	99.78%	Com-SM
180	99.89%	Com-SM
169	99.67%	Com-DR
179	99.73%	Com-DR
393	99.94%	Com-SM
257	99.94%	Com-DR

(3) USPS

Similar experiments are performed on USPS database with both versions.

FTRM and SVMM are selected as weight parameters.

For the first version (V1) with 7291 training samples and 2007 testing samples, the results of MCSs built by both SM and DR are presented in Figures 18 (a) and (b) with FTRM as weight parameter. Figures 19 (a) and (b) show the MCSs built by the same methods with SVMM selected as weight parameter.

Same work has been completed on the second version (V2) of USPS with equal amount of samples in both of training and testing sets. MCSs are created by both of DR and SM methods. Results are displayed as ROC curves in Figures 20 (a) and (b) with FTRM as weight parameter while Figures 21 (a) and (b) use the SVMM.

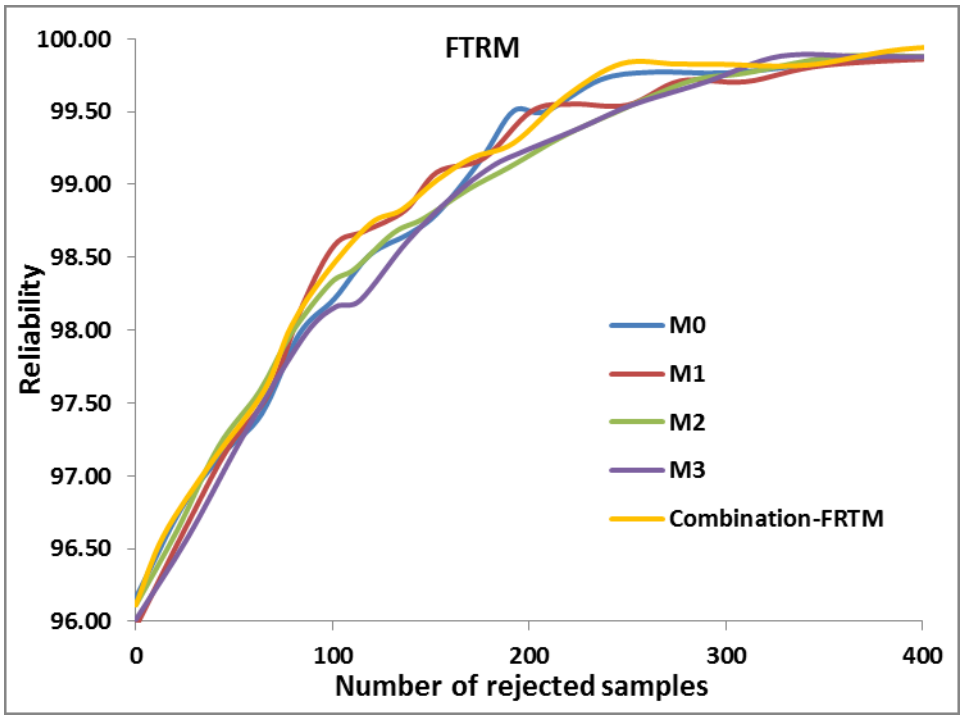

Figure 18 (a) ROC curves of MCS (SM) and single models with FTRM on USPS-V1 database

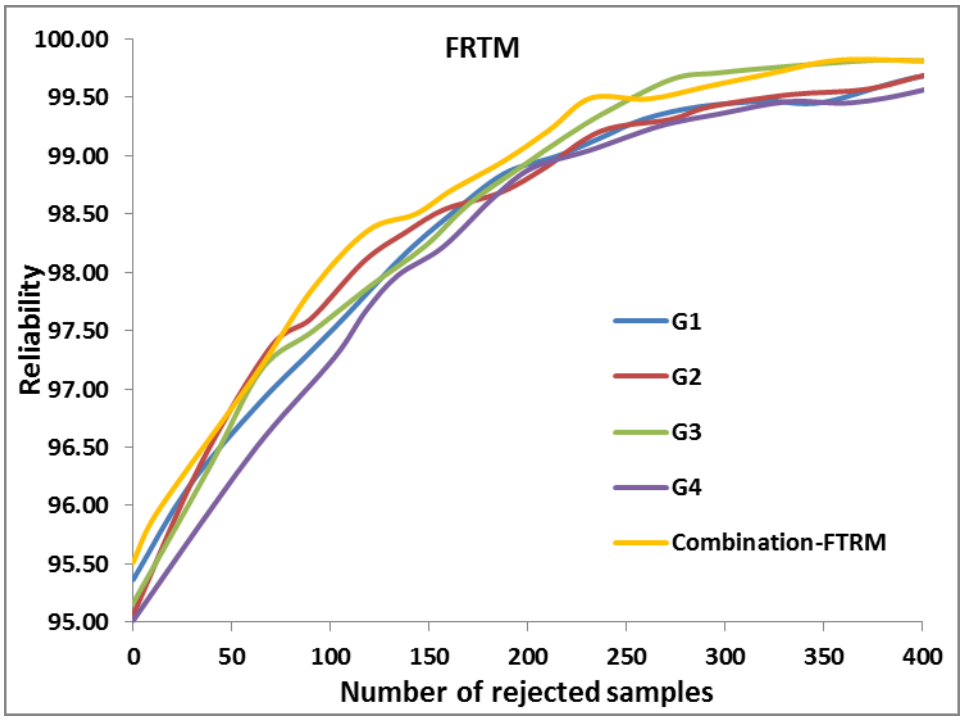

Figure 18 (b) ROC curves of MCS (DR) and single models with FTRM on USPS-V1 database

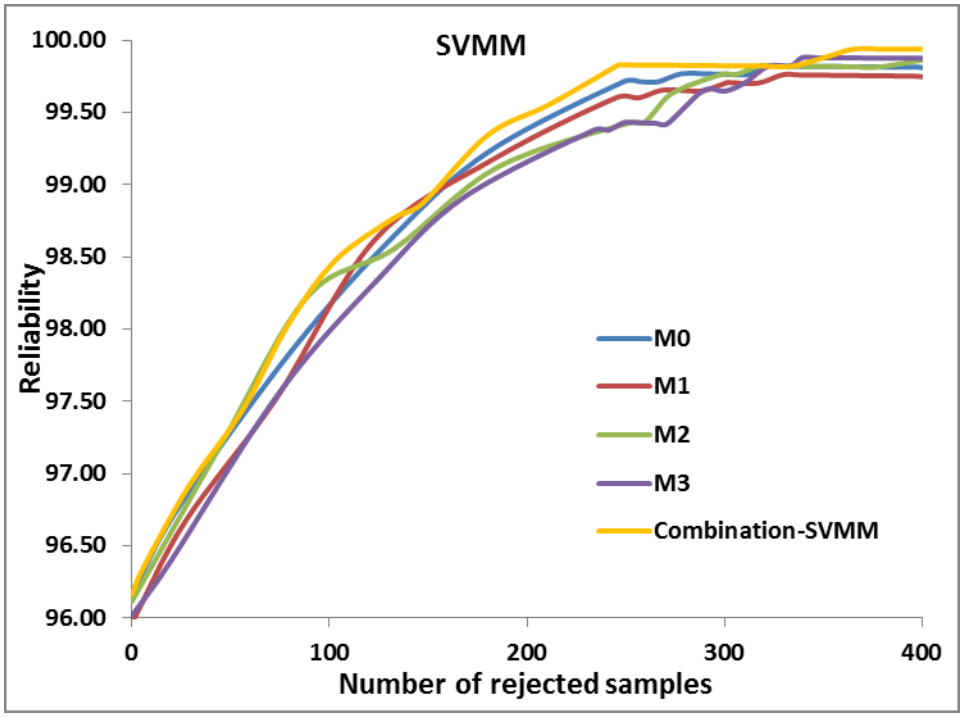

Figure 19 (a) ROC curves of MCS (SM) and single models with SVMM on USPS -V1 database

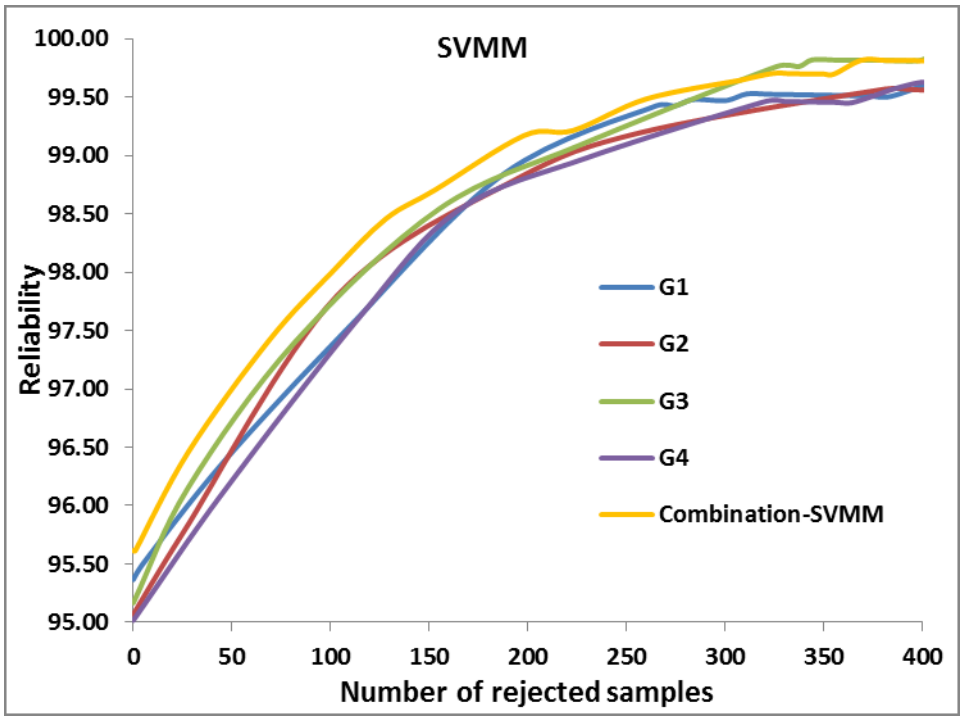

Figure 19 (b) ROC curves of MCS (DR) and single models with SVMM on USPS-V1 database

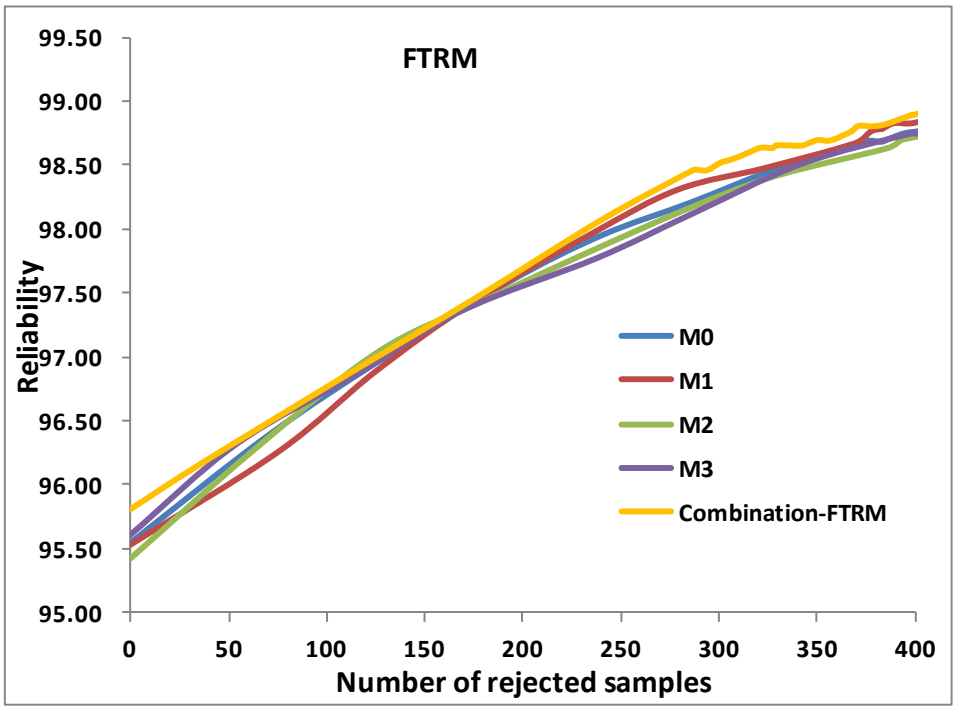

Figure 20 (a) ROC curves of MCS (SM) and single models with FTRM on USPS –V2 database

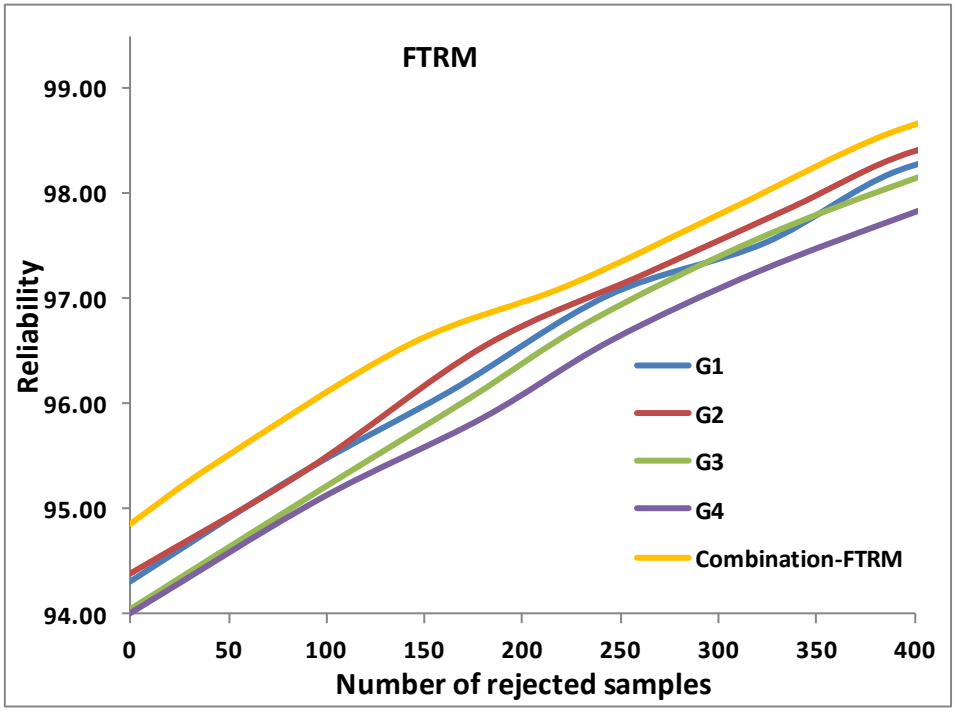

Figure 20 (b) ROC curves of MCS (DR) and single models with FTRM on USPS –V2 database

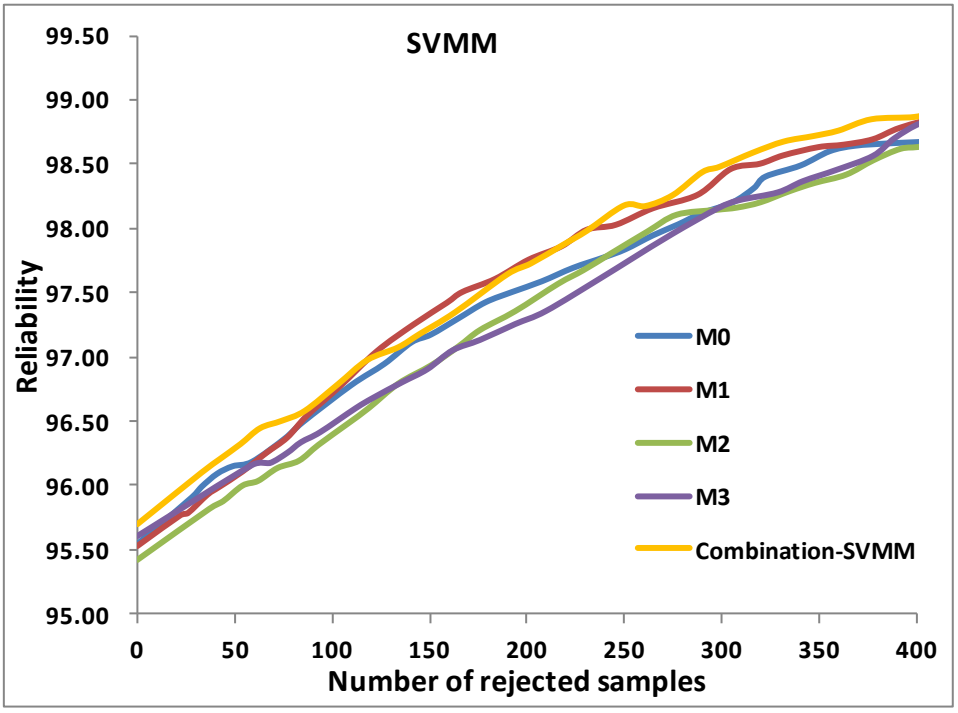

Figure 21 (a) ROC curves of MCS (SM) and single models with SVMM on USPS –V2 database

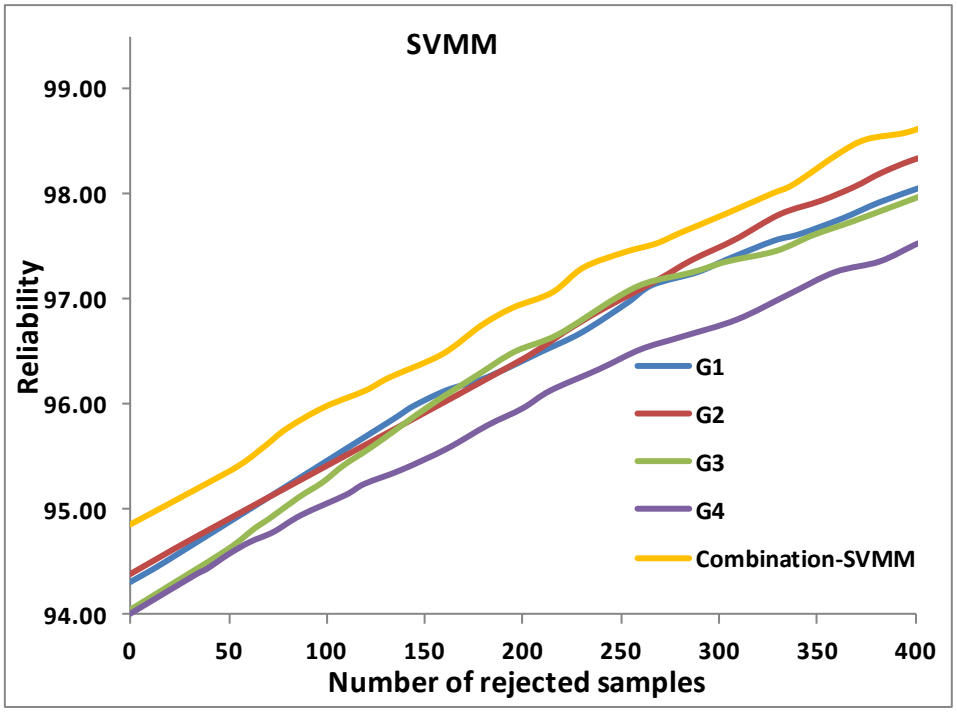

Figure 21 (b) ROC curves of MCS (DR) and single models with SVMM on USPS –V2 database

Although the gaps between the rejection performances of MCSs and single models are not so distinctive in Figures 18 and 19, they prove that the decisions given by various MCSs with the proposed combination method enhance the rejection performances. That is because their curves are above the single models' curves along their entire paths in all the four cases. Figures 20 and 21 display the results of MCSs built by SM and DR with different criteria as weight parameters on USPS version 2. In all of the graphs, improvements in rejection performances remain consistent proving once again the effectiveness of proposed combinations. In addition, the progresses generated in the cases of MCSs built by DR is always more recognizable than those by SM. The most obvious improvements are in the cases of MCSs built by DR with both SVMM and FTRM. All the results confirm that MCS rejection methods work more effectively than the criteria with single

classifiers; meanwhile, MCSs built by DR outperforms their SM counterparts in rejection.

Chapter 5: Combination with Class-specialist

In the previous chapter, outcomes provided by the MCS classifiers were combined with a soft-voting method for rejection where different rejection criteria, reflecting the single classifiers' rejection performances, were selected as the weight parameters. The results showed that this combination method can consistently improve single classifiers' rejection performances.

Considering single models in MCS have their specific strengths in the classification process, it is not advisable to treat all predicted results given by various classifiers at the same level. For example, if classifier A outperforms classifier B in recognizing samples from class "4", the predicted label "4" given by classifier A should be treated with a higher confidence level than a predicted label "4" from classifier B. Thus, it is necessary to consider the specialist capability of single models as a new type of confidence value and incorporate it into the voting based combination process, in order to enhance the rejection system. In this chapter, class-specialist information will be integrated into the proposed combination method for MCS rejection.

5.1 Method with Class-specialist Information

Confusion matrix is an effective tool in representing the specialist categories of various classifiers. It is calculated for each classifier based on the training set in order to identify the classifier with fewest errors and help determine the specialist for each possible category. Some specialist information can be extracted from these matrices

and used as part of the weight parameter in the combination process.

In order to represent the specialist information, a new type of confidence weight parameter is introduced as Conf_{ij} $(i = 1,2,\dots,N;\ j = 1,2,\dots,c)$, which is derived from the confusion matrices on the training set. We created two different ways to compute this confidence weight: a simple one which reflects only the specialist classifier of each category and a complex one which reflects the specialist degree of different classifiers in each category.

In the first designing method (referred to as S1), Conf_{ij} has only two values including 0 and 1. Suppose there are N different classifiers in the MCS, denoted by $g_1, g_2, \dots, g_N$ and c possible classes, shown as $l_1, l_2, \dots, l_c$. For a specific category l_j $(j = 1,2,\dots,c)$, there is a specialist classifier $g_i (i = 1,2,\dots,N)$ with fewest errors among all classifiers. For a pattern, g_i would provide a prediction of the label y_i, where Conf_{ij} equals 1 if g_i is the specialist of the predicted category y_i or else, it gives 0, as seen in Eq. (15):

$$\mathrm{Conf}_{ij} = \begin{cases} 1 & \text{if } y_i = l_j \\ 0 & \text{else} \end{cases} \tag{15}$$

It is noted that, for each classifier, there may be several specialist categories and the confidence values Conf_{ij} are 1 for all of them.

In the second method (referred to as S2) of Conf_{ij}, it reflects the specialist degree of single classifier in each category. It is mentioned above that for each category l_j $(j = 1,2,\dots,c)$, there is a specialist classifier with the fewest errors, denoted as g_m, while a classifier with the most errors is represented by g_n. In

addition, we analyze the number of mistakes made by each of the classifiers in this category l_j as $e_{ij}(i = 1, \dots, n, m, \dots, N)$. $Conf_{ij}$ is calculated in Eq. (16):

$$Conf_{ij} = 1 - \frac{e_{ij}}{e_{nj}} \tag{16}$$

In this case, the classifier making more errors in a specific category will get a smaller confidence weight value $Conf_{ij}$ when compared to the one with fewer errors, since e_{nj} is a fixed number. So, $Conf_{ij}$ reflects the specialist degree of a classifier in a specific category through the number of produced errors.

During the combination process, the same soft voting combination method is performed alongside with specialist information which is added as a new confidence weight parameter. Each model $g_i(i = 1,2, \dots, N)$ would provide a prediction of the label y_i as well as an output vector $\{f_1^i, f_2^i, \dots, f_c^i\}$ for a random pattern. Then, for each classifier, the selected rejection measurement (FRM, FTRM, SVMM and so forth) can be calculated based on the output vector, denoted as $t_i(i = 1,2, \dots, N)$. After that, soft voting is performed and a voting value $V_j(j = 1,2, \dots, c)$ is calculated for each of the classes as seen in Eq. (17).

$$V_j = \sum_{i=1}^{N} t_i I(y_i, j)(1 + Conf_{ij}), \quad I(y_i, j) = \begin{cases} 1 & \text{if } y_i = j \\ 0 & \text{else} \end{cases} \tag{17}$$

Within V_j, the maximum voting value $V_{max} = \max_{1 \le j \le c} V_j$ can be determined and then, thresholds T_{com} are searched and applied for V_{max}. If V_{max} is smaller than a threshold, the pattern will be rejected.

5.2 Experiment with Class-specialist Information

In Section 4.1, we construct MCS by SM method on the MNIST dataset. To evaluate the effectiveness of different classifiers in this system, confusion matrices are displayed in Tables 13(a~g), with specialist categories marked:

Table 13(a). Confusion matrix of M0

predict / true	0	1	2	3	4	5	6	7	8	9	Sum
0			1			1	2		2		***6***
1								9	1	1	11
2		2						5	6	2	15
3		1	2			6		3	8	7	27
4		2					3	1		16	22
5		1	1	3	1		11		8	0	25
6	5		1		3	3			7	1	20
7		6	3		2					3	14
8		1	1	4	7	5	2	1		9	30
9	3	1		3	19	1		14	5		46

Table 13(b). Confusion matrix of M1

predict / true	0	1	2	3	4	5	6	7	8	9	Sum
0		1		1			4		2		8
1								11			11
2		2						6	3	3	14
3			4			4		5	8	7	28
4		2					4	1	1	16	24
5		1	1	3			10		11	3	29
6	2				3	1			6	1	13
7		5	5	1	3					1	15
8		1	1	2	3		2	1		9	19
9	2			2	13	4		11	8		40

Table 13(c). Confusion matrix of M2

predict / true	0	1	2	3	4	5	6	7	8	9	Sum
0		1	1				3		1		***6***
1			5					10	1		16
2	1	3		1				9	3	1	18
3			3			2		4	6	5	***20***
4		3					3	1	1	11	***19***
5		1	1	5	1		8		7	4	27
6	3				4	3			3		13
7		5	1		1				1	2	10
8		1	3	2	3	3	4			6	22
9	3	1		1	14	2		8	4		33

Table 13(d). Confusion matrix of M3

predict / true	0	1	2	3	4	5	6	7	8	9	Sum
0		1				2			3		***6***
1						1	1	10	1		13
2	2	3						9	3	1	18
3	1		6			3		4	6	8	28
4		3					3	2		19	27
5		1	1	5	1		8		6	2	24
6	2		1	1	3	3			5		15
7		4	5		1				1	3	14
8	1	1	2	1	5	4	4	1		8	27
9	1		1		14	3		10	4		33

Table 13(e). Confusion matrix of M4

predict / true	0	1	2	3	4	5	6	7	8	9	Sum
0		1				1	2		2		***6***
1								9		1	10
2		4						3	3		***10***
3			2			4		5	6	5	22
4		3					2	3		12	20
5		1		3			7		4	1	***16***
6	2				2				6		***10***
7		4	3		2					1	***10***
8	1	1			4	3	3	1		5	***18***
9	2			1	13	5		8	6		35

Table 13(f). Confusion matrix of M5

predict / true	0	1	2	3	4	5	6	7	8	9	Sum
0		1	1				1		3		***6***
1							2	10		2	14
2	1	4		2				6	3	2	18
3			2			3		3	5	8	21
4		2		1			3	2		14	22
5	1	1		6	2		9		8	3	30
6	3	1		1	3	2			4	1	15
7		3	7							4	14
8	1	1	1	3	3	3	7	1		9	29
9	5			3	10	3		8	5		34

Table 13(g). Confusion matrix of M6

predict / true	0	1	2	3	4	5	6	7	8	9	Sum
0		1		1			2		3		7
1								8			***8***
2		3		1				7	3	1	15
3			2			3		3	11	5	24
4		2				1	2	1	1	13	20
5	1	1		5	1		9		6	2	25
6	2				3	3			6	1	15
7		4	3		3				1	1	12
8		1	1	2	4	2	4	1		7	22
9	1		1		13	2		6	4		***27***
Sum	4	12	7	9	24	11	17	26	35	30	175

Table 14. Different models with least and most errors in each category

Class / Model	0	1	2	3	4	5	6	7	8	9
Fewest errors (No. errors)	M0,M2, M3,M4, M5 (6)	M6 (8)	M4 (10)	M2 (20)	M2 (19)	M4 (16)	M4 (10)	M4 (10)	M4 (18)	M6 (27)
Most errors (No. errors)	M1 (8)	M2 (16)	M2 (18)	M3 (28)	M3 (27)	M5 (30)	M0 (20)	M1 (15)	M0 (30)	M0 (46)

The information about the classifiers with fewest and most errors as well as the

number of errors in each category is extracted from the matrices and listed in Table 14.

There are two ways (referred to as S1 and S2) to calculate the confidence weight parameter $Conf_{ij}$, which can integrate the class specialist information into the combination process, as mentioned in Section 5.1. The experiments with both of these two methods are conducted respectively. Their results along with the original combination result without the specialist information are presented as ROC curves in Figures 22. In this case, MCS is built using method SM and FTRM is chosen as weight parameter for combination.

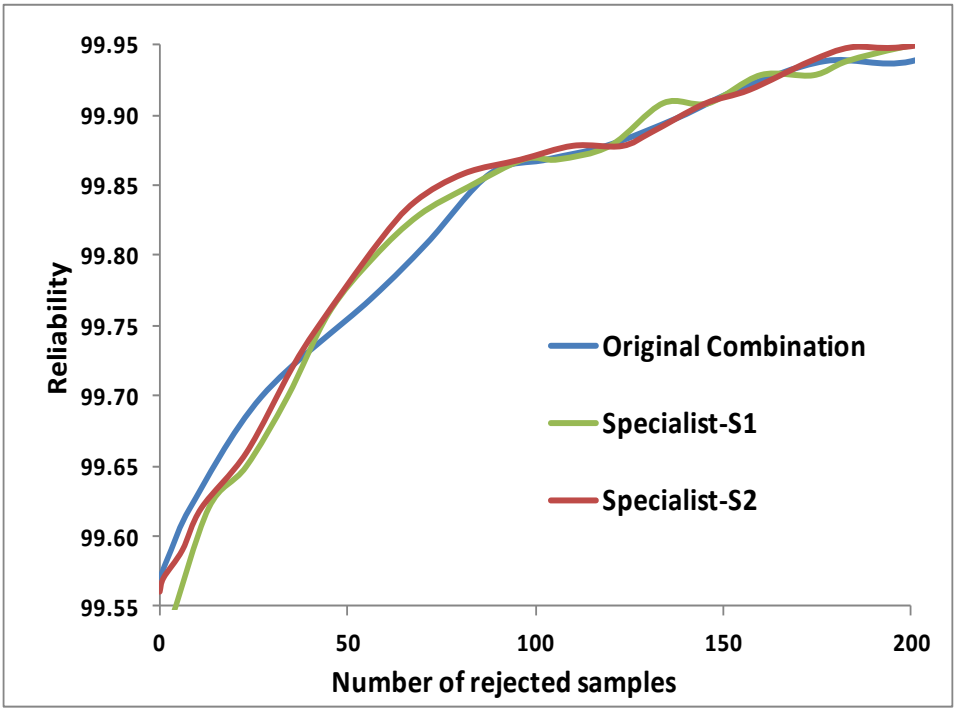

Figure 22. ROC curves of original combination and combination with specialist information calculated by S1 and S2 in MCS(SM)

Figure 22 shows that the combination with specialist information can improve the rejection performance of the original method to a certain extent. At the stage where a small amount of samples are rejected, the original combination works more

effectively when compared to S1 and S2. But, as more samples are rejected, the methods with specialist information surpass their original counterpart and then, the three lines start to perform in a similar manner. The comparison of these two designing methods for confidence weight parameter demonstrates that it is very difficult to determine which one is better in this case. From the graph, it is observed that recognition rate of S2 is higher than S1 without any rejection; yet, their ROC curves intertwine as the number of rejections increases. Very similar result appears with SVMM used as weight parameter, which is not shown to avoid redundancy.

The same experiment is conducted to the MCS built by DR. All the result ROC curves of S1, S2 and original method with FTRM and SVMM used as weight parameters are displayed in Figures 23 and 24 respectively. From these figures, it is observed that the curves of the original combination along with those with specialist information are too much overlapped to compare their performances. However, the combinations with specialist information can actually reduce the number of rejected samples to achieve 100% reliabilities. By comparing the performances of MCS rejections integrated with specialist information, S1 outperforms S2 when two types of weight parameters (FTRM and SVMM) are applied. With FTRM as a weight parameter, the combination with S1 reaches the 100% reliability point at the expense of 4.09% rejection rate, while the one with SVMM reaches the 100% reliability point at the expense of 4.10% rejection rate, as marked with blue circles in these two figures. The best rejection performance on MNIST, which rejects 409 samples to reach 100% reliability, comes out in the combination system of S1 with FTRM

selected as weight parameter. Similar result, which rejects 410 samples to achieve 100% reliability, appears in the combination system of S1 with SVMM.

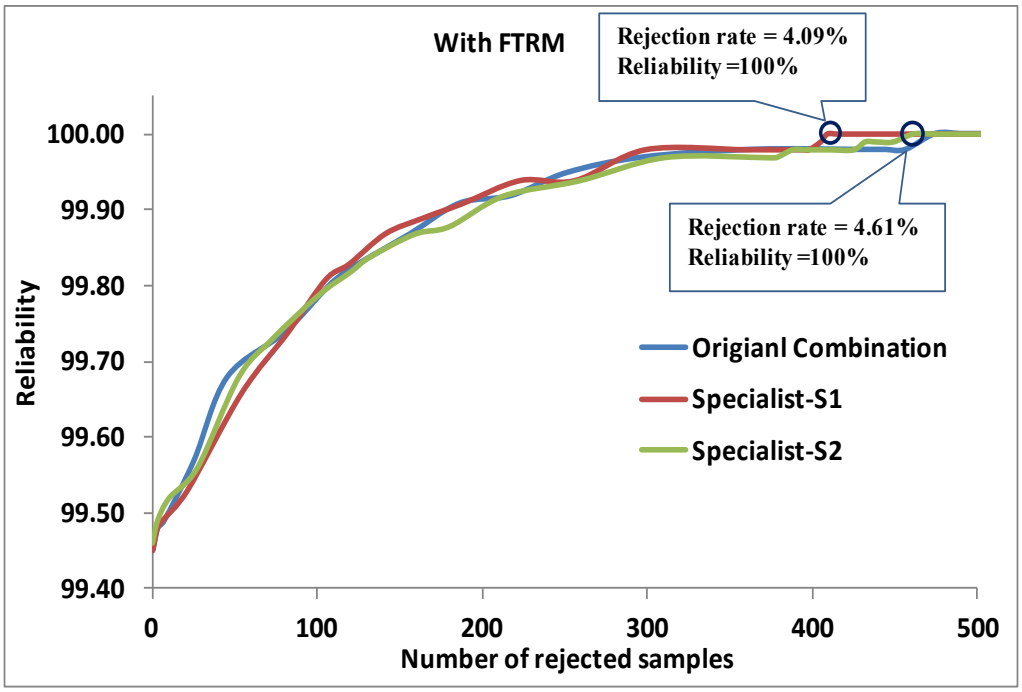

Figure 23 ROC curves of original combination and combinations with specialist information with FRTM as weight parameter in MCS(DR)

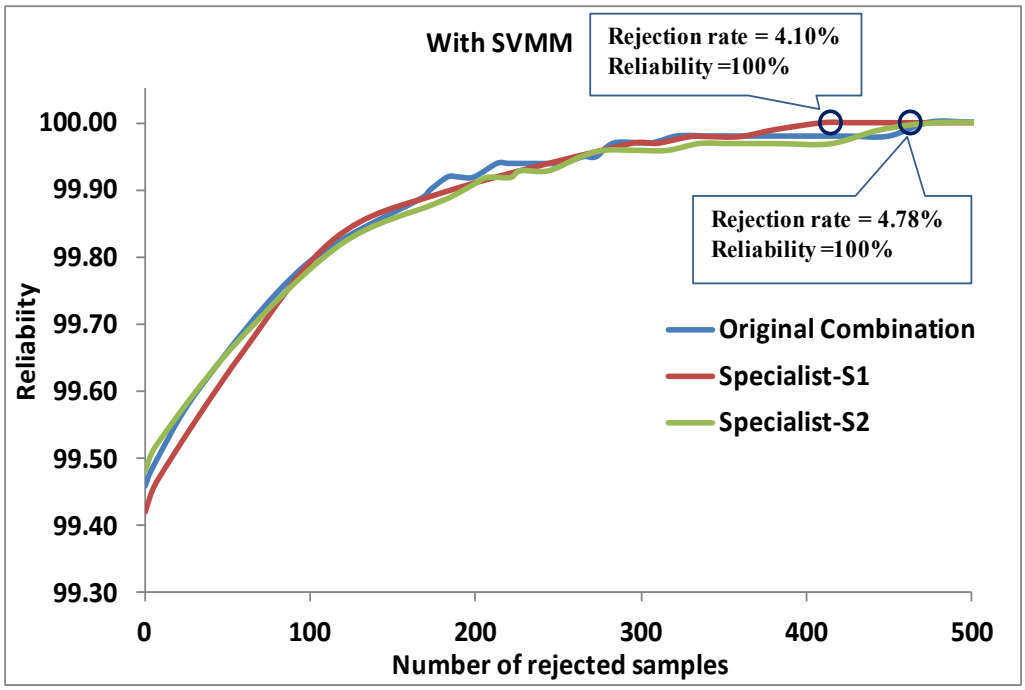

Figure 24 ROC curves of original combination and combinations with specialist information with SVMM as weight parameter in MCS(DR)

In the analysis, we can see that, in method S1, the predict label provided by the specialist classifier in a category contributes much more than other predictions to the process of combination; meanwhile, in method S2, the predictions of all the classifiers contribute to the combination process to a certain extend according to their specialist degrees. At last, the rejection of fewer samples to obtain 100% reliability in both S1 and S2 when compared to the original combination method demonstrates that combination with specialist information can achieve a better performance after decreasing the need of non-specialist classifier's information which may interfere with the final results. In addition, the emphasis on the specialist information allows S1 to outperform S2, in which case, some non-specialist information is still taken into account.

Chapter 6: Conclusion

By focusing on the rejection process of offline handwritten numeral recognition, we hope to enhance existing recognition systems in order to decrease processing errors for handwritten documents, such as cheques. Having a highly reliable recognition system can potentially reduce losses at financial institutions while improving employees' productivity, since the machines can complete time consuming tasks with greater accuracy.

In order to increase recognition systems' reliabilities, we looked at two novel learning-based rejection criteria for single classifier and rejection methods with MCS based on soft voting combination. The newly proposed rejection criteria with single models are then compared with several traditional criteria on the benchmark MNIST handwritten digit database. The voting based rejection methods with MCS are evaluated on three handwritten digit databases including MNIST, CENPARMI and USPS based on MCSs constructed by Structure Modification (SM) and Dataset Re-sampling (DR). Experimental results are quite encouraging and will be presented in Section 6.1. Also, it is seen that the work with rejection can be further improved, as summarized in Section 6.2 under Future Work.

6.1 Contribution

This research contributes to the field through rejection criteria designing which aims to improve the reliability of recognition systems, by looking at two novel rejection criteria for single classifiers including SVM-based measurement (SVMM)

and Area Under the Curve measurement (AUCM). Also, voting based combination methods of multiple classifier system (MCS) are proposed for pattern rejection. The main contributions of this thesis are summarized in the following paragraphs.

Firstly, in order to evaluate the rejection performance of a criterion or a system, two factors have to be considered simultaneously: the number of rejected samples and the reliability. Since there is always a tradeoff between these two factors, it is insufficient to verify rejection performance based on one of them exclusively. Therefore, we introduce a ROC space consisting of these two factors and curves in it represent the performances of different rejection processes. A good rejection criterion can achieve a higher reliability with fewer samples rejected. As a result, we expect a good ROC curve to be as close to the top left corner as possible and this is applied to evaluate all the rejection criteria proposed through the whole thesis.

Secondly, we propose two novel rejection criteria for single classifiers: SVMM and AUCM. Both of them are learning-based rejection criteria, meaning that they are obtained based on the training data. Unlike the traditional criteria based on heuristic ideas, these two extend the rejection process into the training procedure. SVMM locates a linear optimal rejection boundary between confusing samples and clear samples by learning from the training data in order to predict the rejections on testing samples. AUCM determines a linear combination of FR and SR, seen as the most meaningful ones among all these confidence values, for rejection based on all training samples. The optimal combination is the one that maximizes the area under the ROC curve used for representing the performance of the rejection system. Both of them are

more straight-forward than the heuristic criteria and can retrieve more information from the data, especially the training data. With a CNN classifier based on the MNIST database, these two rejection criteria are compared with three traditional rejection criteria that have been proven to be very effective. The results demonstrate that SVMM always works better than FRM and LDAM, as the ROC curves for it are always above those of the other two (refer back to Figures 7 and 9). Although performances of SVMM and FTRM are too close to determine which one is better, SVMM is still proven to be a good rejection criteria since FTRM has distinguished itself in this model. Moreover, The ROC curve of AUCM is much closer to the left-top corner than those of FTRM and FRM, and the reliability values of AUCM remain higher than those of FTRM and FRM in almost their entire paths (refer back to Figure 11). It means that with the same number of patterns rejected, AUCM always achieves higher system reliability than FTRM and FRM. All the results show that the newly proposed learning-based rejection criteria reach higher performance than the heuristic designed ones, demonstrating the effectiveness of the learning-based rejection idea.

Thirdly, voting based combination methods for MCS rejection are presented. It is a preliminary attempt to adopt MCS for the purpose of rejection. MCSs are constructed in two different ways including DR and SM. Both hard voting and soft voting are considered for combination. In the hard voting process, experiment is performed with MCS built by SM on the MNIST database. A range limitation problem makes the rejection process inflexible since the thresholds can only be set to

limited values. As a result, once the maximum value is reached, the reliability cannot be improved anymore. It cannot yield a ROC curve either. To solve the problem, the soft voting method is introduced and different rejection criteria (FRM, FTRM and SVMM) are used as weight parameters for different models since they can reflect the rejection effectiveness. Experiments are conducted on MNIST, CENPARMI and USPS. Different MCSs are constructed with SM and DR. The results show that no matter what building method is chosen or what criterion is selected as weight parameter in soft voting, rejection based on MCS can improve the rejection performance of the system consistently (refer back to Figures 14~20). They also demonstrate that MCSs built by DR work better than those by SM in rejection (refer back to Figures 16~21). In order to further improve the performance of MCS for rejection, the class-specialist information is integrated into the soft voting process by introducing a new confidence weight parameter. With two different designing ways of this new parameter, the soft voting process is slightly changed, leading to improvements of the rejection performance (refer back to Figures 22~24). The best result appears in the case of MCS built by DR with specialist information integrated by S1. Expenses of 4.09% and 4.10% rejection rates to reach a reliability of 100% are accomplished with FTRM and SVMM selected as weight parameters respectively.

6.2 Future Work

Until it is possible to eliminate recognition errors, there will always be research on rejection in handwritten recognition. The following are proposed methods which can push the way forward:

The SVMM is designed based on the well-known classifier, SVM, which has achieved extraordinary recognition rates. However, for the rejection problem, SVM does not work as effectively as in the regular recognition field. The main reason is that in regular recognition, there are nearly the same amount of samples from each possible category, making the boundary locating process much easier and more accurate. However, in this case, a serious unbalancing problem appears because the baseline accuracy of the classifier is high. We believe that if we can figure out a way to solve the unbalancing problem, SVMM can achieve a superior performance.

The other learning-based rejection criterion is AUCM which attempts to find an optimal combination of FR and SR for rejection. In our model, we interpret the optimal combination to be the one which maximizes the area under the ROC curve representing the criterion's performance on training data. It can be interpreted in other ways as well. Also, the combination can be derived from the five top ranks or all the rank values rather than just the first two ranks, allowing it to be much more representative. The third way to improve AUCM is in the part of determining the optimal parameters. We use a simple method of setting one fixed and conducting exhaustive search in the range of (-5.0, 5.0) for the other one, since we only have two parameters to determine. With more parameters, back-propagation algorithm can be applied which we believe will produce better results.

Furthermore, the rejection method with MCS is a task worth more exploring. In our research, the MCSs are constructed in two simple ways including SM and DR. Although the results demonstrate that combining several classifiers with the proposed

soft voting method can improve the rejection performances of single classifiers consistently, the final result still depends on the single classifiers. If the rejection performances of the committees that compose the MCS are better, the result after combination will improve accordingly. On the other hand, if we enlarge the variety of the errors between different classifiers in the system, the combination result can also be enhanced. That is why MCSs built by DR always achieve better results than those by SM. Therefore, MCSs with committees consisting of different classifier models can achieve much higher performances, because they recognize patterns based on different types of features using different algorithms, making the errors more diverse.

References:

[1] G. Dzuba, A. Filatov, D. Gershuny, I. Kil, and V. Nikitin. Check amount recognition based on the cross validation of courtesy and legal amount field. *International Journal of Pattern Recognition and Artificial Intelligence*, 11(4): 639-655, 1997.

[2] N. Gorski, V. Anisimov, E. Augustin, O. Baret, and S. Maximov. Industrial bank check processing: the A2iA CheckReaderTM. *International Journal of Document Analysis and Recognition*, 3(4): 196-206, 2001.

[3] C.Y. Suen and J. Tan. Analysis of errors of handwritten digits made by a multitude of classifiers. *Pattern Recognition Letters*, 26(3): 369-379, 2005.

[4] Y. LeCun, L. Jackel, L. Bottou, A. Brunot, C. Cortes, J. Denker, H. Drucker, I. Guyon, U. Muller, E. Sackinger, P. Simard, and V. Vapnik. Comparison of learning algorithms for handwritten digit recognition. In: *Proceedings of the International Conference on Artificial Neural Networks*, pp. 53-60, 1995.

[5] C.Y. Suen, C. Nadal, R. Legault, T.A. Mai, and L. Lam. Computer recognition of unconstrained handwritten numerals. *Proceedings of IEEE*, 80(7): 1162-1180, 1992.

[6] J.J. Hull. A database for handwritten text recognition research. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 16(5): 550-554, 1994.

[7] P.J. Grother. NIST special database19: handprinted forms and character database. *Technical report*, National Institute of Standards and Technology, March, 1995.

[8] C.L. Liu, K. Nakashima, H. Sako, and H. Fujisawa. Handwritten digit recognition:

benchmarking of state-of-the-art techniques. *Pattern Recognition*, 36(10): 2271-2285, 2003.

[9] C.L. Liu, K. Nakashima, H. Sako, and H. Fujisawa. Handwritten digit recognition: investigation of normalization and feature extraction techniques. *Pattern Recognition*, 37(2): 265-279, 2004.

[10] R.O. Duda, P.E. Hart, and D.G. Stork. *Pattern Classification 2nd Edition*. Wiley, New York, NY, 2000.

[11] F. Kimura, K. Takashina, S. Tsuruoka, and Y. Miyake. Modified quadratic discriminant functions and the application to Chinese character recognition. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 9(1): 149-153, 1987.

[12] D.E. Rumelhart, G.E. Hinton, and R.J. Williams. Learning representations by back-propagation errors. *Nature*, 323(9): 533-536, 1986.

[13] L. Tarassenko and S. Robert. Supervised and unsupervised learning in radial basis function classifiers. *IEE – Proceedings of Vision, Image and Signal Processing*, 141(4): 210-216, 1994.

[14] J. Schürmann. *Pattern Classification: A unified view of statistical and neural approaches*. Wiley, New York, NY, 1996.

[15] U. Kressel and J. Schürmann. Pattern classification techniques based on function approximation. In: H. Bunke, P.S.P. Wang (Eds.), *Handbook of character recognition and document image analysis*, World Scientific, Singapore, pp.49-78, 1997.

[16] G. Guo, S.Z. Li, and K. Chan. Face recognition by support vector machine. In: *Proceedings of IEEE International Conference on Automatic Face and Gesture*

Recognition, pp.196-201, 2000.

[17] T. Joachims. Text categorization with support vector machine: learning with many relevant features. In: *Proceedings of the European Conference on Machine Learning*, pp.137-142, 1998.

[18] Y.P. Estevan, V. Wan, and O. Scharenborg. Finding maximum margin segments in speech. In: *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 15-20, 2000.

[19] J.X. Dong, A. Krzyzak, and C.Y. Suen. Fast SVM training algorithm with decomposition on very large datasets. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 27(4): 603-618, 2005.

[20] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of IEEE*, 86(11): 2278-2324, 1998.

[21] O.D. Trier, A.K. Jain, and T. Taxt. Feature extraction methods for character recognition -- a survey. *Pattern Recognition*, 29(4): 641-662, 1996.

[22] Z.L. Bai and Q. Huo. A study on the use of 8-derectional features for online handwritten Chinese character recognition. In: *Proceedings of International Conference on Document Analysis and Recognition*, pp. 262-266, 2005.

[23] C.L. Liu, H. Sako, and H. Fujisawa. Discriminative learning quadratic discriminant function for handwriting recognition. *IEEE Transaction on Neural Networks*, 15(2): 430-444, 2004.

[24] P.Y. Simard, D. Steinkraus, and J.C. Platt. Best practices for convolutional neural networks applied to visual document analysis. In: *Proceedings of International*

Conference on Document Analysis and Recognition, pp. 958-962, 2003.

[25] F. Lauer, C.Y. Suen, and G. Bloch. A trainable feature extractor for handwritten digit recognition. *Pattern Recognition*, 40(6): 1816-1824, 2007.

[26] L. Lam and C.Y. Suen. Optimal combinations of pattern classifiers. *Pattern Recognition Letters*, 16(9):945-954, 1995.

[27] C.Y. Suen and L. Lam. Multiple classifier combination methodologies for different output levels. In: *Proceedings of International Workshop on Multiple Classifier System*, pp. 52-66, 2000.

[28] D.C. Ciresan, U. Meier, and J. Schmidhuber. Multi-column deep neural networks for image classification. In: *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3642-3649, 2012.

[29] C. Wu, W. Fan, Y. He, J. Sun, and S. Naoi. Cascaded heterogeneous convolution neural networks for handwritten digit recognition. In: *Proceedings of International Conference on Pattern Recognition*, pp. 657-660, 2012.

[30] X.X. Niu and C.Y. Suen. A novel hybrid CNN-SVM classifier for recognizing handwritten digits. *Pattern Recognition*, 45(4): 1318-1325, 2012.

[31] C.L. He, L. Lam, and C.Y. Suen. Rejection measurement based on linear discriminant analysis for document recognition. *International Journal of Document Analysis and Recognition*, 14(3): 263-272, 2011.

[32] C.L. He and C.Y. Suen. A hybrid multiple classifier system of unconstrained handwritten numeral recognition. *Pattern Recognition and Image Analysis*, 17(4): 608-611, 2007.

[33] P. Zhang, T.D. Bui, and C.Y. Suen. A novel cascade ensemble classifier system with a high recognition performance on handwritten digits. *Pattern Recognition*, 40(12): 3415-3429, 2007.

[34] W. Pan, T.D. Bui, and C.Y. Suen. Isolated Handwritten Farsi Numerals Recognition Using Sparse and Over-Complete Representations. In: *Proceedings of International Conference on Document Analysis and Recognition*, pp. 586-590, 2009.

[35] The MNIST Database of Handwritten Digits. http://yann.lecun.com/exdb/mnist/

[36] D. Keysers, T. Deselaers, C. Gollan, and H. Ney. Deformation models for image recognition. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 29(8): 1422-1435, 2007.

[37] M. Ranzato, C. Poultney, S. Chopra, and Y, LeCun. Efficient learning of sparse representation with an energy-based model. In: *Advances in Neural Information Processing Systems 19*, pp.1137-1144, 2006.

[38] C.Y. Suen, K. Liu, and N.W. Strathy. Sorting and Recognizing Cheques and Financial Documents. In: *Proceedings of Document Analysis System: Theory and Practice*, pp.173-187, 1999.

[39] The USPS Database of Handwritten Digits. http://gaussianprocess.org/gpml/data/

[40] P. Simard, Y. LeCun, and J. Denker. Efficient pattern recognition using a new transformation distance. In: *Advances in Neural Information Processing System 5, [NIPS Conference]*, pp. 50-58, 1992.

[41] H. Drucker, R. Schapire, and P. Simard. Boosting performance in neural networks. *International Journal of Pattern Recognition and Artificial Intelligence*,

7(4): 705-719, 1993.

[42] P.Y. Simard, Y. LeCun, J. Denker, and B. Victorri. Transformation invariance in pattern recognition -- Tangent distance and tangent propagation. In: *Neural Networks: Tricks of the Trade*, pp.239-274, 1998.

[43] M.E. Tipping. The Relevance Vector Machine. In: S. Solla, T. Leen, and K. Müller (Eds.), *Advances in Neural Information Processing Systems*, 12(1), MIT Press, Cambridge, MA, pp. 332-388, 2000.

[44] J. Dahmen, D. Keysers, H. Ney, and M.O. Güld. Statistical image object recognition using mixture densities. *Journal of Mathematical Imaging and Vision*, 14(3): 285-296, 2001.

[45] M. Karic and G. Martinovic. Improving offline handwritten digit recognition using concavity-based features. *A Bimonthly Journal with Emphasis on the Integration of Three Technologies*, 183, 2013.

[46] T. Fawcett. An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8): 861-874, 2006.

[47] C.C. Chang and C.J. Lin. LIBSVM: A library for support vector machine. *Journal ACM Transactions on Intelligent Systems and Technology*, 2(3): 27:1-27:27, 2011.

[48] L. Breiman. Bagging predictors. *Machine Learning*, 24(2): 123-140, 1996.

[49] Y. Freund and R.E. Schapire. Experiments with a new boosting algorithm. In: *Proceedings of International Conference on Machine Learning*, pp. 148-156, 1996.

[50] T.G. Dietterich. Ensemble methods in machine learning. In: *Proceedings of*

International Workshop on Multiple Classifier System, pp. 1-15, 2000.