International Journal of Computer Sciences and Engineering Systems
Vol. 9 No. 2 (December, 2015)

Speeding Up Isolated Vietnamese Handwritten Recognition by Combining SVM and Statistical Features

Pham Anh PHUONG[1], Ngo Quoc TAO[2] and Luong Chi MAI[2]

[1]Information Technology Department, College of Sciences, Hue University Hue, 084, Vietnam
E-mail:paphuong@yahoo.com

[2]Institute of Information Technology, Institute of Science and Technology of Vietnam, Ha Noi, 084, Vietnam
E-mail: nqtao, lcmai@ioit.ac.vn

Abstract: *Based on SVMs (Support Vector Machines) classification where statistical features are extracted, this paper proposes an efficient recognition model for isolated Vietnamese handwritten character recognition. We apply combining reduced set method and dimension reduction of input feature vectors to improve the speed of recognition model. Our test results over Vietnamese handwriting with 52,485 character samples collecting from 655 individuals show that the accuracy of our recognition model is over 90%, our improving model increase speed up to 9 times compared to original model with the acceptable accuracy.*

Keywords: *Vietnamese handwritten recognition*

1. INTRODUCTION

Handwriting recognition has been a big challenge to researchers. This problem still has no complete solution because it depends on writers, writing styles and mood of each writer. Up to now, the problem of handwriting recognition also has several satisfactory results, but only focus in the data of standard handwritten numbers such as USPS and MNIST [6, 7, 10] and other results focus in the Grecian, Latin alphabet…[3,11]. Especially, recognition of Vietnamese handwriting faces many difficulties due to a presence of different diacritical marks with the same similar shape. Hence, there are very few works on Vietnamese handwriting recognition [4, 5, 9].

Vietnamese alphabet includes a set of mark-less letters {A, B, C, D, Đ, E, G, H, I, K, L, M, N, O, P, Q, R, S, T, U, V, X, Y}, and a set of diacritical mark letters that includes vowels with marks represented different pronunciation tones {Ă, Â, À, Ả, Ã, Á, Ạ, Ằ, Ẳ, Ẵ, Ắ, Ặ, Ầ, Ẩ, Ẫ, Ấ, Ậ, Ê, Ề, Ể, Ễ, Ế, Ệ, Ề, Ể, Ễ, Ế, Ệ, Ì, Ỉ, Ĩ, Í, Ị, Ô, Ơ, Ò, Ỏ, Õ, Ó, Ọ, Ồ, Ổ, Ỗ, Ố, Ộ, Ờ, Ở, Ỡ, Ớ, Ợ, Ư, Ù, Ủ, Ũ, Ú, Ụ, Ừ, Ử, Ữ, Ứ, Ự, Ỳ, Ỷ, Ỹ, Ý, Ỵ}. We limit the scope of the problem in some regulations such as handwritten characters, it must have relatively equal sizes, characters , are not too much slant, there must be separation between mark part and character part. Support Vector Machine is the state-of-art learning machine that has been very fruitful not only in pattern recognition, but also in data mining areas, such as text classification, image, speech recognition... SVMs were already used to recognize handwritten [6, 7, 8, 10, 11], so that we use SVMs classifier using specific feature extraction methods, which are efficient for each character part and mark part to recognize isolated Vietnamese handwriting. However, SVMs are currently considerably slower in test phase than other approaches with similar generalization performance. Hence, we apply combining reduced set method [1] and dimension reduction for input feature vectors to improve the speed of recognition model. Based on connected regions of image we divide Vietnamese letters into three groups and split mark characters to isolated parts. Then, we use SVM classifier using specific feature extraction methods, which are efficient for each character part and mark part to recognize isolated Vietnamese handwriting. Our test results show that the accuracy of our recognition model is relatively high.

The rest of this paper has the following structure: Section 2 presents architecture of the recognition model, specific extraction methods, SVM classifying machines serving recognition. Section 3 proposes a solution to reduce dimension of input feature vectors and apply the reduced set method to improve the speed of recognition model. Section 4 is test results over data of isolated Vietnamese handwriting, which we collected. The last section is conclusion.

2. MODEL OF ISOLATED VIETNAMESE HANDWRITTEN RECOGNITION

In this section, we present the architecture of model of isolated Vietnamese handwritten in detail (Figure 1).

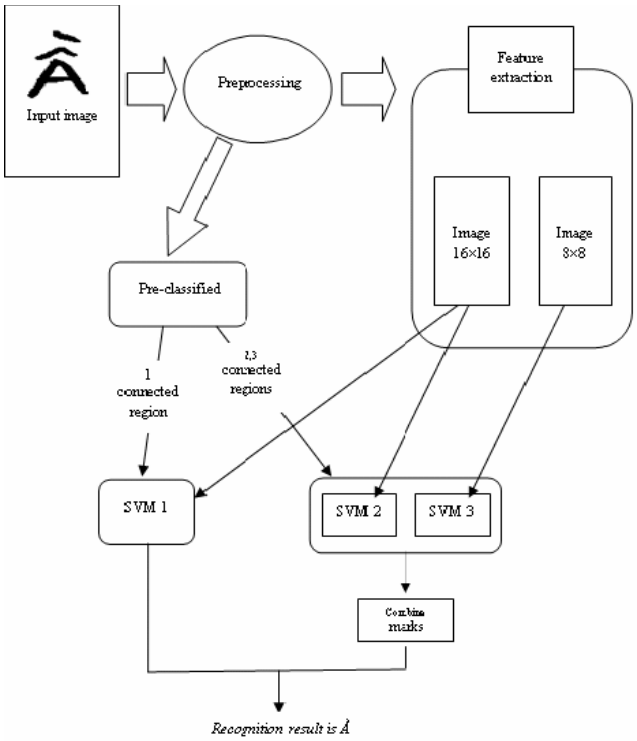
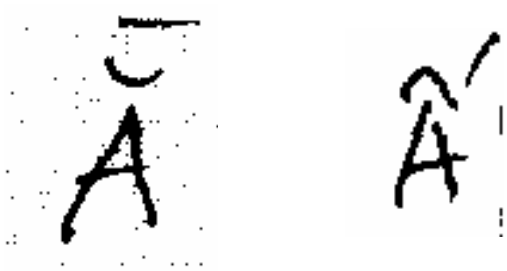


Figure 1: Architecture of Recognition System

2.1 Preprocessing

The target of preprocessing phase is to increase the accurate degree of recognition system; it uses several techniques such as image binarization that converts gray-scale image into binary one; noise filtering techniques to eliminate some kind of noise, such as mean filter and medium filter to deal with spot noise, the method of eliminating small connected regions with line noise… Figure 2 shows an example of noises that often appear after scanning process.



(a) Spot noise (b) Line noise

Figure 2: Some Frequent Noises after Scanning Image

Normalizing Image According to Connected Region

Image normalization is to create favorable conditions for the phase of dividing image into character part and mark parts.

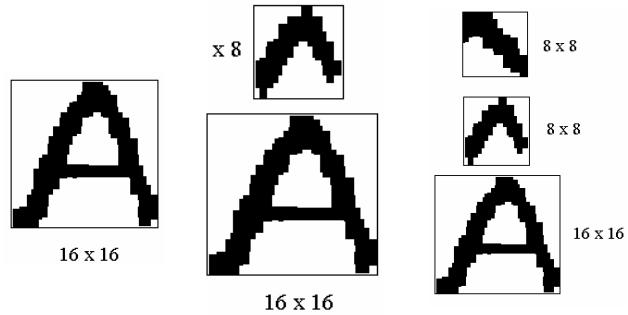
Step 1: Image normalization bases on the determination of connected regions (Figure 3).

Step 2: Arranging connected regions according to top - down order (Figure 3b).

(a) Root image (b) Determine connected regions.

Figure 3: Image Normalization

Step 3:



(a) (b) (c)

Figure 4: Normalizing Connected Regions

– If the image has only one connected region, normalize the image to the size of 16´16 (Figure 4a).
– If the image has two connected regions, let S(i) be the region of ith connected region.

If S(1) > S(2), the sign of the second connected region will be the drop tone mark (.) and we only normalize the first connected region to the size of 16×16.

Otherwise, divide the image into two parts: character part and mark part. Normalize the character part to the size of 16×16 and the mark part to the size of 8×8 (Figure 4b).

– If the image has three connected regions:

If $S(3) = \text{Min}\{S(i)\}$, the sign of this connected region will be the drop tone mark (.). Hence, we only normalize the first connected region to the size of 8×8 and the second connected region to the size of 16×16.

Otherwise, divide the image into three parts based on connected regions. Normalize the first and the second connected region to the size of 8×8 and the third connected region to the size of 16×16 (Figure 4c).

2.2 Feature Extraction

Statistical features that we extract from character image are: projection histograms, contour profiles and zones [6,7,11].

Projection histograms: The basic idea of method of feature extraction is to project black points in the two dimensional image according to the directions of horizontal, vertical, and two diagonals so that forming one dimensional sequence of signals. An advantage of this method is that it does not depend on noises; however, it still depends on the slant of characters.

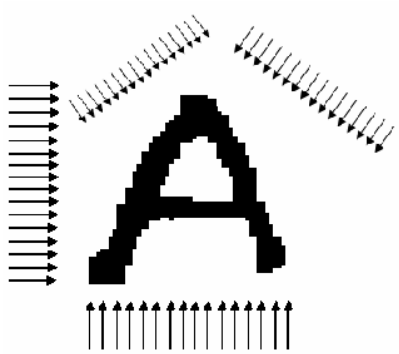


Figure 5: Extract Projection Histograms of Horizontal, Vertical, 2 Diagonals

In practice, with 16×16 image, there are 16 horizontal + 16 vertical + 2×16 diagonal = 64 features. With 8×8 image, there are 8 horizontal + 8 vertical + 2×8 diagonal = 32 features.

Zoning: The character image is divided into N´N zones. The total black points of each zone will be selected to form vector of features.

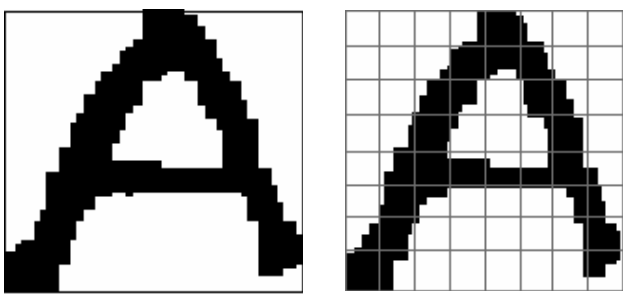


Figure 6: Extraction Feature of Zone

In practice, with 16×16 image, we select N = 8, hence, there are 8×8 = 64 features. With 8×8 image, we select N = 4, hence, there are 4×4 = 16 features.

Contour profiles: The profile is the distance from the image border to the first black point of the character in a same scanning line. The profiles well describes parts outside the character and allows to distinguish a big amount of characters.

In practice, with 16×16 image, there are 16 left + 16 right + 16 top + 16 bottom = 64 features. With 8×8 image, there are 8 left + 8 right + 8 top + 8 bottom = 32 features.

Figure 7: Extract the Contour Profiles of the Character

2.3 Preliminary Classification

Based on connected regions, we divide Vietnamese alphabet into 3 groups.

- Group 1: the group has one connected region {A, B, C, D, Đ, E, G, H, I, K, L, M, N, O, P, Q, R, S, T, U, V, X, Y, Ơ, Ư}.
- Group 2: the group has two connected regions {Ă, Â, À, Ả, Ã, Á, Ạ, Ê, È, Ẻ, Ẽ, É, Ẹ, Ì, Ỉ, Ĩ, Í, Ị, Ô, Ò, Ỏ, Õ, Ó, Ọ, Ờ, Ở, Ỡ, Ớ, Ợ, Ù, Ủ, Ũ, Ú, Ụ, Ừ, Ử, Ữ, Ứ, Ự, Ỳ, Ỷ, Ỹ, Ý, Ỵ}.
- Group 3: the group has three connected regions { Ằ, Ẳ, Ẵ, Ắ, Ặ, Ầ, Ẩ, Ẫ, Ấ, Ậ, Ề, Ể, Ễ, Ế, Ệ, Ồ, Ổ, Ỗ, Ố, Ộ}

2.4 Building SVM Classifiers

In experimental, we shall build up three SVM classifiers that cooperatively use three sets of features to train for classification and recognition.

With 16×16 character part, there are totally 196 features. With 8×8 mark part, there are 80 extracted features.

- SVM1: classifies group which has one connected region {A, B, C, D, Đ, E, G, H, I, K, L, M, N, O, P, Q, R, S, T, U, V, X, Y, Ó, Ú}.
- SVM2: With characters having mark, the character part is all vowels. Hence, this machine only classifies vowels {A, E, I, O, U, Y}.
- SVM3: classifies diacritical marks such as {/, \, ?, ~, ^, ∨} (mark of high tone, mark of falling tone, question mark, tilde, ^ sign, ã sign).

Three SVM classifiers SVM1, SVM2, SVM3 were built up from binary SVMs. We choose Sequential Minimal Optimization (SMO) algorithm [1] to train binary classifiers with the strategy One Versus One (OVO) and with the parameter C = 100. We use Gaussian function ($\sigma = 0.5$) as multiplicative function.

3. IMPROVING SPEED OF RECOGNITION MODEL

In this section, we reduce dimension of input feature vectors and apply the reduced set method [1] to improve the speed of recognition model.

3.1 Dimension Reduction of Feature Vectors

Instead of using features of the projection histogram and contour profile, we choose alternately features to reduce a half of dimension of input vectors for each row, column or diagonal. Therefore, there are only 128 features for input vector with 16×16 image and 48 features for input vector with 8×8 image.

3.2 Improving Speed of SVM Classifiers

SVMs work in feature space indirectly via a kernel function $K(x, y) = <\Phi(x)\cdot\Phi(y)>$, where Φ: $R^d \rightarrow F$ *is a* map from the *d*-dimensional input space to a possibly high-dimensional

feature space. For a two-class classification problem, the decision rule takes the form:

$$y = sign\left(\sum_{i=1}^{Ns} \alpha_i K(x_i, z) + b\right) \quad (1)$$

where α_i is weights of support vectors x_i, x is the input vector needed to classify, $K(x\, x_i) = <\Phi(x)\cdot\Phi(x_i)>$ is a kernel function calculating the dot product of two vectors $\Phi(x)$ and $\Phi(x_i)$ in the feature space, b is the bias, and N_S is the number of support vectors. The task of the SVMs training process is to determine all the parameters (x_i, α_i, b, N_S); the resulting x_i, $i = 1...N_S$ are a subset of the training set and are called *support vectors*.

The main idea of reduced set method is then to choose the smallest $N_Z < N_S$, and corresponding reduced set $\{(z_i, \beta_i)\}$, $i = 1...N_Z$, such that minimize $\rho = |\Psi' - \Psi|$, that mean try to approximate the normal vector Ψ of the separating hyperplane by a reduced set expansion Ψ' such that any resulting loss in generalization performance remains acceptable. Ψ and Ψ' are given by

$$\Psi = \sum_{i=1}^{Ns} \alpha_i \Phi(x_i) \text{ and } \Psi' = \sum_{j=1}^{Ns} \beta_i \Phi(z_j)$$

where $N_Z < N_S$, $z_i \in R^d$, $\beta_i \in R$.

To classify a new test point x, calculation (1) is replaced by

$$y = sign\left(\sum_{i=1}^{Ns} \beta_i K(x, z_i) + b\right) \quad (2)$$

In this paper we apply Bottom-Up method is described in (Nguyen Duc Dung and Ho Tu Bao, 2005) [8]. Bottom-Up approach iteratively considered and replaced two nearest support vectors belonging to the same class by a newly constructed vector. This approach leads to a conceptually simpler and computationally less expensive method, the local extremum problem does not exist, and it also makes the vectors in the simplified solution look more meaningful.

4. THE EXPERIMENTAL RESULTS

Our data of Vietnamese handwritten samples were collected from 655 individuals in which the majority comes from students. Each person wrote about 200 upper case characters, these characters are separate. We chose 52,485 samples for experiment that includes 2,485 mark samples and 20,925 marks - less characters in which 13,782 mark-less characters were used for training, the rest were for recognition.

We built three datasets serving for training three coresponding SVM classifiers:

- TrainData1: A set of marks of Vietnamese language {/, \, ?, ~, ^, ∨}, with 2,485 samples.
- TrainData2: A set of vowels in Vietnamese language {A, E, I, O, U, Y}, with 4,128 sample are extracted from 13,782 mark-less characters.
- TrainData3: A set of mark-less Vietnamese letters{A, B, C, D, Đ, E, G, H, I, K, L, M, N, O, P, Q, R, S, T, U, V, X, Y, Ó, Ú}, with 13,782 samples.

Figure 8: Samples Extracted from Set of Vietnamese Handwritten Characters

The test data includes 36,218 samples, in which 7143 mark-less and 29,075 mark characters. To evaluate accurately each group of letters in the connected regions, we constructed five datasets as the following:

- TestData1: A set of Vietnamese letters which have one connected region, with 7,143 samples.
- TestData2: A set of vowels in Vietnamese language, with 1,500 sample are extracted from 7,143 mark-less characters.
- TestData3: A set of Vietnamese letters which have two connected regions, with 16,856 samples.
- TestData4: A set of Vietnamese letters which have three connected regions, with 12,219 samples.
- TestData5 = TestData1 È TestData3 È TestData4.

Table 1
Test Results from Data Sets of Vietnamese Handwritten Characters

Dataset	*Samples*	*Accuracy*
TestData1	7143	84.26%
TestData2	1500	99.67%
TestData3	16856	92.18%
TestData4	12219	91.85%
TestData5	36218	90.51%

The test results in Table 1 showed that the accuracy of TestData2 give us the higher accuracy than that of TestData1. The reason for that, while TestData2 only includes six vowels with distinguish shape so it has less mistaken, the TestData1 must be differentiate too many the letters with the similar shapes ({B,Đ},{C,G},{U,U,V},...) so it is more likely to be mistaken, which results in low accuracy. Since the separation and combination of the marks are not perfect, the accuracy of TestData3 and TestData4 is not higher than TestData2. The final test result of TestData5 showed that the combination SVMs classifiers and statistical

features into isolated Vietnamese handwritten recognition have relatively high accuracy.

Table 2 presents test results of four methods: Original method (ORG), dimension reduction method (DRM), reduced set method (RSM), combining both DRM and RSM.

Table 2
Test Results on TestData4

Method	*Accuracy*	*Testing (secs)*
ORG	90.51%	2985
RDM	90.51%	2279
RSM	90.38%	763
RDM+RSM	90.38%	338

The test results in Table 2 showed that with the original model, it will take 2.7 minutes to recognize one A4 page (about 2000 letters) which is so slow for a recognition system, while the improvement model will take 19 seconds for one A4 page (not include the time of letter cutting). It is 9 times faster than the original model with acceptable accuracy.

5. CONCLUSION

The paper proposed an efficient model for isolated Vietnamese handwritten character recognition. Our recognition model was built up from SVM classifiers combining with specific statistical feature extraction. Our test results showed that the accuracy of our recognition model is over 90%. Base on reduced set method combining with dimension reduction of feature vectors, our improving model increase speed up to 9 times compared to original model with acceptable accuracy.

REFERENCES

[1] Burges, C. J. C. "Simplified Support Vector Decision Rules", *Proc. 13th International Conference on Machine Learning,* San Mateo, CA, 1996, 71–77.

[2] J. Platt, "Fast Training of Support Vector Machines Using Sequential Minimal Optimization", *In Advences in Kernel Methods-Support Vector Learning,* Cambridge, M.A, MIT Press, 1999, 185-208.

[3] Le Hoai Bac, Le Hoang Thai, "Neural Network & Genetic Algorithm in Application to Handwritten Character Recognition", *Journal of Computer Science and Cybernetics*, Vietnamese Academy of Science and Technology, **17**(4), 2001, 57-65.

[4] Nguyen Huu Hoa, Luong Chi Mai, "Recognition of Vietnamese Handwritten Isolated Characters using Fuzzy Neural Network", *Proceeding of IOIT' Workshop on R&D*, 12/2001, pp. 623-631 (in Vietnamese).

[5] Vu Hai Quan, Pham Nam Trung, Nguyen Duc Hoang Ha, "A Robust Method for the Vietnamese Handwritten and Speech Recognition", 16th International Conference on Pattern Recognition, ICPR 2002, *IEEE Computer Society, Quebec, Canada*, **3**, 732-735.

[6] Gorgevik D., Cakmakov D., "An Efficient Three-Stage Classifier for Handwritten Digit Recognition", *Proceedings of 17th Int. Conference on Pattern Recognition, ICPR 2004,* **4**, *IEEE Computer Society*, Cambridge, UK, 23-26, 2004, 507-510.

[7] Cakmakov D., Gorgevik D., "Handwritten Digit Recognition Using Classifier Cooperation Schemes", *Proceedings of the 2nd Balkan Conference in Informatics, BCI 2005*, Ohrid, 2005, 23-30.

[8] D. D. Nguyen, T. B. Ho, "An Effcient Method for Simplifying Support Vector Machine", The 22th International Conference on Machine Learning, ICML 2005, Bonn, Germany, 2005.

[9] Ngo Quoc Tao, Pham Van Hung, "Online Continues Vietnamese Handwritten Character Recognition based on Microsoft Handwritten Character Recognition Library", *IEEE Asia Pacific Conference on Circuits and Systems, APCCAS 2006*, Singapore, 2024-2026.

[10] Pham Anh Phuong, "Handwritting Recognition with a SVM Model", *Journal of Science, Hue University*, ISSN: 1859-1388, Num 42, 2007, pp. 157-165.

[11] G. Vamvakas, B. Gatos, I. Pratikakis, N. Stamatopoulos, A. Roniotis and S.J. Perantonis, "Hybrid Off-Line OCR for Isolated Handwritten Greek Characters", *The Fourth IASTED International Conference on Signal Processing, Pattern Recognition, and Applications* (SPPRA 2007), ISBN: 978-0-88986-646-1, Innsbruck, Austria, February 2007, 197-202.