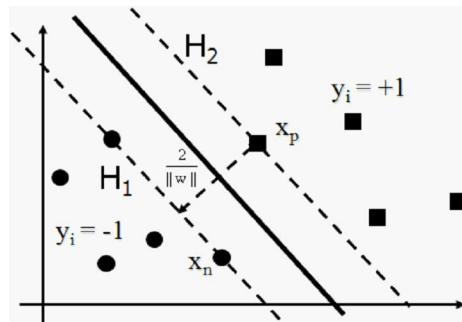


nhiều bài toán thực tế và cho nhiều kết quả rất khả quan [1, 3]. Vì vậy, chúng tôi sử dụng phương pháp véc tơ tựa cho mô hình nhận dạng mà chúng tôi đề xuất. Trong một hệ thống nhận dạng thì việc trích chọn đặc trưng là một bước quan trọng, nó có ảnh hưởng rất lớn đến độ chính xác của hệ thống nhận dạng. Có rất nhiều phương pháp trích chọn đặc trưng hiệu quả có thể áp dụng cho chữ viết tay như: ma trận trọng số, toán tử Kirsch, các biểu đồ chiều...[4, 5, 7], trong bài báo này chúng tôi đã sử dụng và cài đặt thử nghiệm trên một số các loại đặc trưng đó và quyết định sử dụng ý tưởng của phương pháp trích chọn đặc trưng wavelet Haar [8] cho mô hình nhận dạng chữ Việt viết tay rời rạc.

Tiếp theo, Mục 2 sẽ tóm tắt những ý tưởng cơ bản của phương pháp véc tơ tựa. Mục 3 trình bày các kết quả thực nghiệm trên dữ liệu chữ viết tay tiếng Việt với một số phương pháp trích chọn đặc trưng thông dụng. Mục 4 phát họa kiến trúc của mô hình nhận dạng chữ Việt viết tay rời rạc và các kết quả thực nghiệm theo mô hình này. Cuối cùng là phần kết luận và hướng phát triển.

2. PHƯƠNG PHÁP VÉCTƠ TỰA

Cho tập mẫu huấn luyện $x_i \in R^N$, $i = 1, \dots, N$ và các nhãn tương ứng $y_i \in \{-1, +1\}$, mục tiêu của SVM là tìm một siêu phẳng phân cách (được xác định bởi w) sao cho khoảng cách (margin) giữa hai lớp đạt cực đại (Hình 1).



Hình 1. Siêu phẳng tách với khoảng cách cực đại

Hàm mục tiêu của một máy phân lớp SVM nhị phân có thể được phát biểu như sau:

$$g(x) = w \cdot \Phi(x) + b, \quad (1)$$

trong đó, véc tơ đầu vào $x \in R^N$, w là véc tơ chuẩn của siêu phẳng phân cách trong không gian đặc trưng được sinh ra từ ánh xạ của hàm $\Phi(x) : R^N \rightarrow R^M$ ($M > N$, $\Phi(x)$ có thể tuyến tính hoặc phi tuyến) và b là độ lệch so với gốc tọa độ [1]. SVM gốc được thiết kế cho bài toán phân lớp nhị phân, vì vậy dấu của $g(x)$ cho biết véc tơ x thuộc lớp $+1$ hay lớp -1 .

Việc tìm siêu phẳng phân cách chính là việc giải bài toán qui hoạch toàn phương (QP-Quadratic programming):

$$\max_{\alpha} \left(\alpha_1 - \frac{1}{2} \alpha^T H \alpha \right) \quad (2)$$

thỏa mãn $0 \leq \alpha_i \leq C$, $\forall i$ và $\sum_{i=1}^N \alpha_i y_i = 0$, trong đó $\alpha = [\alpha_1, \dots, \alpha_N]^T$ và \mathbf{H} là ma trận kích thước $N \times N$ được gọi là ma trận nhân (kernel matrix) với mỗi phần tử $\mathbf{H}(i, j) = y_i y_j \Phi(x_i) \cdot \Phi(x_j)$.

Giải bài toán QP (3) ta thu được:

$$w = \sum_{i=1}^N \alpha_i y_i \Phi(x_i). \quad (3)$$

Mỗi mẫu huấn luyện x_i tương ứng với một hệ số Lagrange α_i . Các mẫu có $\alpha_i > 0$ được gọi là véc tơ hỗ trợ.

Thế (3) vào (1), ta có:

$$g(x) = \sum_{i=1}^N \alpha_i y_i \Phi(x_i) \cdot \Phi(x) + b. \quad (4)$$

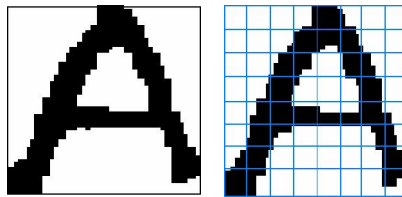
Giả sử $\Phi(x_i) \cdot \Phi(x_j) = K(x_i, x_j)$. Nghĩa là, tích vô hướng trong không gian đặc trưng tương đương với một hàm nhân K của không gian đầu vào. Vì vậy, ta không cần phải tính trực tiếp các giá trị $\Phi(x_i)$, $\Phi(x_j)$ mà chỉ cần tính tích vô hướng $\langle \Phi(x_i), \Phi(x_j) \rangle$ gián tiếp thông qua hàm nhân $K(x_i, x_j)$. Như vậy, ta có hàm mục tiêu cho bài toán phân lớp SVM có dạng

$$g(x) = \sum_{i=1}^N \alpha_i y_i K(x_i, x) + b. \quad (5)$$

3. NHẬN DẠNG CHỮ VIẾT TAY VỚI CÁC ĐẶC TRƯNG THÔNG DỤNG

Trong phần này, chúng tôi giới thiệu một số đặc trưng đã được sử dụng rất hiệu quả trong các bài toán nhận dạng chữ viết tay [4, 5, 7].

3.1. Trọng số vùng (Zoning)

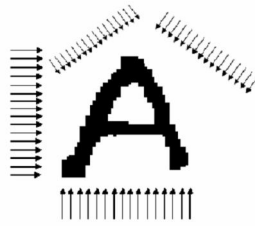


Hình 2. Trích chọn đặc trưng trọng số vùng

Ảnh ký tự được chia thành $N \times N$ vùng (zones). Tổng số điểm đen của mỗi vùng sẽ được chọn để tạo thành véc tơ đặc trưng.

Trong thực nghiệm, với ảnh kích thước 16×16 , chúng tôi chọn $N = 8$, như vậy có $8 \times 8 = 64$ đặc trưng.

3.2. Biểu đồ chiếu (Projection histograms)

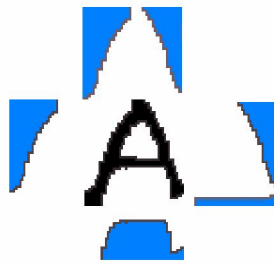


Hình 3. Trích chọn các biểu đồ chiếu ngang, dọc và 2 đường chéo

Ý tưởng cơ bản của phương pháp trích chọn đặc trưng này là chiếu các điểm đen trên ảnh 2 chiều theo các hướng ngang, dọc và hai đường chéo thành một dãy các tín hiệu 1 chiều. Ưu điểm của các đặc trưng này là không phụ thuộc vào nhiễu, tuy nhiên nó vẫn phụ thuộc vào độ nghiêng của chữ.

Trong thực nghiệm, với ảnh kích thước 16×16 , chúng tôi chọn 16 ngang + 16 dọc + 2 \times 16 chéo = 64 đặc trưng.

3.3. Trích chọn chu tuyến (Contour profiles)



Hình 4. Trích chọn các khối bên ngoài của chữ

Phần được trích chọn là khoảng cách từ biên của khung chứa ảnh tới điểm đen đầu tiên của chữ trên cùng 1 dòng quét. Phương pháp trích chọn này mô tả tốt các khối bên ngoài của chữ và cho phép phân biệt một số lượng lớn các ký tự.

Trong thực nghiệm, với ảnh kích thước 16×16 , có 16 trái + 16 phải + 16 trên + 16 dưới = 64 đặc trưng.

3.4. Trích chọn đặc trưng wavelet Haar

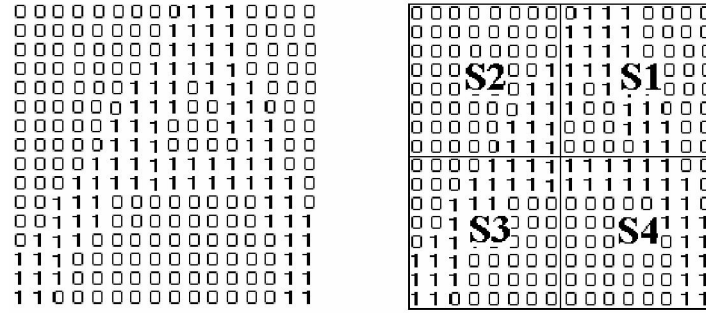
Chúng tôi sử dụng ý tưởng của phương pháp trích chọn đặc trưng wavelet Haar [8] để chọn tập đặc trưng cho mỗi ảnh ký tự đầu vào.

Từ ảnh nhị phân kích thước $2n \times 2n$ (Hình 5), quá trình trích chọn đặc trưng được mô tả theo thuật toán sau:

Procedure HaarFeature

Input Ma trận vuông (A, n) cấp 2^n .

Output Tập các đặc trưng $\{F_1, F_2, \dots, F_{2^n \times 2^n}\}$.



Hình 5. Trích chọn đặc trưng wavelet Haar

Method

1. Khởi tạo: $Queue = \Phi$; $i = 1$;
2. - Tính F_i bằng tổng các điểm đen trong toàn bộ ma trận (A, n) ;
- $PUSH((A, n), Queue)$;
3. While $Queue \neq \Phi$ Do
 - {
 - $POP(Queue, (A, n))$;
 - if $(n > 1)$
 - {
 - Chia ảnh thành 4 phần: A_1, A_2, A_3, A_4 ;
 - for $(j = 1; i = 4; j++)$
 - $PUSH((A_j, ndiv2), Queue)$;
 - Gọi $S1, S2, S3, S4$ là tổng các điểm đen tương ứng với A_1, A_2, A_3, A_4 ;
 - Tính $F_{i+1} = S1 + S2$;
 - $F_{i+2} = S2 + S3$;
 - $F_{i+3} = S4$;
 - $i = i + 3$;

Trong thực nghiệm, với phần chữ chúng tôi chọn $n = 4$, như vậy ta có: $1 + 3 + 4 \times 3 + 4 \times 4 \times 3 + 4 \times 4 \times 4 \times 3 = 256$ đặc trưng, còn với phần dấu chúng tôi chọn $n = 3$, như vậy có tất cả 64 đặc trưng.

Phương pháp trích chọn đặc trưng này sẽ tạo ra một dãy số các đặc trưng giảm dần. Với cùng một chữ thì các giá trị lớn ở đầu dãy tương đối ổn định, có thể đại diện cho hình dạng khái quát của chữ; còn các giá trị ở cuối dãy nhỏ dần và không ổn định, thể hiện sự đa dạng trong từng chi tiết của chữ.

3.5. Kết quả thực nghiệm

Dữ liệu chữ viết tay tiếng Việt của chúng tôi được thu thập từ 655 người viết khác nhau, đối tượng chủ yếu là sinh viên. Mỗi người viết khoảng 200 chữ in hoa, các ký tự được viết

rời rạc. Chúng tôi chọn lọc ra 50000 mẫu để tiến hành thực nghiệm, trong đó sử dụng 30000 mẫu dùng để huấn luyện và 20000 mẫu phục vụ cho việc nhận dạng.

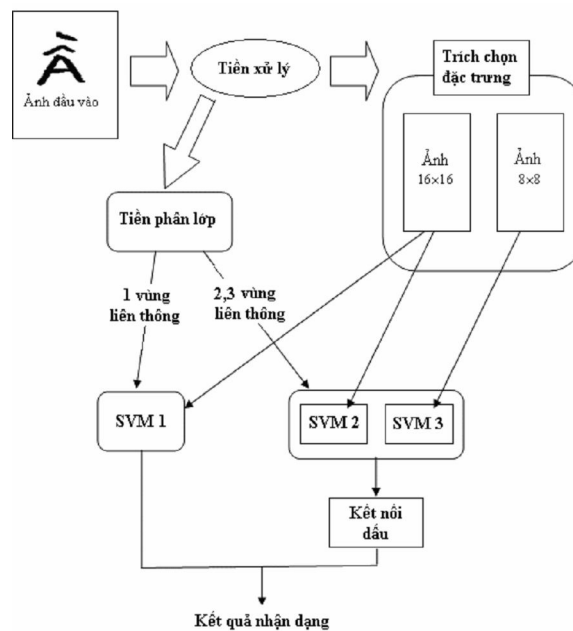
Công cụ phân lớp được xây dựng từ các SVM nhị phân. Chúng tôi chọn thuật toán SMO [1] để huấn luyện các máy phân lớp nhị phân theo chiến lược một chống một (OVO- One versus One) với các tham số $C = 100$ và sử dụng hàm nhân là hàm Gaussian với $\sigma = 0,5$.

Bảng 1. Kết quả nhận dạng theo các loại đặc trưng khác nhau

| Tập đặc trưng | Độ chính xác |
|-----------------------|--------------|
| Zoning | 75.37% |
| Projection histograms | 73.82% |
| Contour profiles | 73.14% |
| wavelet Haar | 77.28% |

Các kết quả thực nghiệm ở bảng 1 cho thấy đối với bài toán nhận dạng chữ viết tay tiếng Việt, áp dụng đặc trưng Haar cho độ chính xác cao hơn so với sử dụng các đặc trưng khác. Tuy nhiên, do bộ ký tự tiếng Việt quá đa dạng nên việc áp dụng các đặc trưng này lên tập dữ liệu viết tay tiếng Việt vẫn chưa đạt hiệu quả cao. Vì vậy cần phải có một mô hình hiệu quả cho bài toán nhận dạng chữ viết tay tiếng Việt. Trong phần sau chúng tôi sẽ đề xuất một mô hình hiệu quả cho việc nhận dạng chữ viết tay tiếng Việt.

4. MÔ HÌNH NHẬN DẠNG CHỮ VIỆT VIẾT TAY RỜI RẠC



Hình 6. Kiến trúc của hệ nhận dạng chữ viết tay tiếng Việt

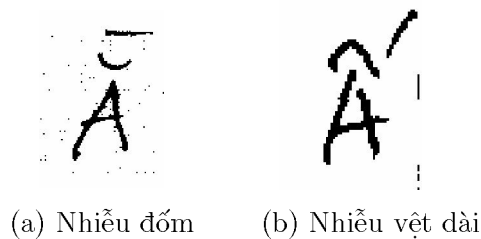
Trong phần này, chúng tôi sẽ trình bày chi tiết kiến trúc của mô hình nhận dạng chữ Việt viết tay rời rạc. Trên cơ sở các thành phần liên thông của ảnh, chúng tôi phân tập ký tự tiếng Việt thành 3 nhóm và tách các ký tự có dấu thành các phần rời nhau. Sau đó chúng tôi

xây dựng các máy phân lớp SVM để nhận dạng cho từng phần chữ và dấu, cuối cùng ghép nối các kết quả phân lớp của các phần chữ và phần dấu để có được kết quả nhận dạng cuối cùng.

4.1. Tiền xử lý

Mục đích của giai đoạn tiền xử lý nhằm tăng độ chính xác của hệ thống nhận dạng. Khi quét ảnh thường gặp các loại nhiễu, vì vậy chúng tôi sử dụng một số kỹ thuật lọc nhiễu để khử các nhiễu đốm và nhiễu vệt dài. Đối với nhiễu đốm, sử dụng các bộ lọc trung bình và lọc trung vị, còn với các nhiễu vệt dài thì chúng tôi sử dụng phương pháp khử các vùng liên thông nhỏ (Hình 7).

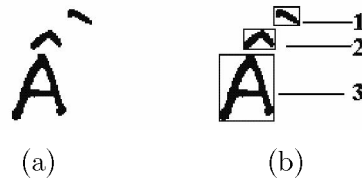
Để thuận tiện cho việc xử lý sau này, chúng tôi biến đổi ảnh đầu vào từ ảnh đa cấp xám thành ảnh nhị phân.



Hình 7. Một số nhiễu thường gặp khi quét ảnh

Chuẩn hóa ảnh theo vùng liên thông Chuẩn hóa ảnh nhằm mục đích tạo điều kiện thuận tiện cho công đoạn tách ảnh thành từng phần chữ và dấu.

Bước 1: Xác định các vùng liên thông trên ảnh (Hình 8).



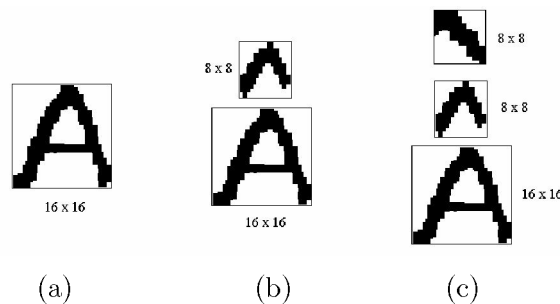
Hình 8. Chuẩn hóa ảnh

(a) Ảnh gốc

(b) Xác định các vùng liên thông và đánh thứ tự các vùng liên thông

Bước 2: Sắp xếp các vùng liên thông theo thứ tự từ trên xuống (Hình 8b).

Bước 3:



Hình 9. Chuẩn hóa các vùng liên thông

4.5. Kết quả thực nghiệm

Với 50000 mẫu mà chúng tôi thu thập để tiến hành thực nghiệm, trong đó 13782 mẫu chữ không dấu dùng để huấn luyện, phần còn lại phục vụ cho việc nhận dạng.

Ba tập dữ liệu được xây dựng phục vụ cho việc huấn luyện:

- TrainData1: Tập các dấu tiếng Việt $\{/, \backslash, ?, \sim, \wedge, \vee\}$, với 2485 mẫu.
- TrainData2: Tập các chữ cái nguyên âm tiếng Việt $\{A, E, I, O, U, Y\}$, với 4128 mẫu.
- TrainData3: Tập các chữ cái tiếng Việt không dấu $\{A, B, C, D, Đ, E, G, H, I, K, L, M, N, O, P, Q, R, S, T, U, V, X, Y, Ơ, Ư\}$, với 13782 mẫu.



Hình 10. Các mẫu trích từ tập ký tự viết tay tiếng Việt

Bốn tập dữ liệu được xây dựng phục vụ cho việc nhận dạng, kết quả thực nghiệm thể hiện ở Bảng 1:

- TestData 1: Tập các ký tự tiếng Việt có 1 vùng liên thông, với 7143 mẫu.
- TestData 2: Tập các ký tự tiếng Việt có 2 vùng liên thông, với 16856 mẫu.
- TestData 3: Tập các ký tự tiếng Việt có 3 vùng liên thông, với 12219 mẫu.
- TestData 4 = TestData 1 \cup TestData 2 \cup TestData 3.

Bảng 2. Kết quả nhận dạng trên các tập dữ liệu tiếng Việt viết tay rời rạc

| Tập mẫu | Số mẫu | Độ chính xác |
|------------|--------|--------------|
| TestData 1 | 7143 | 82.24% |
| TestData 2 | 16856 | 90.69% |
| TestData 3 | 12219 | 87.78% |
| TestData 4 | 36218 | 88.04% |

Các kết quả thực nghiệm ở Bảng 2 cho thấy mô hình nhận dạng chữ Việt viết tay rời rạc mà chúng tôi đề xuất đạt độ chính xác cao hơn rất nhiều so với các kết quả thực nghiệm ở Bảng 1.

5. KẾT LUẬN

Bài báo đã sử dụng một số phương pháp trích chọn đặc trưng để áp dụng vào bài toán nhận dạng chữ viết tay rời rạc. Các kết quả thực nghiệm trên dữ liệu chữ viết tay tiếng Việt cho thấy áp dụng đặc trưng wavelet Haar cho độ chính xác cao nhất trong số các đặc trưng được lựa chọn. Vì vậy chúng tôi đã đề xuất mô hình nhận dạng chữ Việt viết tay rời rạc dựa trên cơ sở phương pháp véc tơ tựa kết hợp với trích chọn đặc trưng wavelet Haar. Các

kết quả thực nghiệm của chúng tôi trên mô hình nhận dạng chữ Việt viết tay rời rạc đạt độ chính rất khả quan. Trong tương lai, chúng tôi tiếp tục nghiên cứu phần hậu xử lý, áp dụng mô hình ngôn ngữ N-Gram để cải thiện độ chính xác của mô hình nhận dạng.

TÀI LIỆU THAM KHẢO

- [1] J. Platt, Fast training of support vector machines using sequential minimal optimization, *Advances in Kernel Methods - Support Vector Learning*, Cambridge, M.A, MIT Press, 1999 (185–208).
- [2] Lê Hoài Bắc, Lê Hoàng Thái, Neural network & genetic algorithm in application to handwritten character recognition, *Tạp chí Tin học và Điều khiển học* **17** (4) (2001) 57–65.
- [3] Chih-Chung Chang and Chil-Jen Lin, “LIBSVM: a Library for Support Vector Machines”, National Taiwan University, 2004.
- [4] D. Gorgevik, D. Cakmakov, An efficient three-stage classifier for handwritten digit recognition, *Proceedings of 17th Int. Conference on Pattern Recognition, ICPR2004*, Vol. 4, IEEE Computer Society, Cambridge, UK, 23-26 August 2004 (507–510).
- [5] D. Cakmakov, D. Gorgevik, Handwritten digit recognition using classifier cooperation schemes, *Proceedings of the 2nd Balkan Conference in Informatics, BCI 2005*, Ohrid, November 17-19, 2005 (23–30).
- [6] Phạm Anh Phương, Nhận dạng chữ viết tay hạn chế với mô hình SVM, *Tạp chí khoa học Đại học Huế* (42) (2007) 157–165.
- [7] 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).
- [8] P. Viola, M. Jones, Rapid object detection using a boosted cascade of simple features, *Proc. Intl. Conf. on Computer Vision and Pattern Recognition (CVPR)* **1**, Kauai, HI, USA, 2001 (511–518).

Nhận bài ngày 3 - 3 - 2008

Nhận lại sau sửa ngày 6 - 2 - 2009