NHẬN DẠNG DẤU THANH VÀ MŨ TRONG KÝ TỰ TIẾNG VIỆT VIẾT TAY

RECOGNIZING ACCENT IN VIETNAMESE HANDWRITING CHARACTERS

Huỳnh Hữu Hưng, Nguyễn Trọng Nguyên

Trường Đại học Bách khoa, Đại học Đà Nẵng; Email: hhhung@dut.udn.vn, ntnguyen.dn@gmail.com

Tóm tắt – Nhận dạng ký tự viết tay là chủ đề đang được nghiên cứu rộng rãi hiện nay. Nhiều giải pháp được áp dụng cho ký tự tiếng Anh và mang lại hiệu quả cao. Tuy nhiên, bài toán tiếng Việt đang đối mặt với nhiều thách thức, phần lớn vấn đề đến từ dấu mũ và dấu thanh (gọi chung là dấu). Bài báo này đề xuất một giải pháp nhận dạng dấu, kể cả trong trường hợp mũ và thanh trong ký tự dính liền nhau – một vấn đề rất phổ biến và ảnh hưởng đáng kể đến kết quả nhận dạng. Chúng tôi bắt đầu với việc tách riêng dấu và chữ cái bằng kĩ thuật gán nhãn vùng liên thông. Tiếp theo, dấu được phân vào hai lớp: dấu đơn (một dấu mũ hoặc dấu thanh) hoặc dấu kép (mũ và thanh dính liền nhau). Dấu kép (nếu có) tiếp tục được tách ra thành các dấu đơn để chuẩn bị cho bước xử lý cuối cùng. Quá trình nhận dạng được thực hiện bằng phương pháp mô hình Markov ẩn. Giải pháp đề xuất đã được thử nghiệm và thu được kết quả khả quan.

Từ khóa – ký tự viết tay tiếng việt; dấu mũ; dấu thanh; bộ phát hiện góc; tách nhánh; mô hình Markov ẩn.

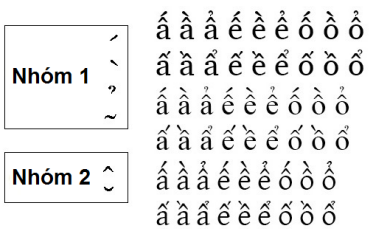
Abstract – Handwriting character recognition is one of the most common research topics. Many approaches have applied to English characters and achieve high accuracy. However, the complexities in the language of each country are not same. Recognizing Vietnamese handwriting character is facing many problems, most of them come from the accent. This paper focuses on accent recognition, especially when there is a connection between two accents - a common problem which affects the identification result. Our approach starts with separating accent from character using the connected-component labeling method. The obtained accent then is checked if it is single or multiple (the combination of many accents). In the second case, each accent is extracted for the last processing step. The recognition is performed using hidden Markov models with each single accent. Proposed solution has been tested and obtained high accuracy.

Key words – vietnamese handwriting character; accent; corner detector; branch separating; hidden Markov models.

1. Đặt vấn đề

Chủ đề nhận dạng ký tự viết tay đã được nghiên cứu trong suốt hơn 40 năm qua. Ngôn ngữ được hướng đến nhiều nhất là tiếng Anh, trong đó mỗi ký tự chỉ gồm một chữ cái, do đó kết quả thường cao hơn các ngôn ngữ khác. Phần lớn các nghiên cứu đều tập trung vào bảng chữ cái Latinh hơn là ký tự có dấu, nguyên nhân là bởi độ phức tạp trong cách biểu diễn và nhận dạng dấu. Tiếng Việt là một ngôn ngữ phức tạp với mỗi ký tự có thể chứa một hoặc hai dấu. Do đó, kết quả nghiên cứu trên tiếng Việt khá hạn chế, độ chính xác chỉ cao khi ký tự được viết một cách lý tưởng, nghĩa là các dấu trong một ký tự được tách bạch rõ ràng. Phương pháp do chúng tôi đề xuất hướng đến việc giải quyết vấn đề tồn tại sự liên kết của các dấu trong một ký tự, được thực hiện bởi sự kết hợp của nhiều kĩ thuật xử lý ảnh trên miền không gian.

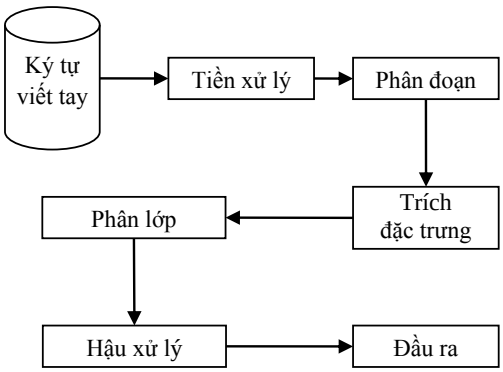
Dấu trong tiếng Việt được chia thành hai nhóm (mũ và thanh) như trong Hình 1, trong đó dấu nặng không được đề cập vì nó xuất hiện bên dưới chữ cái chính trong ký tự, nên không có sự liên kết với bất kỳ dấu khác.



Hình 1: *Các dấu được nghiên cứu và một số ví dụ về ký tự có sự xuất hiện nhiều dấu*

2. Các nghiên cứu liên quan

Hiện nay, nhiều hướng tiếp cận cho vấn đề nhận dạng ký tự viết tay bằng kĩ thuật xử lý ảnh đã được đưa ra. Mục tiêu chung được đặt ra là làm cho máy tính thu nhận và hiểu được ký tự viết tay đầu vào. Mô hình tổng quan của một hệ thống nhận dạng ký tự viết tay được trình bày trong Hình 2.



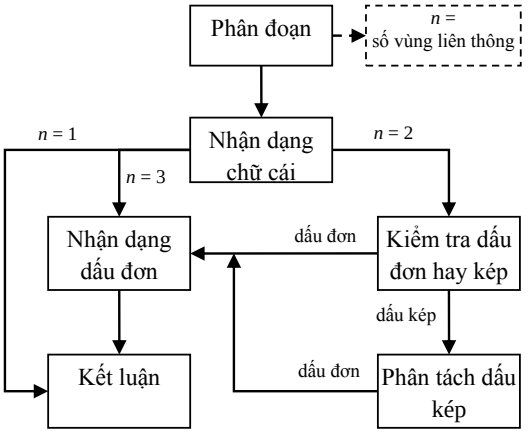
Hình 2: *Tổng quan hệ thống nhận dạng ký tự viết tay*

Hai nghiên cứu gần đây về nhận dạng ký tự có dấu được trình bày trong [1][2]. Sau khi thực hiện bước phân đoạn để thu được dấu, các tác giả sử dụng 45 đặc trưng và nhận dạng bằng máy vectơ tựa. Kết quả thử nghiệm cho thấy giải pháp này đạt hiệu quả khá cao và ổn định. Tuy nhiên, cách viết ký tự được đề cập trong 2 bài báo này quá lý tưởng, cụ thể là các dấu được viết tách biệt, điều này không phải luôn đúng trong các văn bản viết tay. Ngoài ra, hệ thống cần nhiều thời gian để tính toán bởi số lượng đặc trưng lớn và tương đối phức tạp.

Giải pháp được đề xuất trong bài báo này hướng đến khả năng giải quyết cả hai vấn đề trên: nhận dạng dấu kép và yêu cầu chi phí tính toán thấp hơn.

3. Phương pháp đề xuất

Nội dung phần này trình bày cụ thể từng bước xử lý để nhận dạng dấu trong ký tự tiếng Việt. Mô hình tổng quan của giải pháp đề xuất được thể hiện trong Hình 3.



Hình 3: *Tổng quan giải pháp đề xuất*

Đầu tiên, ký tự đầu vào được tách riêng dấu và chữ cái chính bằng phương pháp gán nhãn vùng liên thông, trong đó chữ cái nằm bên dưới dấu. Số lượng vùng lớn nhất có thể trong ký tự tiếng Việt là 3, bao gồm 1 chữ cái và 2 dấu. Một số ví dụ cụ thể cho từng trường hợp là: 'o', 'ó', và 'ố' tương ứng với 1, 2, và 3 vùng.

Trường hợp đơn giản nhất là ký tự chỉ bao gồm 1 vùng. Khi đó, ký tự được mặc nhiên xem như một chữ cái. Các phương pháp nhận dạng áp dụng cho chữ viết tay tiếng Anh có thể được dùng để xác định ký tự. Với trường hợp ký tự gồm 3 thành phần, ta có hai dấu đơn và một chữ cái cần nhận dạng. Quá trình nhận dạng dấu đơn được thực hiện bởi phương pháp mô hình Markov ẩn, sẽ được trình bày trong phần sau.

Trường hợp còn lại, tương ứng với ký tự gồm hai thành phần, có độ phức tạp cao nhất bởi vùng dấu có thể là một dấu đơn duy nhất hoặc hai dấu đơn dính liền nhau. Để giải quyết vấn đề này, chúng tôi thực hiện kiểm tra và tách dấu thành các dấu đơn nếu dấu đang xét là dấu kép. Sau đó, các dấu đơn này được nhận dạng bằng phương pháp mô hình Markov ẩn như đã đề cập.

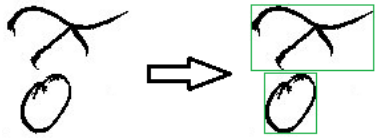
Các bước xử lý cụ thể trong Hình 3 được trình bày trong nội dung tiếp theo.

3.1. Tiền xử lý

Nhiễu thường xuất hiện khi sử dụng thiết bị quét quang học làm giảm độ chính xác khi nhận dạng. Ngoài ra, nhiễu cũng có thể xuất hiện trên ảnh trong quá trình thu nhận và truyền dẫn. Do đó, khử nhiễu là bước cần thiết trước khi thực hiện các công đoạn xử lý chính. Chúng tôi sử dụng bộ lọc trung vị có mặt nạ kích thước 3×3 kết hợp với các phép biến đổi hình thái là đóng và mở ảnh [3].

3.2. Phân đoạn

Đây là bước xử lý quan trọng vì ký tự tiếng Việt thường bao gồm một chữ cái cùng với một hoặc hai dấu, ví dụ như 'á', 'â', 'ă', 'ấ' với chữ cái tương ứng là 'a'. Đầu tiên, phương pháp gán nhãn vùng liên thông [4] được sử dụng để tách riêng từng phần trong ký tự. Sau đó, các phần này được phân vào 2 lớp, tương ứng với dấu và chữ cái, dựa trên vị trí của mỗi phần với đối tượng nằm dưới cùng là chữ cái. Một ví dụ minh họa được trình bày trong Hình 4.



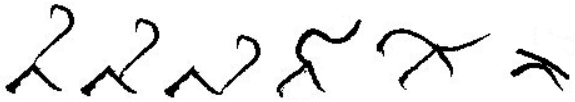
Hình 4: *Ký tự 'ỗ' với chữ cái 'o', dấu '∧' và '∼'*

3.3. Nhận dạng chữ cái

Các chữ cái riêng biệt trong bảng chữ tiếng Việt có thể được nhận dạng hiệu quả thông qua các giải pháp được áp dụng cho bảng chữ tiếng Anh. Bước này không nằm trong nội dung trọng tâm mà bài báo muốn đề cập nên phần này không được trình bày chi tiết.

3.4. Kiểm tra dấu "đơn" hay "kép"

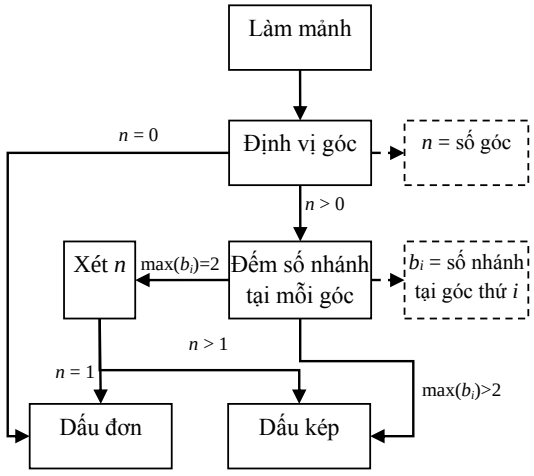
Một số trường hợp phổ biến của dấu kép được thể hiện trong Hình 5.



(a) (b) (c) (d) (e) (f)

Hình 5: *Một số ví dụ về sự liên kết giữa các dấu gồm '?', '~' và '^'*

Ta thấy rằng mỗi dấu kép ở Hình 5 tồn tại một góc với 3 (a, b, d) hoặc 4 (e, f) nhánh tương ứng, hoặc hai góc với mỗi góc gồm 2 nhánh (c), trong khi mỗi dấu đơn chỉ có thể xuất hiện tối đa một góc với 2 nhánh (dấu '∧'). Do đó, một dấu là "đơn" hay "kép" có thể được xác định dựa trên đặc điểm này. Quá trình xử lý tương ứng được trình bày chi tiết trong Hình 6.



Hình 6: *Quá trình kiểm tra dấu*

Bộ phát hiện (định vị) góc được xây dựng dựa trên toán tử Plessey được mô tả trong [6]. Số lượng nhánh tại mỗi góc được tính bằng một phương pháp đơn giản. Chúng tôi định nghĩa một đường tròn có bán kính 3 điểm ảnh và tâm đặt tại góc. Tiếp theo ta xác định các điểm ảnh của đối tượng nằm trên đường tròn. Lúc này, dấu đã được làm mảnh nên số điểm ảnh được xác định tương ứng với số nhánh tại góc đang xét.

Nếu dấu được xác định là dấu đơn, việc nhận dạng được thực hiện ngay sau đó với các mô hình Markov ẩn. Ngược

lại, dấu kép được phân tách thành các dấu đơn trước khi nhận dạng.

3.5. Phân tách dấu kép

Bước này được thực hiện dựa trên cả ảnh ban đầu và ảnh làm mảnh của dấu. Các điểm ảnh tại vị trí góc trong ảnh ban đầu được loại bỏ dựa trên tọa độ góc (và vùng lân cận của góc) thu được từ ảnh làm mảnh nhằm mục đích tách riêng từng nhánh tương ứng với góc đang xét. Tiếp theo, các nhánh được gom vào 2 nhóm, tương ứng với 2 dấu đơn cần nhận dạng.

3.5.1. Tách các nhánh

Tọa độ điểm góc đã xác định ở quá trình kiểm tra dấu được sử dụng để tách riêng các nhánh như sau:

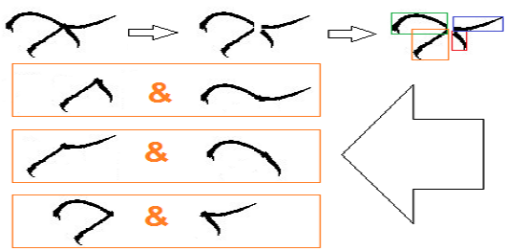
- Loại bỏ điểm góc hiện tại
- Đặt r = 1 (bán kính vùng lân cận)
- Đặt n = số vùng liên thông
- Lặp lại các bước sau đến khi n > 1
  - r = r + 1
  - Xóa điểm ảnh thuộc vùng lân cận r×r
  - n = số vùng liên thông hiện tại
- Lấy tất cả nhánh đã tách

Sau khi thực hiện quá trình tách nhánh cho các góc trong mỗi dấu, ta thu được một tập hợp các nhánh riêng lẻ. Lưu ý rằng với dấu kép có góc nối 3 nhánh (hình 5a) thì chỉ góc này được tách nhánh; ngược lại, với dấu kép có hai góc 2 nhánh (Hình 5c), bước tách nhánh được thực hiện với cả hai góc.

3.5.2. Phân nhóm các nhánh

Việc phân chia các nhánh vào hai nhóm được thực hiện bằng cách chọn 2 nhánh đưa vào nhóm 1 và các nhánh còn lại được đưa vào nhóm 2. Do đó, ta có cách phân nhóm.

Các nhánh ở mỗi nhóm mang ý nghĩa biểu diễn một dấu đơn trong dấu kép đang xét. Nếu ta chỉ đơn thuần liên kết các nhánh dựa trên vị trí của chúng trong ảnh ban đầu, dấu đơn thu được trở nên không đầy đủ do các điểm ảnh tại góc đã bị loại bỏ. Vì vậy, chúng tôi sử dụng phép trừ ảnh để xác định dấu đơn tương ứng với mỗi nhóm. Cụ thể, với mỗi nhóm, chúng tôi tiến hành trừ đi các nhánh trong nhóm đó trên ảnh gốc để thu được ảnh đơn tương ứng với các nhánh trong nhóm còn lại. Một ví dụ cụ thể được trình bày trong Hình 7.



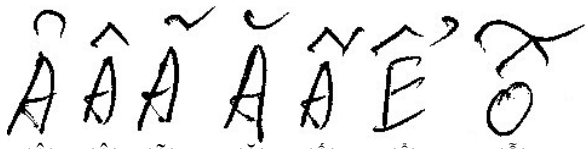
Hình 7: *3 cách chia nhóm cho 1 dấu kép ('^'và '~')*

3.5.3. Nhận dạng từng cặp dấu đơn

Quá trình này được thực hiện dựa trên mô hình Markov ẩn [7] và đặc trưng mã xích [5] (chain code). Phương pháp học máy này được chọn vì mức độ hiệu quả cao và cần thiết có sự so sánh kết quả nhận dạng các cặp dấu để chọn ra kết quả chính xác nhất. Với từng cặp dấu, 6 xác suất (tương ứng với 6 lớp) của mỗi dấu đơn được tính toán và dấu này được gán vào lớp cho xác suất đầu ra cao nhất.

Sau khi thu được các giá trị xác suất cao nhất tương ứng với mỗi cặp dấu, chúng tôi tính tổng xác suất từng cặp và so sánh để chọn ra cặp có xác suất tổng cao nhất. Kết quả nhận dạng của cặp dấu đơn này được xem là kết quả cuối cùng.

4. Kết quả thử nghiệm



Hình 8: *Một số ký tự có dấu trong tập dữ liệu*

Hiện nay chưa có bộ dữ liệu chuẩn về ký tự viết tay tiếng Việt nên giải pháp đề xuất được thử nghiệm với dữ liệu cục bộ. Các ký tự này được lấy từ nguồn dữ liệu mở của các sinh viên trường Đại học Bách khoa Hà Nội [8] và một phần do nhóm tác giả tự thu thập. Mỗi ảnh trong bộ dữ liệu chứa một ký tự tiếng Việt có dấu. Quá trình tách dấu được thực hiện thủ công để phục vụ cho việc huấn luyện. Một số ký tự được trình bày trong Hình 8.

Bảng 1: *Một số ký tự có dấu trong tập dữ liệu*

Dấu						
Mẫu thử	50	50	50	50	50	50
Nhận dạng	48	49	46	42	43	41
Tỉ lệ	96%	98%	92%	84%	86%	82%
Trung bình	**89.7%**					

Bảng 2: *Kết quả phân tách dấu kép*

Dấu	Số mẫu thử	Tách chính xác	Tỉ lệ
	40	38	95.0%
	40	40	100%
	40	39	97.5%
	40	38	95.0%
	40	36	90.0%
	40	40	100%
	40	35	87.5%
	40	37	92.5%
Độ chính xác trung bình: **94.7%**			

Việc huấn luyện được thực hiện với 120 mẫu cho mỗi dấu đơn, tương ứng với số lượng mẫu huấn luyện là 720.

Tập mẫu thử nghiệm được chia thành 2 nhóm, nhóm đầu tiên chứa các dấu đơn với tổng số mẫu là 300 chia đều cho 6 lớp, nhóm còn lại gồm 320 mẫu tương ứng với 8 dấu kép. Kết quả thử nghiệm với nhóm dấu đơn được trình bày ở Bảng 1.

Với nhóm dấu kép, kết quả tách dấu khá khả quan khi độ chính xác trung bình lên đến hơn 94% (Bảng 2).

5. Kết luận

Bài báo này trình bày một cách tiếp cận mới để giải quyết vấn đề nhận dạng dấu trong ký tự viết tay tiếng Việt. Giải pháp đề xuất bắt đầu bằng việc xác định dấu trong ký tự dựa trên kĩ thuật gán nhãn vùng liên thông. Tiếp theo, dấu đang xét được kiểm tra và phân vào 2 lớp: dấu đơn và dấu kép. Quá trình nhận dạng bằng mã xích kết hợp với mô hình Markov ẩn được áp dụng cho dấu đơn hoặc dấu kép đã phân tách. Ưu điểm lớn nhất ở cách tiếp cận này là các dấu có thể được nhận dạng cả khi có sự liên kết tại cặp dấu trong cùng một ký tự. Ngoài ra, chi phí tính toán cũng đáp ứng được yêu cầu nhận dạng theo thời gian thực.

Trong các nghiên cứu tiếp theo, giải pháp này sẽ được cải tiến để tăng khả năng nhận dạng, áp dụng được cho một số trường hợp cách viết không chính xác. Bên cạnh đó, một số luật về chính tả cũng có thể được kết hợp để nâng cao hiệu quả nhận dạng.

Tài liệu tham khảo

[1] De Cao Tran (2012), "An efficient method for on-line Vietnamese handwritten character recognition", Proceedings of the Third Symposium on Information and Communication Technology, pp. 135-141.

[2] De Cao Tran, Patrick Franco, Jean-Marc Ogier (2010), "Accented Handwritten Character Recognition Using SVM – Application to French", International Conference on Frontiers in Handwriting Recognition (ICFHR), Kolkata, pp. 65-71.

[3] Gary Bradski, Adrian Kaehler (2008), Learning OpenCV, pp. 115-120.

[4] Sivaramakrishnan Rajaraman, Arun Chokkalingam (2013), "Connected Components Labeling and Extraction Based Interphase Removal from Chromosome Images", International Journal of Bio-Science and Bio-Technology, vol. 5, no. 1, pp. 81-90.

[5] Mark Nixon, Alberto Aguado (2008), Feature Extraction & Image Processing 2nd, Academic Press, UK.

[6] D.Parks, J.P.Gravel (2004), "Corner Detection", International Journal of Computer Vision.

[7] Sung-Jung Cho (2005), "Introduction to Hidden Markov Model and Its Application", Samsung Advanced Institute of Technology (SAIT).

[8] Sapphire-ocr data, http://sapphire-ocr.googlecode.com /files/samples-full.zip.

(BBT nhận bài: 21/12/2013, phản biện xong: 27/12/2013)