Tạp chí Tin học và Điều khiển học, T.16, S.3 (2000), 1–6

DẠNG ĐIỂM VÀ ĐỐI SÁNH DẠNG ĐIỂM - ỨNG DỤNG TRONG NHẬN DẠNG KÝ TỰ TIẾNG VIỆT

NGUYỄN NGỌC KỶ

Abstract. In many cases, a structural object can be represented as characteristic point set extracted from some local distinctive properties of image. This paper presents some theoretic treatment and experimental results of some point set matching methodes using interdistances or path-length for Vietnamese character recognition problem. Proposed methodes of matching are invariant to translation, rotation, and scaling and furthermore low sensitive to various types of noise and distortion.

Tóm tắt. Trong nhiều trường hợp, một đối tượng có cấu trúc phức tạp thường có thể biểu diễn được bằng một tập điểm được trích chọn trên cơ sở một số tính chất cục bộ của ảnh tại mỗi điểm. Bài này trình bày kết quả xử lý lý thuyết và thực nghiệm một số phương pháp đối sánh các dạng điểm sử dụng các khoảng cách hay đường dẫn giữa các điểm và ứng dụng cho bài toán nhận dạng ký tự tiếng Việt. Các phương pháp được áp dụng không chỉ có tính chất bất biến đối với phép quay, tịnh tiến, tỷ lệ mà còn chịu được sai số do ảnh hưởng của nhiễu, biến dạng và sai số định vị.

1. MỞ ĐẦU

Trong lý thuyết nhận dạng thống kê, mỗi đối tượng thường được biểu diễn bằng một điểm trong không gian nhiều chiều với mỗi chiều là một thuộc tính định lượng. Phương pháp biểu diễn này không còn phù hợp đối với các dạng hình học. Bởi vậy, gần đây người ta đã đưa ra khái niệm dạng điểm (point pattern). Theo cách biểu diễn mới này thì mỗi đối tượng được thể hiện bằng một tập điểm trên không gian n chiều. Một thí dụ đơn giản của dạng điểm là các chùm sao Đại hùng tinh, Tiểu hùng tinh, tuy vị trí từng ngôi sao thay đổi theo mùa song người ta vẫn có thể dễ dàng nhận biết ra chúng theo tương quan vị trí giữa các ngôi sao. Đối với ký tự ta cũng có thể biểu diễn chúng bằng dạng điểm trên không gian hai chiều. Các điểm biểu diễn ở đây có thể là các điểm đặc trưng như điểm ngã ba hay ngã tư, điểm bắt đầu hay kết thúc, điểm cực trị. Quan hệ giữa các điểm có thể là khoảng cách, tính liên thông hay số giao điểm của đường nối hai điểm với các nét chữ.... Ta dễ dàng nhận thấy rằng các tính chất thu được trên dạng điểm là bất biến đối với các phép quay, tịnh tiến và nếu chuẩn hóa tốt còn bất biến đối với tỷ lệ. Trong [5] chúng tôi đã có dịp khảo cứu kỹ một loạt các phương pháp nhận dạng điểm tổng quát. Sau đây chúng tôi chỉ chọn lọc và xem xét một vài phương pháp nhận dạng điểm thích hợp cho bài toán nhận dạng chữ: ít nhạy cảm với sai số do biến dạng, do nhiễu hoặc do định vị.

2. MỘT SỐ KHÁI NIỆM CƠ BẢN

Định nghĩa 1. Cho E là trục số thực, cho $P = \{x_1, x_2, \ldots, x_n\}$ và $P' = \{y_1, y_2, \ldots, y_n\}$ là hai dạng điểm. P' được gọi là α-xáo trộn của P nếu tồn tại hàm $f : \{1, 2, \ldots, n\} \to \{1, 2, \ldots, n\}$ sao cho

$$\mathrm{abs}(d(x_i, x_j) - (y_{f(i)}, y_{f(j)}))/d(x_i, x_j) \leq \alpha$$

với mọi $i, j = 1, 2, \ldots, n$; hằng số $\alpha \in E$, và hàm $d(\cdot)$ có thể chọn là khoảng cách Euclide, $f(\cdot)$ là hàm xáo trộn.

Định nghĩa 2. Cho X, X' là các trục thực, $g(.)$ là ánh xạ 1–1 từ X đến X'. Ta gọi X' là α-biến dạng của X nếu với mọi cặp điểm x_1, x_2 trên X ta có:

$$\mathrm{abs}(d(x_1, x_2) - d(g(x_1), g(x_2)))/d(x_1, x_2) \leq \alpha,$$

THƯ VIỆN TRUNG TÂM KHOA HỌC TỰ NHIÊN VÀ CÔNG NGHỆ

$d(\cdot)$ là một hàm khoảng cách và $g(\cdot)$ được gọi là hàm biến dạng và α là một hằng thực dương.

Định nghĩa 3. Cho E^2, $E^{2\prime}$ là các không gian Euclide hai chiều; x, y là các trục tọa độ của E^2 và x', y' là các trục tọa độ của $E^{2\prime}$. Ta gọi $E^{2\prime}$ là không gian α-biến dạng của E^2 nếu: các trục x', y' theo thứ tự là α-biến dạng của các trục x, y.

Hàm biến dạng $G : E^2 \to E^{2\prime}$ là hàm vectơ, tạo thành từ hai hàm biến dạng thành phần của các trục tọa độ tương ứng.

Bổ đề 1. *$P = \{a_1, a_2, \ldots, a_n\}$ là một dạng điểm trên E^2 và $P' = \{b_1, b_2, \ldots, b_n\}$ là biến dạng của P' trên F^2 theo nghĩa:*

$$b_k = G(a_k) = (g_x(a_k), g_y(a_k)),$$

$k = 1, 2, \ldots, n$; $G(\cdot)$ là hàm biến dạng, $d(\cdot)$ là khoảng cách Euclide, $\alpha \in [0, 1]$. Khi đó P' là α-xáo trộn của P.

3. MỘT SỐ PHƯƠNG PHÁP NHẬN BIẾT DẠNG ĐIỂM

3.1. Phương pháp nhận dạng theo vectơ sắp xếp tất cả các khoảng cách giữa các điểm SIDV (sorted interdistances vectors)

Nội dung cơ bản của phương pháp này là sử dụng vectơ sắp xếp tất cả các khoảng cách giữa các điểm như là một bất biến của dạng điểm. Sự giống nhau giữa hai dạng điểm P và P' được thể hiện bằng các số lượng các thành phần giống nhau của hai vectơ tương ứng.

3.2. Phương pháp nhận dạng theo vectơ sắp xếp tất cả các khoảng cách tới láng giềng gần nhất SNNV (sorted nearest neighbourhood vectors)

Phương pháp nhận dạng theo vectơ sắp xếp tất cả các khoảng cách giữa các điểm đòi hỏi quá nhiều thời gian tính toán. Phương pháp nhận dạng theo vectơ sắp xếp tất cả các khoảng cách tới láng giềng gần nhất là một trong những hướng giảm bớt độ phức tạp tính toán.

Định nghĩa 4. Cho P là một dạng gồm n điểm $x_1, x_2, \ldots, x_n$. Ứng với mỗi $i = 1, 2, \ldots, n$ ta có:

$$d_i = \min d(x_i, x_j), \quad 1 < j < n, \ i \neq j,$$

ở đây, $d(x_i, x_j)$ là khoảng cách Euclide tính từ x_i tới x_j. Ta định nghĩa vectơ $\mathrm{SNNV}(P) \in \mathbb{R}^n$ là vectơ sắp xếp các khoảng cách từ mỗi điểm tới điểm láng giềng gần nhất của nó. Thành phần thứ i của $\mathrm{SNNV}(P)$ ứng với phần tử thứ i của tập $\{d_1, d_2, \ldots, d_n\}$.

Định nghĩa 5. Giả sử P là dạng mẫu và P' là dạng cần nhận biết, mỗi dạng đều có n phần tử và α là một số thực, dương. P' được gọi là α-SNNV trùng hợp với P nếu

$$\mathrm{abs}(\mathrm{SNNV}(P)_i - \mathrm{SNNV}(P')_i)/\mathrm{SNNV}(P)_i \leq \alpha \text{ với mọi } i = 1, 2, \ldots, n.$$

Định lý 1. *Giả sử $P = \{a_1, a_2, \ldots, a_n\}$ và $P' = \{b_1, b_2, \ldots, b_n\}$ là một α-xáo trộn của P. Khi đó P' là α-SNNV trùng hợp của P.*

Phần chứng minh định lý dựa trên kết quả của Bổ đề 1 đã được trình bày trong [5].

3.3. Phương pháp nhận dạng theo vectơ sắp xếp các tổng khoảng cách tới mỗi điểm

Phương pháp sử dụng SIDV là quá tốn kém thời gian thì SNNV thể hiện ít phức tạp hơn. Song như ta sẽ thấy, phương pháp sử dụng vectơ sắp xếp các tổng khoảng cách tới mỗi điểm còn đơn giản hơn.

Định nghĩa 6. Giả sử P là dạng n điểm, $P = \{x_1, x_2, \ldots, x_n\}$, vectơ $\mathrm{SRSV}(P) \in \mathbb{R}^n$ là vectơ sắp xếp các tổng khoảng cách từ mỗi điểm tới tất cả các điểm còn lại của P. Thành phần thứ i của $\mathrm{SRSV}(P)$ ứng với phần tử thứ i theo thứ tự sắp xếp tăng dần của tập $\{S_1, S_2, \ldots, S_n\}$, với

$$S_i = \sum_{j=1}^{n} d(x_i, x_j), \quad i = 1, 2, \ldots, n.$$

Định nghĩa 7. Giả sử P là dạng mẫu và P' là dạng cần nhận biết, mỗi dạng đều có n phần tử và α là một số thực dương. P' được gọi là α-SRSV trùng hợp với P nếu

$$\text{abs}(\text{SRSV}(P)_i - \text{SRSV}(P')_i/\text{SRSV}(P)_i \leq \alpha \text{ với mọi } i = 1, 2, \ldots, n.$$

Định lý 2. *Giả sử $P = \{a_1, a_2, \ldots, a_n\}$ là một dạng đã cho và $P' = \{b_1, b_2, \ldots, b_n\}$ là một α-xáo trộn của P. Khi đó P' là α-SRSV trùng hợp của P.*

Phần chứng minh định lý đã được trình bày chi tiết trong [5].

3.4. Phương pháp nhận dạng theo vectơ sắp xếp các khoảng cách tới trọng tâm

So với các phương pháp trên, phương pháp sử dụng vectơ sắp xếp các khoảng cách tới trọng tâm không đòi hỏi phải tính ma trận khoảng cách mà chỉ cần tính trọng tâm và n khoảng cách từ các điểm tới nó.

Định nghĩa 8. Giả sử P là một dạng n điểm, $P = \{a_1, a_2, \ldots, a_n\}$, x_t là trọng tâm của P. Ta định nghĩa vectơ $\text{SRV}(P) \in \mathbb{R}^n$ là vectơ sắp xếp các khoảng cách từ tâm x_t tới các điểm của P. Thành phần thứ i của $\text{SRV}(P)$ ứng với phần tử thứ i theo thứ tự sắp xếp tăng dần của tập $\{r_1, r_2, \ldots, r_n\}$ với $r_i = d(x_i, x_t)$, $i = 1, 2, \ldots, n$.

Định nghĩa 9. Giả sử P là dạng mẫu và P' là dạng cần nhận biết, mỗi dạng đều có n phần tử và α là một số thực dương. P' được gọi là α-SRV trùng hợp với P nếu

$$\text{abs}(\text{SRV}(P)_i - \text{SRV}(P')_i/\text{SRV}(P)_i \leq \alpha \text{ với mọi } i = 1, 2, \ldots, n.$$

Định lý 3. *Giả sử $P = \{a_1, a_2, \ldots, a_n\}$ là một dạng đã cho trên mặt phẳng E^2 và $P' = \{b_1, b_2, \ldots, b_n\}$ là một điểm α-biến dạng của P, $\alpha \in [0, 1]$, trên mặt phẳng F^2, với hàm biến dạng $g(a_i) = b_i$, $i = 1, 2, \ldots, n$. Khi đó P' là α-SRV trùng hợp của P.*

Phần chứng minh định lý đã được trình bày chi tiết trong [5].

4. MỘT SỐ PHƯƠNG PHÁP ĐỐI SÁNH DẠNG ĐIỂM CHO TRƯỜNG HỢP THÔNG TIN KHÔNG ĐẦY ĐỦ

Trong phần này ta chỉ quan tâm tới các bất biến chịu được dữ liệu thiếu, đó là:

- Xâu các khoảng cách từ các điểm đến một điểm chốt chọn trước nào đó.
- Xâu tất cả các khoảng cách giữa các điểm, sắp xếp theo thứ tự tăng dần.
- Xâu các cạnh của đồ thị láng giềng gần nhất.

Trong các trường hợp này, nói chung việc đối sánh hai dạng điểm được qui về vấn đề đối sánh hai xâu số thực để tìm số các thành phần chung. Nếu số các thành phần chung vượt quá một ngưỡng cho trước, hai dạng được coi là tương hợp với nhau. Riêng trường hợp đối với xâu các cạnh của đồ thị láng giềng gần nhất, có thể sử dụng để nhận dạng dựa trên hai giả thiết bổ sung sau:

- Xác suất một điểm không cùng xuất hiện với điểm láng giềng gần nhất của nó là nhỏ.
- Có thông tin phụ trợ, khẳng định hay không khẳng định điểm láng giềng gần nhất của từng điểm, dựa trên cơ sở quan sát hay xử lý tự động các vùng không xác định trên ảnh.

Như vậy, bằng cách chọn những bất biến chịu được dữ liệu thiếu, và thay vì đối sánh theo các vectơ có thể đối sánh theo xâu, dù tính toán tốn kém hơn, song vẫn đảm bảo có thể mở rộng các kết quả đã nghiên cứu ở mục trên. Theo hướng này, đối với trường hợp nhận dạng chữ, chúng tôi bổ sung thêm phương pháp sau đây:

Phương pháp nhận dạng theo xâu sắp xếp các độ dài đường dẫn ngắn nhất giữa các cặp điểm

Cùng cách xử lý trên, đối với dạng chữ in có thể bổ sung phương pháp nhận dạng theo xâu sắp xếp các độ dài đường dẫn ngắn nhất giữa mọi cặp điểm. Ưu điểm của phương pháp này là nó cho phép đối sánh được cả cho các trường hợp xuất hiện các ngã ba, ngã tư cùng điểm bắt đầu / kết thúc ký sinh. Cũng giống như các phương pháp trên, phương pháp này không nhạy đối với sai

số. Đặc biệt là nó có thể cho phép nhận biết được cả trường hợp các chữ cái dính nhau vì chúng được mô tả bởi một xâu chứa trong nó hai xâu con của từng ký tự. Tuy nhiên, phương pháp này đòi hỏi phải tính được độ dài các nét chữ và trên cơ sở đó tính được đường đi ngắn nhất giữa các cặp điểm. Song như trong [4] đã phân tích, với cách tiếp cận vectơ, độ dài nét có thể hoàn toàn tính được đồng thời với quá trình vectơ hóa.

5. KẾT QUẢ CÀI ĐẶT THỬ NGHIỆM

Để tổ chức cài đặt thực nghiệm đánh giá khả năng đối sánh của từng kiểu biểu diễn, chúng tôi đưa ra khung đồ tổng quát sau đây:

1. Chọn một phông chữ tiếng Việt tiêu biểu: Để đơn giản ta có thể chọn phông VnArial, sau đó soạn thảo và liệt kê lần lượt tất cả các chữ cái, mỗi ký tự 4 lần xuất hiện và in trên máy in HP LaserJet 4 Plus, độ phân giải máy in: 300-600 dpi.

2. Học và biểu diễn: Tiến hành tách chữ, vectơ hóa và trích chọn đặc điểm cho từng mẫu ảnh ký tự, sau đó tổng hợp chọn biểu diễn chung cho từng dạng ký tự. Bộ dấu hiệu mô tả từng mẫu ký tự bao gồm:

Phần dấu hiệu khái quát:

+ Chiều rộng, chiều dài hình chữ nhật ngoại tiếp từng ký tự.
+ Số chu trình.
+ Các điểm bắt đầu / kết thúc trên, dưới, trái, phải.
+ Số điểm ngã ba, ngã tư.
+ Số và vị trí tương đối của điểm cực trị, điểm cong gấp.

Phần dấu hiệu đồ thị:

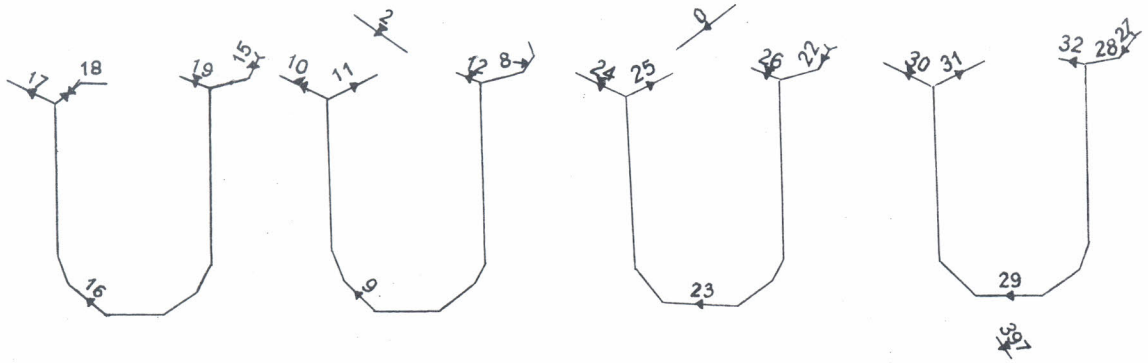
+ Tổng độ dài tất cả các cạnh.
+ Chọn một trong các phương pháp đối sánh dạng điểm trình bày ở trên, chẳng hạn xâu sắp xếp các độ dài đường dẫn ngắn nhất.

3. Tổ chức sắp xếp, phân cấp, tối thiểu hóa và đánh chỉ số bộ dấu hiệu mẫu chữ để gia tăng tốc độ đối sánh.

4. Tổ chức nhập một file ảnh ký tự tiếng Việt có xuất hiện đầy đủ ký tự tiếng Việt có cùng phông đã học để nhận biết thử.

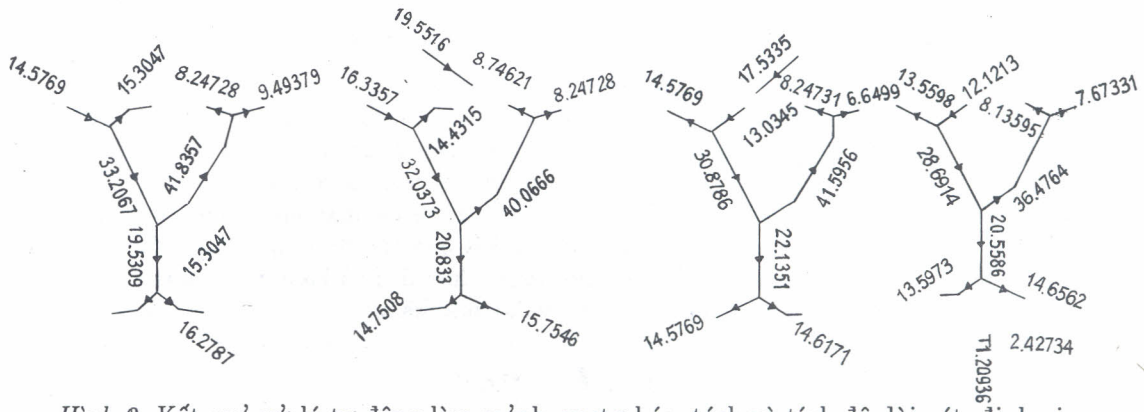
5. Đánh giá mức độ phân biệt, đề xuất các giải pháp xử lý bổ trợ cho những trường hợp còn nhận biết nhầm lẫn và nhập nhằng.

Các hình 1 và 2 trình bày kết quả xử lý và biểu diễn một số ký tự tiếng Việt.



Hình 1. Kết quả xử lý tự động làm mảnh, vectơ hóa, tách nét, đánh số thứ tự và định vị điểm đặc trưng (điểm bắt đầu / kết thúc, điểm ngã ba) của các ký tự

Ư , Ừ , Ứ , Ự



Hình 2. Kết quả xử lý tự động làm mảnh, vectơ hóa, tách và tính độ dài nét, định vị điểm đặc trưng (điểm bắt đầu / kết thúc, điểm ngã ba) và biểu diễn cấu trúc ký tự dưới dạng đồ thị có hướng của các ký tự

Y , Ỳ , Ý , Ỵ

Phần tiếp theo sẽ trình bày kết quả thử nghiệm thành công của hướng tiếp cận vectơ thông qua kết quả cài đặt thử nghiệm ứng dụng đọc và kiểm tra hộ chiếu tự động theo tiêu chuẩn ICAO.

Cài đặt phần mềm đọc và kiểm tra hộ chiếu sử dụng phông chữ theo tiêu chuẩn ICAO

1. Công dụng của phần mềm

Phần mềm kiểm tra hộ chiếu đọc máy là phần mềm chuyên dụng được thiết kế nhằm mục đích đọc và kiểm tra tự động các thông tin ghi trên trang nhân thân của quyển hộ chiếu in theo tiêu chuẩn của Hiệp hội Hàng không quốc tế (ICAO). Theo tiêu chuẩn này, ngoài phần thông tin nhân thân, mỗi quyển hộ chiếu đều có in thêm hai dòng mã gồm các ký tự 0-9, A-Z và một vài ký tự đặc biệt khác như "<", ">" bằng một phông chữ chuẩn OCRB. Nếu chương trình đọc được các dòng mã này và sau khi giải mã ra sẽ thu được bản ghi dữ liệu ghi trên trang nhân thân. Tính tiện lợi cúa việc sử dụng hộ chiếu đọc máy thể hiện trên hai mặt: ngăn ngừa nạn làm giả và cho phép tăng nhanh tốc độ kiểm tra, đối chiếu tại các cửa khẩu.

2. Các modul chức năng

Chương trình gồm các modul chính sau đây:
- Hiển thị ảnh trang nhân thân của hộ chiếu.
- Nhận dạng ký tự trên hai dòng mã và giải mã.
- Hiển thị bản ghi thông tin nhân thân sau khi đọc và giải mã để đối chiếu.
- Lưu và Tra tìm trên cơ sở dữ liệu "danh sách đen".
- Hỗ trợ kiểm tra ảnh, chữ ký.

Như vậy, bên cạnh các modul nhận dạng ảnh, chữ ký, modul có tầm quan trọng bậc nhất ở đây là đọc ký tự OCR.

3. Kết quả thực nghiệm

Mẫu thử nghiệm: 200 hộ chiếu đọc máy, có in các dòng mã ICAO theo đúng phông chữ chuẩn ICAO, trong đó có 100 mẫu in bằng máy in laser, 100 mẫu in bằng máy in kim.

+ Trường hợp đối sánh theo phương pháp cũ (xử lý ảnh bitmap):
Đối với các mẫu in laser: chính xác 100%.
Đối với các mẫu in kim: chính xác 90%.
Tốc độ 1 phút/trang.

+ Trường hợp ứng dụng kỹ thuật vectơ:
Đối với các mẫu in laser: chính xác 100%.
Đối với các mẫu in kim: chính xác 95,5%.

Tốc độ 15 giây/trang (nhờ kết hợp đồng thời các giai đoạn tách chữ và vectơ hóa).

4. Đánh giá

Tương tự như các ứng dụng dựa trên kỹ thuật mã vạch một chiều, hai chiều, vấn đề đọc chữ in chuẩn để gia tăng tốc độ và hiệu quả dịch vụ là một vấn đề tưởng đơn giản nhưng có nhiều đòi hỏi thực tế rất gay gắt: đó là tính đồng bộ trong nghiên cứu triển khai các giải pháp. Chương trình của nhóm chúng tôi qua thử nghiệm đã khẳng định được rằng về phần mềm chúng ta có đầy đủ khả năng để thực thi với hiệu quả nhận biết cao. Những kết quả thu được qua triển khai viết chương trình đọc hộ chiếu hoàn toàn có thể mở rộng cho trường hợp đọc và kiểm tra chứng minh thư nhân dân, bằng lái xe, thẻ kiểm soát, và các loại giấy tờ đặc biệt khác.

6. KẾT LUẬN

Nhận dạng chữ Việt là một vấn đề rất phức tạp song qua kết quả xử lý lý thuyết và thực nghiệm sơ bộ bước đầu có thể thấy hướng tiếp cận vectơ và biểu diễn ký tự bằng dạng điểm là một hướng triển vọng vì nó cho ta một cách giải quyết vấn đề nhận dạng chữ Việt một cách cơ bản, xét về khả năng trích chọn đặc điểm chi tiết tới các đường nét cũng như tính chịu nhiễu, chịu sai số trong đối sánh dạng điểm. Nhiều vấn đề liên quan đến vấn đề học, tối thiểu hóa biểu diễn và xây dựng chiến lược đối sánh hiệu quả trong những trường hợp gia tăng nhiễu vẫn còn cần phải được tiếp tục nghiên cứu.

TÀI LIỆU THAM KHẢO

[1] Chia-Wei Liao and Jun S. Huang, A transformation invariant matching algorithm for handwritten chinese character recognition, *Pattern Recognition* **23** (11) (1990) 1167-1188.

[2] ICAO, *Machine Readable Passports*, Doc 9303, Part 1, Third edition, 1992.

[3] Lavine D., Lambird B., Kanal L. N., Recognition of spatial point pattern, *Pattern Recognnition* **16** (3) (1990) 1167-1188.

[4] Nguyễn Ngọc Kỷ, Phương pháp biểu diễn cấu trúc chữ Việt theo hướng tiếp cận vectơ, *Tạp chí Tin học và Điều khiển học* **16** (1) (2000) 72-79.

[5] Nguyễn Ngọc Kỷ, "Biểu diễn và đồng nhất ảnh đường nét", Luận án Phó tiến sĩ Toán-Lý, Hà Nội, 1992.

[6] Nguyễn Ngọc Kỷ, "Khảo sát lý thuyết và thực nghiệm phương pháp nhận dạng ký tự tiếng Việt theo hướng tiếp cận vectơ", Báo cáo kết quả thực hiện đề tài NCKH cấp Nhà nước KH01-07, nhánh OCR, Hà Nội, 1998.

[7] Vũ Văn Khoan, "Ứng dụng công nghệ thông tin trong sản xuất và kiểm tra hộ chiếu đọc máy theo tiêu chuẩn ICAO", Báo cáo đề tài NCKH cấp Bộ, Hà Nội, 1995–1997.

Nhận bài ngày 18 - 4 - 1999
Nhận lại sau khi sửa ngày 11 - 4 - 2000

Tổng cục KHKT&CN, Bộ Công an.