

**ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN**

HỒ TRẦN NHẬT THỦY

**XÂY DỰNG HỆ THỐNG TRUY VẤN ẢNH
DỰA VÀO VĂN BẢN NGOẠI CẢNH**

**Chuyên ngành: KHOA HỌC MÁY TÍNH
Mã số: 60.48.01**

LUẬN VĂN THẠC SĨ

**NGƯỜI HƯỚNG DẪN KHOA HỌC:
TS. LÝ QUỐC NGỌC**

Thành phố Hồ Chí Minh – 2012

LỜI CẢM ƠN

Đầu tiên, tôi xin gửi lời cảm ơn chân thành và sâu sắc nhất đến TS. Lý Quốc Ngọc. Thầy đã tận tình hướng dẫn, chỉ bảo, động viên tôi trong suốt thời gian thực hiện đề tài, đưa ra những lời khuyên quý báu và khơi gợi cảm hứng giúp tôi hoàn thành luận văn này.

Tôi xin chân thành cảm ơn quý Thầy Cô trong Khoa Công nghệ Thông tin đã truyền đạt cho tôi những kiến thức quý báu, những kinh nghiệm, suy nghĩ về cuộc sống.

Tôi xin cảm ơn các anh chị, bạn bè trong khoa đã đóng góp những ý kiến quý báu và hữu ích trong thời gian thực hiện luận văn.

Cuối cùng, tôi xin được bày tỏ lòng biết ơn sâu sắc đối với Cha Mẹ, gia đình – những người đã luôn ở bên tôi, là điểm tựa và nguồn cổ vũ trong những khi tôi gặp khó khăn.

Thành phố Hồ Chí Minh, Tháng 8/2012

Người thực hiện đề tài

Hồ Trần Nhật Thủy

MỤC LỤC

LỜI CẢM ƠN.....	i
MỤC LỤC.....	ii
DANH MỤC CÁC KÝ HIỆU, CÁC CHỮ VIẾT TẮT.....	iv
DANH MỤC CÁC BẢNG.....	v
DANH MỤC CÁC HÌNH VẼ	vi
Mở đầu	1
Chương 1 Tổng quan	4
1.1 Động lực nghiên cứu	4
1.2 Phát biểu bài toán	5
1.3 Các đóng góp của luận văn	6
1.4 Tổ chức luận văn	6
Chương 2 Tình hình nghiên cứu và hướng tiếp cận của luận văn.....	7
2.1 Những khó khăn trong bài toán phát hiện văn bản ngoại cảnh trong ảnh.....	7
2.1.1 Văn bản trong ảnh có sự thay đổi về màu sắc, kiểu chữ, kích thước, hướng, vị trí, điều kiện chiếu sáng.....	7
2.1.2 Văn bản được nhúng trên nền phức tạp	8
2.1.3 Ảnh có độ tương phản thấp	9
2.2 Tình hình nghiên cứu trong lĩnh vực phát hiện văn bản trong ảnh.....	10
2.3 Các phương pháp rút trích văn bản	13
2.4 Tình hình nghiên cứu trong lĩnh vực truy vấn ảnh	14
2.5 Hướng tiếp cận	15
Chương 3 Mô hình phát hiện và rút trích văn bản ngoại cảnh trong ảnh ..	17
3.1 Sơ đồ chung	17
3.2 Tiền xử lý.....	18
3.3 Phát sinh vùng văn bản ứng viên	22
3.3.1 Phát sinh các ký tự ứng viên bằng SWT	23
3.4 Gom nhóm các thành phần liên kết.....	27
3.4.1 Nhóm các ký tự thành dòng văn bản	27

3.4.2	Tách dòng văn bản thành các từ	29
3.5	Tinh lọc các từ ứng viên bằng bộ phân lớp SVM	30
3.5.1	Đặc trưng HOG	30
3.5.2	Bộ phân lớp SVM	32
3.5.3	Huấn luyện và phân lớp từ bằng bộ phân lớp SVM	34
3.6	Rút trích văn bản	36
3.7	Hiệu chỉnh kết quả nhận dạng ký tự bằng phần mềm OCR	38
Chương 4	Mô hình truy vấn ảnh dựa vào văn bản ngoại cảnh	44
4.1	Mô hình tổ chức dữ liệu	44
4.1.1	Phát hiện, rút trích và nhận dạng văn bản	45
4.1.2	Gom nhóm văn bản	45
4.1.3	Trích chọn phần tử đại diện nhóm văn bản	47
4.2	Mô hình truy vấn ảnh dựa vào văn bản ngoại cảnh	48
Chương 5	Kết quả thực nghiệm	52
5.1	Kết quả phát hiện và rút trích văn bản	52
5.1.1	Tập dữ liệu thử nghiệm và phương pháp đánh giá	52
5.1.2	Kết quả thực nghiệm	53
5.2	Đánh giá hiệu quả phương pháp hiệu chỉnh kết quả nhận dạng bằng phần mềm OCR	57
5.3	Kết quả truy vấn ảnh	59
5.3.1	Kết quả truy vấn ảnh bằng từ khóa	59
5.3.2	Kết quả truy vấn ảnh bằng ảnh chứa văn bản tự nhiên	61
Chương 6	Kết luận và hướng phát triển	65
6.1	Kết luận	65
6.2	Hướng phát triển	66
Tài liệu tham khảo		67

DANH MỤC CÁC KÝ HIỆU, CÁC CHỮ VIẾT TẮT

Ký hiệu	Ý nghĩa
c_k	Nhóm ký tự thứ k
M_I	Mô hình truy vấn ảnh
n_{SI}	Số ảnh trong tập ảnh
ns_k	Số chuỗi ký tự trong nhóm thứ k
QI	Ảnh truy vấn
s_i	Chuỗi ký tự rút trích từ ảnh thứ i trong tập ảnh
s_{kl}	Từ thứ l của chuỗi ký tự thứ k
SC	Tập các nhóm chuỗi ký tự
SI	Tập dữ liệu ảnh
SRI	Tập ảnh kết quả
SS	Tập các chuỗi ký tự
SS_q	Tập các chuỗi ký tự ứng với câu truy vấn Q
ST	Tập các phần tử đại diện nhóm chuỗi ký tự
ST_q	Tập các phần tử đại diện nhóm chuỗi ký tự ứng với câu truy vấn Q
t_k	Phần tử đại diện của nhóm chuỗi ký tự thứ k

DANH MỤC CÁC BẢNG

Bảng 3.1 Thuật toán reconstruction cơ bản	18
Bảng 3.2 Thuật toán rút trích đặc trưng HOG	31
Bảng 3.3 Thuật toán nhị phân hóa vùng văn bản.....	37
Bảng 3.4 Thuật toán tính khoảng cách Levenshtein	39
Bảng 4.1 Giải thuật gom nhóm phân cấp từ dưới lên	46
Bảng 5.1 Hiệu quả phát hiện văn bản trong tập dữ liệu học của phương pháp đề xuất	53
Bảng 5.2 Hiệu quả của các phương pháp phát hiện văn bản khác nhau	53
Bảng 5.3 Hiệu quả nhận dạng văn bản trước và sau khi hiệu chỉnh.....	57
Bảng 5.4 Một số kết quả nhận dạng văn bản trước và sau khi hiệu chỉnh.....	58
Bảng 5.5 Hiệu quả truy vấn ảnh với độ dị biệt $\xi = 0.0$	64

DANH MỤC CÁC HÌNH VẼ

Hình 0.1 Minh họa văn bản nhân tạo trong ảnh.....	2
Hình 0.2 Minh họa văn bản ngoại cảnh trong ảnh.....	2
Hình 2.1 Minh họa văn bản trong ảnh không nhất quán về màu sắc, kiểu chữ, kích thước, hướng	8
Hình 2.2 Minh họa văn bản có sự chiếu sáng khác nhau	8
Hình 2.3 Minh họa văn bản được nhúng trên nền phức tạp	9
Hình 2.4 Minh họa ảnh có độ tương phản thấp	9
Hình 2.5 Các bước thực hiện trong hệ thống phát hiện và nhận dạng văn bản.....	10
Hình 3.1 Sơ đồ các bước thực hiện trong mô hình phát hiện và rút trích văn bản...	17
Hình 3.2 a) Ảnh mức xám ban đầu I; b) Ảnh khởi tạo J; c) Kết quả phép reconstruction của ảnh a); d) Kết quả khi lấy ảnh a - c.....	19
Hình 3.3 So sánh kết quả các phương pháp nhị phân ảnh. a) Ảnh kết quả reconstruction; b) Nhị phân bằng phương pháp Otsu; c) Nhị phân bằng ngưỡng T_{bin}	21
Hình 3.4 a) Kết quả thực hiện toán tử đóng trên ảnh nhị phân; b) Thực hiện phép giãn nở trên ảnh a); c) Các vùng văn bản ứng viên được lựa chọn	23
Hình 3.5 Minh họa đường nét trong ảnh [4].....	24
Hình 3.6 Các bước tìm độ rộng nét [4]	24
Hình 3.7 Minh họa ảnh SWT cho ký tự W.....	25
Hình 3.8 a) Ảnh SWT của ký tự “e” trước khi làm mịn; b) Ảnh SWT của ký tự “e” sau khi làm mịn	26
Hình 3.9 a) Ảnh SWT; b) Các ký tự ứng viên được chọn lọc.....	27
Hình 3.10 Kết quả các dòng văn bản hệ thống phát hiện được	28
Hình 3.11 Khoảng cách giữa các hình chữ nhật bao quanh ký tự	29
Hình 3.12 Các từ ứng viên.....	30
Hình 3.13 Quá trình rút trích đặc trưng HOG [3]	32
Hình 3.14 Một số mẫu từ tập huấn luyện bộ phân lớp.....	35

Hình 3.15 Kết quả phát hiện văn bản của hệ thống	35
Hình 3.16 Kết quả nhị phân hóa vùng văn bản.....	38
Hình 3.17 Minh họa các bước tính khoảng cách Levenshtein	40
Hình 4.1 Sơ đồ tổ chức dữ liệu ảnh.....	44
Hình 4.2 Minh họa các bước gom nhóm bằng thuật toán HAC	47
Hình 4.3 Sơ đồ truy vấn ảnh	48
Hình 5.1 Minh họa một số kết quả phát hiện văn bản ngoại cảnh trong ảnh	55
Hình 5.2 Minh họa một số trường hợp thất bại	56
Hình 5.3 Kết quả truy vấn ảnh dùng từ khóa “office”	59
Hình 5.4 Kết quả truy vấn ảnh dùng từ khóa (“car park”)	60
Hình 5.5 Kết quả truy vấn bằng ảnh.....	62
Hình 5.6 Kết quả truy vấn bằng ảnh.....	63

Mở đầu

Trong bối cảnh lượng dữ liệu ảnh ngày càng tăng và không ngừng phát triển như hiện nay, con người đang tốn nhiều công sức để quản lý và vẫn đang tìm kiếm phương pháp để khai thác và truy vấn dạng dữ liệu này một cách hiệu quả. Phương pháp chú thích ảnh thủ công không thể đáp ứng được vì tốn nhiều thời gian, đồng thời không thể mô tả hết các thông tin ẩn chứa trong ảnh. Thông tin được lưu trữ trên ảnh bao gồm phần nội dung thị giác và phần nội dung ngữ nghĩa. Phần nội dung thị giác bao gồm các thuộc tính như màu sắc, cường độ, hình dáng, vân. Phần nội dung ngữ nghĩa bao gồm các đối tượng, sự kiện và mối quan hệ giữa chúng. Việc khai thác phần nội dung ngữ nghĩa đến nay vẫn còn là một vấn đề thách thức. Văn bản trong ảnh là một trong những đối tượng mang đến thông tin ngữ nghĩa quan trọng giúp chúng ta hiểu được nội dung ảnh. Việc khai thác được nội dung văn bản trong ảnh có thể mang lại những lợi ích và các ứng dụng phong phú, bao gồm:

- Cung cấp các thông tin về ngữ nghĩa bổ sung hữu ích cho việc lập chỉ mục hay truy vấn ảnh.
- Áp dụng vào các hệ thống truy vấn ảnh theo nội dung mong muốn từ những từ khóa được rút trích từ văn bản trong ảnh.
- Sàng lọc, phân loại ảnh, hoặc ngăn chặn được các ảnh có nội dung xấu.
- Nội dung văn bản trong ảnh có thể được dịch sang nhiều ngôn ngữ khác giúp người dùng hiểu được nội dung ảnh đa ngôn ngữ.
- Áp dụng vào các hệ thống phát hiện các biển chỉ dẫn, bằng lái xe, các thiết bị hỗ trợ người dùng khiếm thị, hệ thống giao tiếp người máy, hệ thống giao thông thông minh,...

Một cách tổng quát, văn bản trong ảnh được chia thành hai loại: văn bản ngoại cảnh (scene text) và văn bản nhân tạo (artificial text). Văn bản nhân tạo là loại văn bản do con người tạo ra với mục đích giải thích, bổ sung, nhấn mạnh hoặc chú thích cho nội dung và ý nghĩa của ảnh. Chúng thường xuất hiện trong các bản tin, phụ đề phim, tỉ số của các trận đấu (Hình 0.1)...Loại văn bản này thường được thể hiện một

cách có tổ chức. Về màu sắc, hình dáng, kích thước, phương hướng thường có xu hướng thống nhất và văn bản không bị biến dạng. Ngược lại với văn bản nhân tạo, văn bản ngoại cảnh(hay còn gọi là văn bản tự nhiên) là văn bản tồn tại một cách tự nhiên trong ảnh. Nó xuất hiện trong ảnh chụp các bảng quảng cáo, áp phích, tên đường, tên cửa hàng, bảng hiệu, nhãn hiệu của các sản phẩm,... trong ảnh (Hình 0.2) .Văn bản ngoại cảnh có cách thể hiện không giới hạn, chúng có thể xuất hiện với bất cứ hình dáng, màu sắc, kích thước, độ nghiêng nào, trong điều kiện ánh sáng bất kỳ, với các bề mặt phẳng hay lượn sóng,... Do đó, nhiều nhà nghiên cứu nhận thấy văn bản ngoại cảnh khó phát hiện hơn văn bản nhân tạo. Mặc dù đã có nhiều kết quả đạt được trong lĩnh vực này, nhưng một số khó khăn vẫn còn tồn tại.



Hình 0.1 Minh họa văn bản nhân tạo trong ảnh



Hình 0.2 Minh họa văn bản ngoại cảnh trong ảnh

Các hệ thống truy vấn ảnh hiện có chủ yếu vẫn dựa vào các đặc trưng thị giác và chưa khai thác nhiều phần nội dung ngữ nghĩa trong ảnh. Đặc biệt, chưa có hệ

thống truy vấn ảnh nào khai thác đối tượng văn bản trong ảnh. Trong bối cảnh như trên, luận văn này tập trung trình bày hai vấn đề chính:

- Xây dựng mô hình phát hiện và rút trích văn bản ngoại cảnh trong ảnh.
- Xây dựng mô hình truy vấn ảnh dựa vào văn bản ngoại cảnh.

Trong mô hình phát hiện và rút trích văn bản ngoại cảnh trong ảnh, chúng tôi xây dựng mô hình nhằm giải quyết các vấn đề về sự thay đổi kích thước, kiểu chữ, màu sắc,... của văn bản ngoại cảnh, cũng như sự phức tạp của vùng nền xung quanh. Đối với mô hình phát hiện văn bản, chúng tôi sử dụng phép reconstruction để loại bỏ phần lớn các đối tượng thuộc vùng nền. Các toán tử hình thái học cũng được sử dụng để phát sinh các vùng văn bản ứng viên và các ký tự ứng viên được tạo thành từ một đặc trưng đủ mạnh. Cuối cùng, chúng tôi dùng bộ phân lớp dựa vào Support Vector Machines (SVM) được huấn luyện bằng đặc trưng Histogram of Oriented Gradient (HOG) để phân loại các từ ứng viên đã phát sinh. Một phương pháp nhị phân hóa vùng ảnh chứa văn bản được áp dụng để rút trích các ký tự từ ảnh nhằm giúp quá trình nhận dạng đạt kết quả tốt hơn.

Trong mô hình truy vấn ảnh dựa vào văn bản ngoại cảnh, chúng tôi tiến hành rút trích và nhận dạng các vùng văn bản trong ảnh từ tập dữ liệu ảnh. Sau đó, các chuỗi ký tự nhận dạng được sẽ được gom thành các nhóm khác nhau với phần tử đại diện cho nhóm. Từ tập dữ liệu ảnh ban đầu, ta thu được các nhóm chuỗi ký tự với phần tử đại diện. Các nhóm chuỗi ký tự và phần tử đại diện được sử dụng để so khớp trong giai đoạn truy vấn.

Tóm lại, với những thông tin ngữ nghĩa hữu ích được cung cấp từ văn bản trong ảnh, chúng tôi mong muốn xây dựng mô hình rút trích được đối tượng quan trọng này từ các ảnh. Từ đó, áp dụng vào bài toán truy vấn ảnh dựa vào văn bản ngoại cảnh với hy vọng có thể kết hợp với các hệ thống truy vấn thông tin thị giác hiện có để tạo thành một mô hình truy vấn thực sự hiệu quả và hữu dụng.

Chương 1 Tổng quan

Trong chương này, chúng tôi trình bày động lực nghiên cứu, mục đích nghiên cứu, các đóng góp của luận văn và sau cùng là các nội dung được trình bày.

1.1 Động lực nghiên cứu

Sự phát triển mạnh mẽ của các thiết bị ghi hình như máy ảnh kỹ thuật số, điện thoại di động, máy tính cá nhân,... dẫn đến việc số lượng các ảnh được tạo ra ngày càng nhiều và đang phát triển một cách nhanh chóng. Từ đó, một vấn đề được đặt ra là làm thế nào để quản lý và truy vấn cơ sở dữ liệu ảnh số lượng lớn một cách hiệu quả, cũng như có thể rút trích được các thông tin hữu ích từ ảnh. Ảnh thường chứa các thông tin quan trọng liên quan đến các sự kiện, vị trí, con người,... Theo cách truyền thống, dữ liệu ảnh được chú thích thủ công với một số lượng nhỏ các từ khóa mô tả ảnh. Tuy nhiên, với số lượng ảnh khổng lồ như hiện nay, việc chú thích ảnh bằng tay là không khả thi vì tốn rất nhiều thời gian, đồng thời không thể mô tả hết thông tin ẩn trong ảnh. Điều đó đã thúc đẩy các nhà nghiên cứu tìm kiếm, thiết kế và phát triển các thuật toán mới nhằm tự động rút trích thông tin từ ảnh và đánh chỉ mục cho hệ thống ảnh giúp việc truy vấn hiệu quả hơn. Trong số các nội dung thường xuất hiện trong ảnh như con người, cảnh vật, ... văn bản là một trong số những thông tin quan trọng giúp chúng ta hiểu được nội dung của ảnh. Văn bản xuất hiện trong ảnh cung cấp những thông tin ngữ nghĩa quan trọng, vì vậy nó có thể được sử dụng để đánh chỉ mục và truy vấn ảnh. Nếu văn bản trong ảnh có thể được rút trích, nó sẽ cung cấp những từ khóa có nghĩa cho việc mô tả nội dung của ảnh.

Truy vấn dữ liệu ảnh là một bài toán rất quan trọng trong lĩnh vực tin học và có ý nghĩa thiết thực trong cuộc sống. Bên cạnh đó, việc rút trích được văn bản trong ảnh cũng góp phần giúp máy tính có thể hiểu được nội dung ảnh và giải quyết một phần trở ngại khi nhận dạng văn bản ngoại cảnh trong ảnh. Từ ý nghĩa thực tiễn và khoa học đó, chúng tôi thực hiện đề tài xây dựng hệ thống truy vấn ảnh dựa vào

văn bản ngoại cảnh với mong muốn rút trích được thông tin quan trọng trong ảnh, đóng góp vào cộng đồng truy vấn ảnh bên cạnh các hệ thống truy vấn ảnh dựa vào nội dung hiện có.

1.2 Phát biểu bài toán

Trong luận văn này, đối tượng mà chúng tôi tập trung nghiên cứu là văn bản ngoại cảnh trong ảnh. Cho trước một tập gồm nhiều ảnh chứa văn bản ngoại cảnh. Luận văn tập trung vào các vấn đề sau:

- Phát hiện, rút trích và nhận dạng văn bản ngoại cảnh xuất hiện trong từng ảnh. Kết quả trả về là tập các hình chữ nhật bao quanh từ có trong ảnh cùng tập ảnh nhị phân tương ứng của các từ phát hiện được và chuỗi ký tự nhận dạng được.
- Cho phép người dùng thực hiện truy tìm các ảnh chứa các từ khóa mong muốn. Cho câu truy vấn dưới dạng từ khóa hoặc ảnh chứa từ khóa, kết quả truy vấn là tập ảnh được sắp hạng theo độ tương đồng (về nội dung văn bản có trong ảnh) so với ảnh truy vấn.

Từ phát biểu trên, các bài toán đề tài cần giải quyết như sau:

- Xây dựng mô hình phát hiện, rút trích và nhận dạng văn bản ngoại cảnh trong ảnh, gồm các giai đoạn:
 - Xác định vị trí các vùng văn bản có trong ảnh
 - Rút trích vùng ảnh văn bản đã định vị
 - Cải tiến kết quả nhận dạng văn bản ngoại cảnh từ phần mềm nhận dạng ký tự quang học (OCR).
- Xây dựng mô hình truy vấn ảnh dựa vào văn bản ngoại cảnh, gồm các giai đoạn:
 - Tổ chức dữ liệu ảnh dựa vào văn bản ngoại cảnh
 - Xác định độ đo dị biệt, sắp hạng kết quả tìm được dựa vào từ khóa

1.3 Các đóng góp của luận văn

Luận văn đã có các đóng góp chính như sau:

- Đề xuất và thử nghiệm mô hình phát hiện và rút trích văn bản ngoại cảnh trong ảnh tự nhiên. Mô hình góp phần vượt qua các trở ngại đối với bài toán phát hiện và rút trích văn bản ngoại cảnh trong ảnh: độ phân giải thấp, nền nhiễu loạn, không biết trước về màu sắc, font chữ, cỡ chữ, bố cục và vị trí của văn bản trong ảnh.
- Đề xuất và thử nghiệm mô hình hiệu chỉnh kết quả nhận dạng ký tự từ phần mềm OCR nhằm đạt kết quả nhận dạng văn bản tốt hơn. Mô hình góp phần vượt qua một phần các trở ngại của hệ thống nhận dạng ký tự quang học khi áp dụng trên văn bản ngoại cảnh.
- Đề xuất và thử nghiệm mô hình truy vấn ảnh dựa vào văn bản ngoại cảnh xuất hiện trong ảnh. Đây là mô hình truy vấn mới, chưa thấy được đề xuất trong các hệ thống truy vấn thông tin thị giác trong ảnh trước đây. Mô hình cho phép vượt qua một phần vấn đề về lỗ hổng ngữ nghĩa giữa dữ liệu lưu trữ ảnh và thông tin truy vấn, cho phép truy tìm các ảnh chứa từ khóa mong muốn cả trong trường hợp không biết ngôn ngữ của từ khóa.

1.4 Tổ chức luận văn

Phần còn lại của luận văn được tổ chức như sau:

- Chương 2 trình bày tình hình nghiên cứu trong lĩnh vực phát hiện và rút trích văn bản, lĩnh vực truy vấn ảnh, từ đó đề xuất hướng tiếp cận của luận văn.
- Chương 3 trình bày mô hình phát hiện và rút trích văn bản ngoại cảnh trong ảnh.
- Chương 4 trình bày mô hình truy vấn ảnh dựa vào văn bản ngoại cảnh gồm hai vấn đề chính là tổ chức dữ liệu và cách thức truy vấn ảnh.
- Chương 5 trình bày kết quả thực nghiệm.
- Chương 6 trình bày kết luận và hướng phát triển.

Chương 2 Tình hình nghiên cứu và hướng tiếp cận của luận văn

Trong chương này, chúng tôi trình bày những thách thức trong bài toán phát hiện văn bản trong ảnh, tình hình nghiên cứu trong lĩnh vực phát hiện và rút trích văn bản trong ảnh tự nhiên, trong lĩnh vực truy vấn ảnh, từ đó đề xuất hướng tiếp cận của luận văn.

2.1 Những khó khăn trong bài toán phát hiện văn bản ngoại cảnh trong ảnh

Trong thực tế, văn bản trong ảnh không phải lúc nào cũng được thể hiện một cách rõ ràng để hệ thống dễ dàng tìm thấy. Như đã trình bày ở trên, văn bản ngoại cảnh thường gặp các vấn đề không thống nhất về cách thức thể hiện. Bên cạnh đó, các yếu tố khách quan khác cũng chi phối rất nhiều đến chất lượng của văn bản trong ảnh, những khó khăn đó thực sự là những thách thức trong quá trình nghiên cứu và đưa ra được các mô hình phát hiện văn bản hiệu quả. Dưới đây là một số thách thức có thể nhìn thấy được rõ ràng nhất.

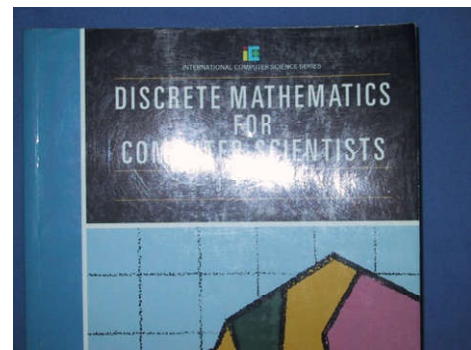
2.1.1 Văn bản trong ảnh có sự thay đổi về màu sắc, kiểu chữ, kích thước, hướng, vị trí, điều kiện chiếu sáng

Văn bản trong ảnh tự nhiên có thể có màu sắc tùy ý và hoàn toàn không thể biết trước. Một số phương pháp thường giả định văn bản trong ảnh có màu sắc giống nhau. Tuy nhiên trong thực tế, các dòng văn bản trong ảnh tự nhiên có thể có màu sắc khác nhau, thậm chí các từ trên cùng một dòng văn bản cũng có thể có màu khác nhau. Việc không xác định được màu của văn bản sẽ gây nhiều khó khăn cho giai đoạn phát hiện văn bản.

Bên cạnh màu sắc, văn bản trong ảnh tự nhiên khi thể hiện còn bao gồm cả kiểu chữ, kích thước, vị trí, hướng trong ảnh. Văn bản ngoại cảnh thường đa dạng về kiểu chữ và kích thước. Ngoài ra, văn bản ngoại cảnh có thể xuất hiện với hướng và vị trí bất kỳ trong ảnh, được chụp trong các điều kiện ánh sáng và góc nhìn khác nhau. Những vấn đề này lại tiếp tục đặt ra những thách thức cho hệ thống phát hiện văn bản. Các thách thức này được minh họa trong Hình 2.1 và Hình 2.2.



Hình 2.1 Minh họa văn bản trong ảnh không nhất quán về màu sắc, kiểu chữ, kích thước, hướng



Hình 2.2 Minh họa văn bản có sự chiếu sáng khác nhau

2.1.2 Văn bản được nhúng trên nền phức tạp

Một thách thức không nhỏ đặt ra cho hệ thống phát hiện văn bản trong ảnh là vùng nền thường có văn phức tạp hoặc có bóng mờ. Nền lúc này có thể sẽ có màu sắc khác nhau, thay đổi tùy ý gây khó khăn cho việc phân biệt giữa nền và văn bản, thậm chí màu nền đôi khi có màu sắc tương tự, gần giống với màu văn bản. Do văn bản nhúng trong ảnh nên việc văn bản xuất hiện trên những nền khác nhau là điều đương nhiên không thể nào tránh khỏi, khi đó nền có thể sẽ có những hình ảnh, hoặc xuất hiện những đường kẻ tương đồng nằm song song hoặc trùng với văn bản làm cho hệ thống không thể phân biệt được đâu là nền và đâu là văn bản.

Đây có thể được xem là một thách thức lớn nhất đặt ra cho giai đoạn nhị phân hóa và tăng cường chất lượng của ảnh bởi nó đặt ra rất nhiều khó khăn cho việc loại bỏ nhiễu xung quanh văn bản. Trường hợp này rất dễ dẫn đến việc nhận dạng sai văn bản bởi các thông tin dư thừa mà hệ thống không loại bỏ được trong quá trình lọc nhiễu.



Hình 2.3 Minh họa văn bản được nhúng trên nền phức tạp

2.1.3 Ảnh có độ tương phản thấp

Độ tương phản thấp là một trong những nguyên nhân khách quan do chất lượng của ảnh mà chúng ta thu nhận được. Hiện tượng này xảy ra khi ánh sáng trong ảnh quá sáng hoặc quá tối, đôi khi cũng do các màu sắc trong ảnh tương tự nhau quá nhiều cũng dẫn đến việc gây nên độ tương phản thấp. Ảnh có độ tương phản thấp cũng gây nhiều khó khăn cho giai đoạn nhị phân hóa và tăng cường chất lượng văn bản.

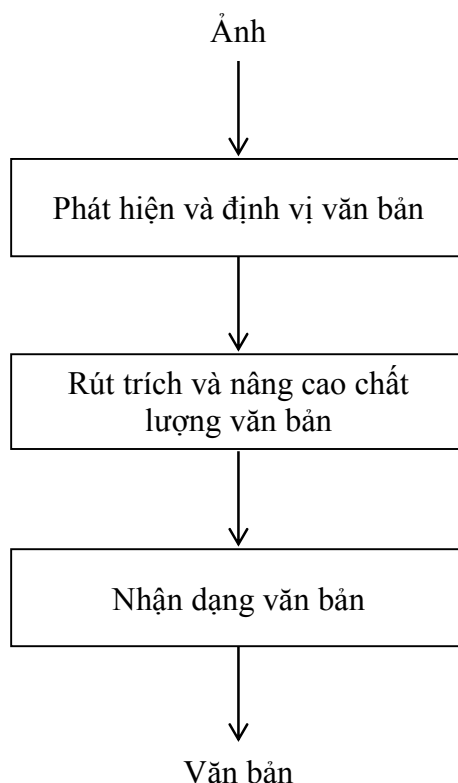


Hình 2.4 Minh họa ảnh có độ tương phản thấp

Các thách thức trên đã tạo nên một vấn đề quan trọng là lỗ hổng ngữ nghĩa giữa văn bản có trong ảnh và kết quả nhận dạng từ các phần mềm OCR. Việc thiết kế và xây dựng các hệ thống phát hiện và rút trích văn bản có khả năng khắc phục được những thách thức đã nêu là rất cần thiết để lấp đầy lỗ hổng ngữ nghĩa đó.

2.2 Tình hình nghiên cứu trong lĩnh vực phát hiện văn bản trong ảnh

Nhiều nghiên cứu trong lĩnh vực rút trích văn bản từ ảnh chụp các bảng tên đường, bằng lái xe, bìa sách, ảnh tự nhiên,... đã được công bố và đạt được một số kết quả nhất định. Nhìn chung, một hệ thống phát hiện và nhận dạng văn bản trong ảnh thường có các giai đoạn: phát hiện và định vị văn bản, rút trích và nâng cao chất lượng văn bản, nhận dạng văn bản (Hình 2.5). Các phương pháp phát hiện và định vị văn bản có thể được chia thành 3 nhóm như sau: dựa trên thành phần liên kết, dựa trên cạnh, dựa trên đặc trưng vân.



Hình 2.5 Các bước thực hiện trong hệ thống phát hiện và nhận dạng văn bản

Các phương pháp dựa trên thành phần liên kết (connected component – based)

Các phương pháp này dựa vào giả thiết các vùng văn bản có tính nhất quán về đặc trưng nào đó, ví dụ có màu sắc tương tự nhau. Thông thường các phương pháp trong hướng tiếp cận này bao gồm các bước xử lý chính như sau:

- i) Tiền xử lý ảnh (giảm nhiễu)
- ii) Gom nhóm các điểm ảnh tương đồng để phát sinh các thành phần liên kết (các ký tự ứng viên)
- iii) Tinh lọc các đối tượng ứng viên bằng các luật heuristic (kích thước, số lượng, ...)
- iv) Nhóm các thành phần liên kết thành vùng văn bản (dòng hoặc từ).

Một số tác giả nổi bật trong hướng tiếp cận này: Lienhart, Nobou Ezaki, Basilios Gatos ... Trong [6], Ezaki et al. đã đề xuất bốn mô hình phát hiện văn bản dựa vào các thành phần liên kết. Mô hình hiệu quả nhất được chứng minh gồm các bước xử lý sau: tạo ảnh biên cạnh bằng bộ lọc Sobel, nhị phân hóa ảnh bằng Otsu, phát sinh các thành phần liên kết và cuối cùng lọc các thành phần liên kết bằng các qui luật. Ưu điểm của các phương pháp này là đơn giản, nhanh và dễ cài đặt. So sánh với các phương pháp dựa trên văn, các phương pháp trong hướng tiếp cận này tính toán nhanh hơn, ít nhạy cảm đối với vấn đề về kích thước văn bản. Tuy nhiên hiệu quả của phương pháp không cao vì trong ảnh có rất nhiều thành phần giống văn bản nếu chỉ dựa vào đặc trưng về màu sắc, và gặp nhiều khó khăn trong trường hợp văn bản được nhúng trên nền phức tạp.

Các phương pháp dựa trên cạnh (edge – based)

Các phương pháp dựa trên cạnh nhìn chung khá giống với các phương pháp dựa trên thành phần liên kết. Điểm khác biệt là đặc trưng được sử dụng là cạnh thay vì màu sắc. Các phương pháp này dựa trên sự tương phản giữa văn bản và vùng nền xung quanh để định vị các vùng văn bản trong ảnh. Các bước thực hiện chính như sau:

- i) Sử dụng bộ lọc biên cạnh để xác định thành phần biên cạnh của ảnh. Các bộ lọc biên cạnh thường dùng là Canny, Sobel, Robert, Prewitt...
- ii) Gom nhóm và nối kết các thành phần biên cạnh, phát sinh vùng văn bản ứng viên.
- iii) Dùng các luật heuristic để loại bỏ các vùng không phải văn bản.

Trong nhóm tiếp cận này, có thể kể đến một số tác giả như Datong Chen, Qixiang Ye ...Chen et al.[1] đã dùng bộ lọc Canny để phát hiện ảnh biên cạnh. Tiếp theo, phép giãn nở (dilation) được sử dụng để kết nối các cạnh thành các cluster (vùng văn bản ứng viên). Một vài luật heuristic như tỉ lệ chiều rộng/chiều cao, kích thước văn bản được dùng để loại bỏ các vùng không phải văn bản. Trong [20], đặc trưng cạnh Sobel và các toán tử hình thái học cũng được áp dụng để xác định các vùng có mật độ cạnh đông đúc. Các vùng văn bản ứng viên được phát sinh dựa vào các qui luật từ thực nghiệm. Các phương pháp trong nhóm này có ưu điểm là nhanh, đơn giản, có thể cho độ phủ cao. Tuy nhiên, nhược điểm của các phương pháp này là độ chính xác không cao khi phần nền trong ảnh cũng có các cạnh tương tự như văn bản.

Các phương pháp dựa trên đặc trưng vân (texture – based)

Các phương pháp trong hướng tiếp cận này phân biệt văn bản với các thành phần khác sử dụng đặc trưng vân. Các phương pháp này thực hiện dựa trên đặc điểm là vùng văn bản trong ảnh thường có thuộc tính vân đặc thù để phân biệt với vùng nền. Các phương pháp dựa trên hướng tiếp cận này thường bao gồm các bước:

- i) Rút trích đặc trưng vân. Các đặc trưng thường dùng có thể kể đến là wavelet [21], bộ lọc Gabor, hệ số DCT, phương sai không gian [1], HOG [7]...
- ii) Thiết kế các bộ phân lớp để xác định vùng nào chứa văn bản, vùng nào không chứa văn bản. Một số phương pháp máy học thường được dùng để huấn luyện bộ phân lớp như mạng neural, SVM [21], Adaboost[2]...
- iii) Phát sinh vùng văn bản ứng viên sau khi đi qua các bộ phân lớp.

Chen et al.[2] tính toán cường độ màu trung bình và thống kê số lượng các điểm biên cạnh trong ảnh từ các mẫu huấn luyện. Các đặc trưng này được sử dụng trong bộ lọc Adaboost để phân loại các vùng ứng viên. Trong[21], Ye et al. sử dụng đặc trưng từ các hệ số wavelet và phân lớp các dòng văn bản ứng viên bằng SVM.

Ưu điểm của các phương pháp này là độ chính xác cao. Tuy nhiên, độ phức tạp tính toán rất lớn vì cần phải quét ảnh với nhiều độ phân giải khác nhau. Ngoài ra, hiệu quả của những phương pháp này phụ thuộc nhiều vào việc lựa chọn tập dữ liệu huấn luyện.

Nhiều tác giả cũng đã đề xuất các phương pháp kết hợp từ các hướng tiếp cận khác nhau nhằm nâng cao hiệu quả của hệ thống. Hầu hết các phương pháp này đều không giải quyết triệt để các yếu tố khác nhau ảnh hưởng đến hiệu quả của hệ thống như vấn đề về ngôn ngữ, kiểu chữ, kích thước, màu sắc, vùng nền phức tạp.

2.3 Các phương pháp rút trích văn bản

Đặc điểm của các phần mềm OCR là được thiết kế để nhận dạng các ký tự chữ in và hiệu quả phụ thuộc vào việc phân đoạn chính xác giữa văn bản và các điểm ảnh thuộc vùng nền. Việc phân đoạn này được thực hiện một cách dễ dàng trong các tài liệu in vì chúng có độ phân giải cao và văn bản thường có màu đen tương phản trên nền trắng. Tuy nhiên, đối với các ảnh tự nhiên thì điều đó hoàn toàn không dễ thực hiện.

Hầu hết các vùng văn bản đã phát hiện và định vị được trong ảnh tự nhiên đều có chất lượng không tốt, độ phân giải thấp và thường nhúng trên nền phức tạp. Điều đó là nguyên nhân khiến các phần mềm OCR không dễ dàng nhận ra các ký tự trong ảnh tự nhiên. Vì thế, sau khi định vị văn bản, người ta thường cố gắng làm tăng chất lượng của văn bản trong ảnh và loại bỏ phần nền từ các vùng văn bản đã phát hiện, nhằm phục vụ cho quá trình nhận dạng văn bản được tốt hơn.

Các phương pháp rút trích văn bản thường dựa vào việc nhị phân hóa các vùng ảnh đã phát hiện bằng cách sử dụng các ngưỡng toàn cục và ngưỡng cục bộ. Một số phương pháp nhị phân thường dùng là:

- Phương pháp ngưỡng toàn cục của Otsu.
- Phương pháp ngưỡng cục bộ của Niblack, Sauvola.
- Phương pháp ngưỡng thích nghi của Bradley.

2.4 Tình hình nghiên cứu trong lĩnh vực truy vấn ảnh

Trong lĩnh vực truy vấn ảnh, nhiều mô hình truy vấn đã được đề xuất và đã có những kết quả đáng kể[4]. Tiêu biểu là hệ thống truy vấn ảnh nổi tiếng của Yahoo, Google,...Mức độ cơ bản của truy vấn ảnh là truy vấn ảnh dựa vào từ khóa. Để tổ chức dữ liệu ảnh, người ta chú thích thủ công trên tập ảnh, sau đó truy vấn dựa vào từ khóa đã chú thích. Kết quả truy vấn dựa vào việc so khớp từ khóa truy vấn và từ khóa chú thích. Khi lượng dữ liệu ngày càng tăng, việc chú thích ảnh thủ công không thể đáp ứng được, đồng thời không khai thác được tối đa nội dung ẩn chứa trong ảnh và chất lượng phụ thuộc vào ý chủ quan của người chú thích.

Đối với các mô hình truy vấn ảnh dựa vào nội dung, để tổ chức dữ liệu, người ta thường rút trích các đặc trưng thị giác như màu sắc, vân, hình dáng của các đối tượng. Các mô hình này có thể được sử dụng trong các trường hợp không thể dùng từ khóa để diễn đạt. Tuy nhiên, khuyết điểm của phương pháp này là sự tương đồng về đặc trưng thị giác không dẫn đến sự tương đồng về mặt ngữ nghĩa.

Trong hệ thống truy vấn ảnh ở mức ngữ nghĩa, người ta tìm cách gán ngữ nghĩa vào ảnh dựa vào một số mô hình như dịch máy, chú thích ảnh tự động, máy học... Các mô hình này chủ yếu vẫn dựa vào các đặc trưng thị giác, các đối tượng, các vùng ảnh trong ảnh để gán ngữ nghĩa một cách tự động cho ảnh. Tuy nhiên, việc gán ngữ nghĩa cho ảnh với độ chính xác cao là không dễ dàng. Ngoài ra, trong các mô hình này, người ta chỉ truy vấn dựa vào từ khóa, nhưng không phải lúc nào các khái niệm muốn truy vấn cũng có thể diễn đạt được bằng từ khóa.

Trong các mô hình truy vấn ảnh hiện có, chưa có hệ thống nào quan tâm khai thác nội dung văn bản xuất hiện trong ảnh – một đối tượng mang nhiều thông tin ngữ nghĩa, và sử dụng cho mục đích tổ chức dữ liệu và truy vấn.

2.5 Hướng tiếp cận

Trong luận văn này, chúng tôi đề xuất mô hình phát hiện và rút trích văn bản dựa trên sự kết hợp giữa các hướng tiếp cận đã nêu trên. Qua phần khảo sát tình hình nghiên cứu trong lĩnh vực phát hiện văn bản trong ảnh, chúng ta có thể thấy để thiết kế hệ thống phát hiện và rút trích văn bản nếu chỉ dùng một trong các phương pháp đã nêu thì rất khó đạt được hiệu quả như mong muốn. Chúng ta cần tận dụng ưu điểm của mỗi phương pháp để xây dựng một mô hình có hiệu quả cao hơn. Đầu tiên, ảnh ban đầu được tiền xử lý ảnh thông qua phép reconstruction để loại bỏ phần lớn các vùng nền trong ảnh, đồng thời làm nổi bật lên các vùng ảnh có khả năng là văn bản. Các đặc trưng cạnh và các toán tử hình thái học cũng được áp dụng để phát sinh các vùng văn bản ứng viên. Từ các vùng văn bản ứng viên, các thành phần liên kết được rút trích bằng đặc trưng độ rộng nét từ phương pháp Stroke Width Transform (SWT) được đề xuất trong [5]. Điểm khác biệt trong phương pháp SWT là đặc trưng được sử dụng để gom nhóm các thành phần liên kết là sự tương đồng về độ rộng nét thay vì sự tương đồng về màu sắc như hầu hết các phương pháp trong hướng tiếp cận dựa trên các thành phần liên kết. Các thành phần liên kết sau đó được gom nhóm để tạo thành các từ ứng viên. Cuối cùng, bộ phân lớp SVM được sử dụng để tinh lọc các từ ứng viên. Phần văn bản trong các vùng ảnh đã phát hiện được rút trích bằng phương pháp nhị phân hóa mới được đề xuất dựa trên ảnh SWT tìm được trong giai đoạn phát hiện văn bản.

Đối với hệ thống truy vấn ảnh dựa vào văn bản ngoại cảnh, các ảnh sau khi phát hiện, rút trích và nhận dạng văn bản sẽ được chú thích tự động bằng chính các từ khóa đã nhận dạng được. Chúng tôi cho phép người dùng truy vấn bằng hai cách: truy vấn bằng từ khóa và truy vấn bằng ảnh chứa từ khóa mong muốn. Điểm khác biệt trong cách thức truy vấn bằng từ khóa so với các mô hình truy vấn khác là hỗ

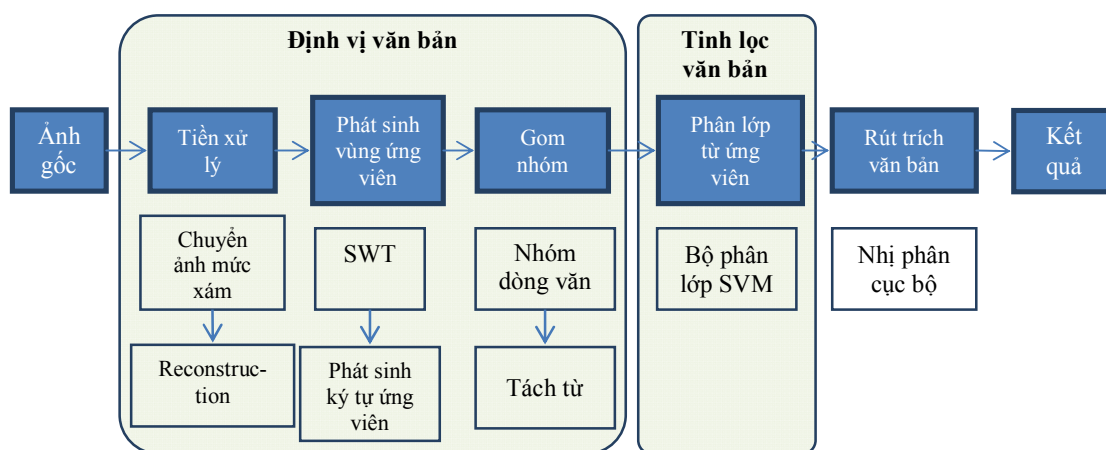
trợ người dùng tìm kiếm các ảnh có xuất hiện từ khóa truy vấn thay vì chỉ dựa vào các đặc trưng thị giác như các mô hình truy vấn trước đây. Trong trường hợp không thể sử dụng từ khóa để truy vấn do người dùng không biết ngôn ngữ của từ khóa (ví dụ khách du lịch không biết ngôn ngữ địa phương), hoặc các thiết bị nhập không hỗ trợ, chúng tôi cho phép người sử dụng đưa vào các ảnh có chứa từ khóa cần truy vấn.

Chương 3 Mô hình phát hiện và rút trích văn bản ngoại cảnh trong ảnh

Trong chương này, chúng tôi trình bày mô hình phát hiện và rút trích văn bản ngoại cảnh cùng phương pháp hiệu chỉnh kết quả nhận dạng từ phần mềm OCR.

3.1 Sơ đồ chung

Mô hình phát hiện và rút trích văn bản đề xuất gồm các bước được minh họa trong Hình 3.1. Mô hình gồm hai bước xử lý quan trọng là định vị văn bản và tinh lọc các vùng văn bản ứng viên. Trong giai đoạn định vị, đầu tiên ảnh gốc được chuyển đổi sang ảnh mức xám và được tiền xử lý để loại bỏ các đối tượng nhiễu. Tiếp theo, các toán tử hình thái học được áp dụng để phát sinh các vùng liên kết. Trong mỗi vùng văn bản ứng viên, đặc trưng độ rộng nét đề xuất trong [5] được sử dụng để tạo thành các ký tự. Các ký tự ứng viên sau đó sẽ được gom nhóm lại thành các dòng văn bản và cuối cùng các dòng văn bản được tách thành các từ. Trong giai đoạn tinh lọc văn bản, một bộ phân lớp SVM được huấn luyện dựa vào đặc trưng HOG được sử dụng để lọc lại các từ ứng viên đã tạo thành. Kết quả của giai đoạn phát hiện và định vị văn bản là tập hợp các hình chữ nhật bao quanh các từ có trong ảnh phát hiện được. Trong giai đoạn rút trích văn bản, các vùng ảnh chứa văn bản được nhị phân hóa bằng thuật toán nhị phân dựa vào ảnh SWT để loại bỏ phần nền.



Hình 3.1 Sơ đồ các bước thực hiện trong mô hình phát hiện và rút trích văn bản

3.2 Tiền xử lý

Tiền xử lý (nhị phân hóa, giảm nhiễu) là bước đầu tiên trong hầu hết các phương pháp theo hướng tiếp cận dựa vào các thành phần kết nối. Giai đoạn này cũng đóng một vai trò quan trọng trong toàn bộ hệ thống phát hiện văn bản. Nếu kết quả trong giai đoạn này không hợp lý thì hiệu quả của hệ thống sẽ bị ảnh hưởng đáng kể. Để giải quyết vấn đề này, chúng tôi áp dụng phép biến đổi hình thái học *reconstruction*[19] để loại bỏ các thành phần nền và làm nổi rõ các thành phần kết nối bên trong ảnh trước khi thực hiện các bước xử lý tiếp theo.

Trong các ảnh tự nhiên, thông tin thường tập trung nhiều bên trong ảnh hơn là ở các vùng biên của ảnh. Bước xử lý này được thực hiện nhằm loại bỏ các đối tượng kết nối với các biên của ảnh và có cường độ sáng hơn vùng xung quanh. Khi các vùng có cường độ sáng hơn vùng nền xung quanh và kết nối với các biên của ảnh được loại bỏ thì phần lớn các đối tượng nhiễu xuất hiện trong ảnh cũng được loại bỏ. Phép *reconstruction* của một ảnh g từ ảnh khởi tạo (marker) f ($\mathcal{D}_f = \mathcal{D}_g$ và $f \leq g$) được định nghĩa là phép giãn nở (dilation) của f đối với $gR_g^\delta(f) = \delta_g^{(i)}(f)$ được lặp lại cho đến khi bền vững $\delta_g^{(i)}(f) = \delta_g^{(i+1)}(f)$ (nghĩa là không có điểm ảnh nào trong f thay đổi giá trị).

Bảng 3.1 Thuật toán reconstruction cơ bản

Thuật toán reconstruction

Input: ảnh I ở dạng mức xám hoặc nhị phân, ảnh J (ảnh *marker*) cùng kích thước với I .

Output: ảnh J

- Khởi tạo ảnh K cùng kích thước với I
- Lặp lại cho đến khi ổn định (không có pixel nào trong J thay đổi giá trị)
 - Bước dilation: Với mỗi pixel $p \in I$

$$K(p) \leftarrow \max \{J(q), q \in N_G(p) \cup \{p\}\}$$

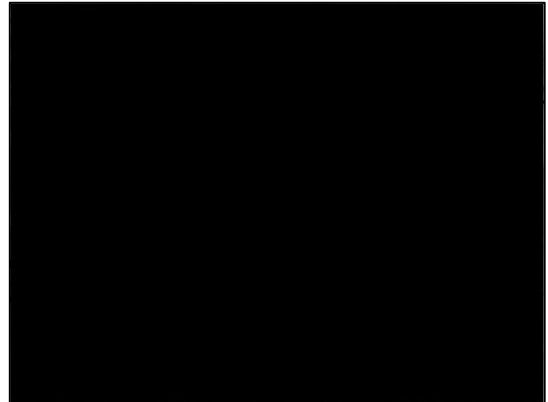
$N_G(p)$: các pixel trong vùng lân cận G của p

– Bước minimum: Với mỗi pixel $p \in I$

$$J(p) \leftarrow \min \{K(p), I(p)\}$$



a)



b)



c)



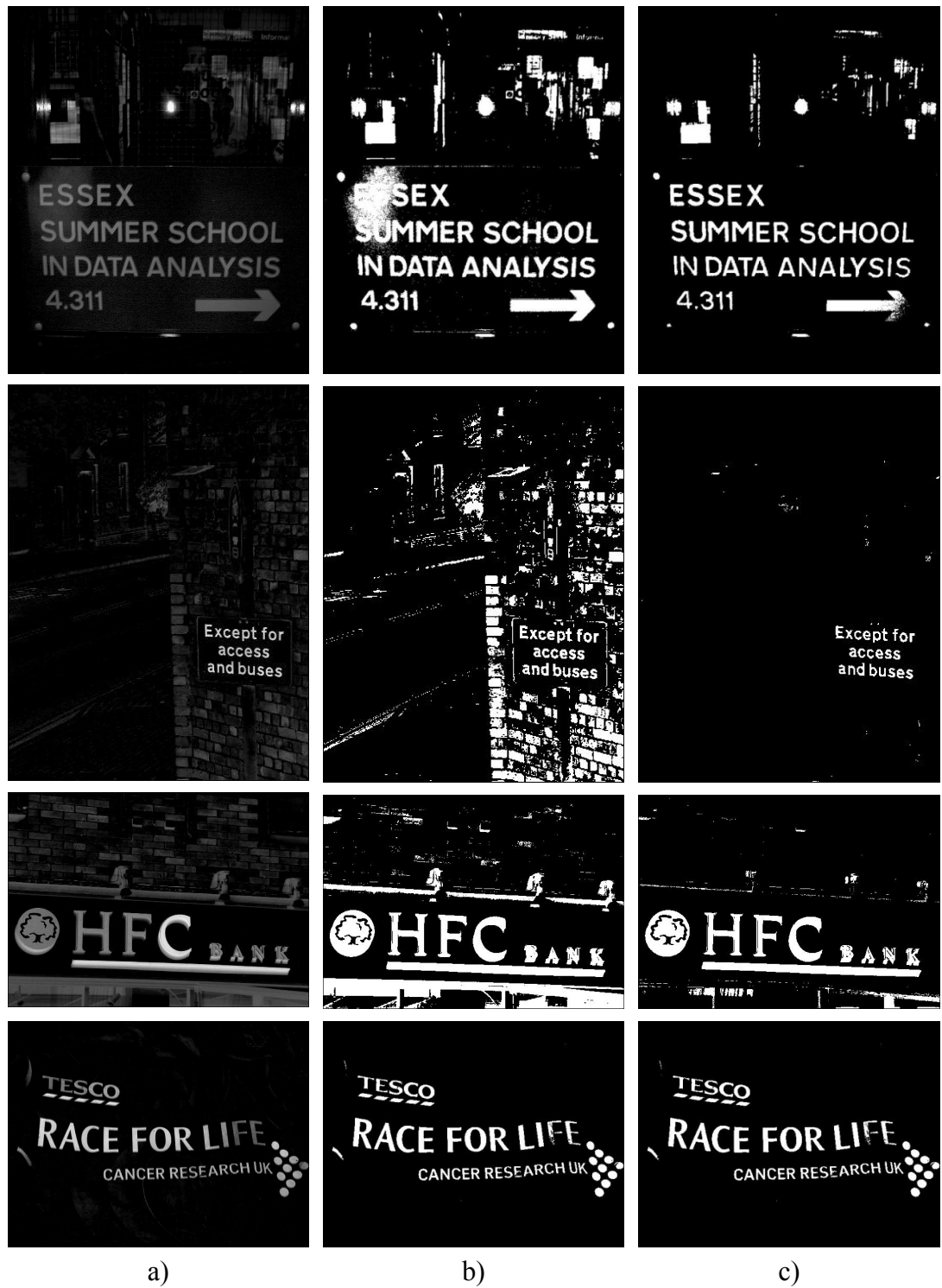
d)

Hình 3.2 a) Ảnh mức xám ban đầu I; b) Ảnh khởi tạo J; c) Kết quả phép reconstruction của ảnh a); d) Kết quả khi lấy ảnh a - c

Theo Soille[16], phép biến đổi *reconstruction* có thể được dùng để rút trích các đối tượng liên kết có cường độ sáng hơn so với phần nền xung quanh trong ảnh gốc (ảnh *mask*), khi ta chọn ảnh *marker* có giá trị 0 ở tất cả các pixel ngoại trừ các pixel ở biên. Tại biên của ảnh *marker*, các pixel sẽ có giá trị bằng với giá trị ở cùng vị trí với pixel trong ảnh gốc. Thuật toán *reconstruction* cơ bản được trình bày

trong Bảng 3.1. Kết quả thực hiện của giai đoạn này được minh họa trong Hình 3.2. Hình 3.2a) là ảnh mức xám ban đầu I , Hình 3.2b) là ảnh khởi tạo J trong đó giá trị tại các điểm ảnh ở bốn biên bằng với giá trị của các điểm ảnh ở cùng vị trí trong ảnh I . Hình 3.2c) là kết quả của ảnh J sau khi thực hiện phép *reconstruction* trên ảnh I . Có thể nhận thấy trong Hình 3.2c), các đối tượng có cường độ sáng hơn vùng nền xung quanh và không liên kết với vùng biên của ảnh đã được làm mờ, trong khi các đối tượng liên kết với vùng biên được giữ lại. Khi đó, thực hiện phép trừ giữa ảnh gốc I và ảnh *reconstruction* J : $I - J$, ta thu được ảnh kết quả của giai đoạn này, trong đó phần lớn vùng nền và các đối tượng nhiễu đã được loại bỏ, chỉ giữ lại hầu hết các đối tượng có khả năng là các vùng văn bản trong ảnh. Vì văn bản có thể có màu sáng hoặc tối hơn so với vùng xung quanh, nên phương pháp này được áp dụng hai lần trên ảnh gốc và ảnh âm bản để phát hiện các vùng văn bản có màu sáng và tối. Thực nghiệm chúng tôi cài đặt *reconstruction* dựa trên thuật toán được mô tả trong [14].

Ảnh *reconstruction* sau đó được nhị phân hóa bằng ngưỡng T_{bin} để thu được ảnh nhị phân. Đặc điểm của các ảnh đã *reconstruction* là các đối tượng có khả năng là văn bản trong ảnh đã được làm nổi rõ và có cường độ tương đối sáng. Chính vì đặc điểm này, phương pháp nhị phân dựa vào ngưỡng toàn cục Otsu hay phương pháp nhị phân dựa vào ngưỡng cục bộ Niblack tỏ ra không hiệu quả khi áp dụng trên các ảnh đã *reconstruction*. Các phương pháp này thường giữ lại quá nhiều các đối tượng nhiễu không mong muốn, hoặc đôi khi lại loại bỏ các đối tượng văn bản thực. Để khắc phục vấn đề này, chúng tôi dựa vào thực nghiệm khảo sát các giá trị điểm ảnh của các vùng văn bản trong ảnh đã *reconstruction* để quyết định một ngưỡng T_{bin} hợp lý. Giá trị ngưỡng được chọn cần không quá cao để đảm bảo giữ lại hầu hết các vùng văn bản thực, cũng như không quá thấp để hạn chế các đối tượng nhiễu xung quanh. Qua thực nghiệm, giá trị ngưỡng cho ảnh nhị phân đạt kết quả tốt nhất là $T_{bin} = 80$. Hình 3.3 minh họa một số kết quả so sánh khi nhị phân ảnh bằng phương pháp Otsu và nhị phân bằng ngưỡng T_{bin} .



Hình 3.3 So sánh kết quả các phương pháp nhị phân ảnh. a) Ảnh kết quả reconstruction; b) Nhị phân bằng phương pháp Otsu; c) Nhị phân bằng ngưỡng T_{bin}

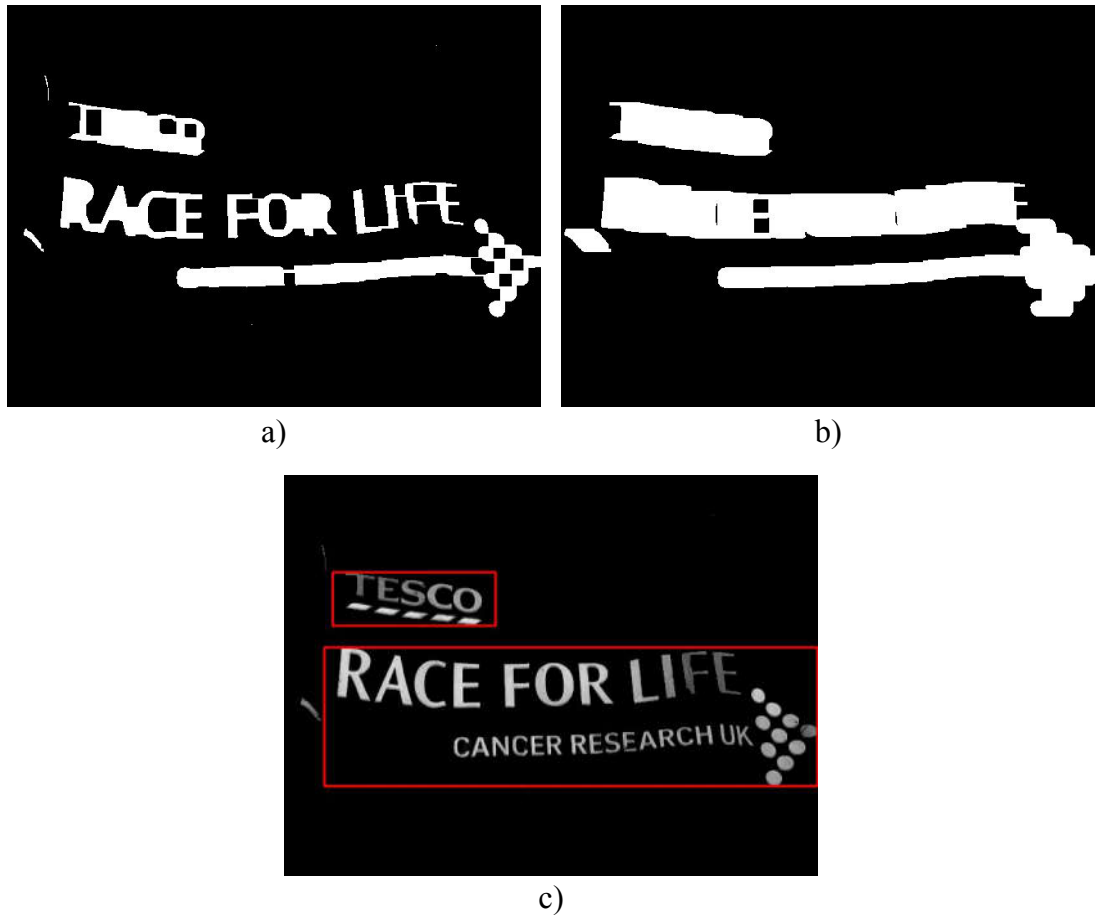
3.3 Phát sinh vùng văn bản ứng viên

Để loại bỏ các thành phần có kích thước quá nhỏ và kết nối các đối tượng gần nhau trong ảnh nhị phân, chúng tôi sử dụng hai toán tử hình thái: phép đóng ảnh (closing) và phép giãn nở (dilation). Trong các ảnh có độ tương phản thấp, một ký tự có thể bị tách ra thành nhiều mảnh. Vì vậy, đầu tiên chúng ta cần kết nối các thành phần này lại để tránh những sai lầm khi phân tích các thành phần kết nối. Chúng tôi sử dụng toán tử đóng ảnh với cấu trúc 13x13 trên ảnh nhị phân thu được từ bước xử lý trước để giải quyết vấn đề này.

Bước tiếp theo, toán tử giãn nở được áp dụng để kết nối các điểm ảnh thành các vùng văn bản ứng viên. Toán tử giãn nở thường được sử dụng để mở rộng và kết nối các vùng gần nhau. Trong phương pháp đề xuất, chúng tôi sử dụng toán tử giãn nở với cấu trúc 33x1 để kết nối các thành phần thành các vùng văn bản ứng viên. Một số quy luật heuristic được áp dụng để lựa chọn các vùng văn bản ứng viên cuối cùng. Các vùng văn bản ứng viên được chọn thỏa mãn các điều kiện sau:

- Tỷ lệ chiều cao/chiều rộng của vùng ứng viên không nhỏ hơn ngưỡng $T_{hv} = 0.5$.
- Chiều cao của vùng ứng viên không nhỏ hơn 8 và diện tích không nhỏ hơn 300. Điều kiện này để loại bỏ các đối tượng có kích thước quá bé, cho kết quả không tốt trong giai đoạn nhận dạng.
- Số lượng điểm ảnh biên cạnh trung bình trên từng dòng không nhỏ hơn 8.

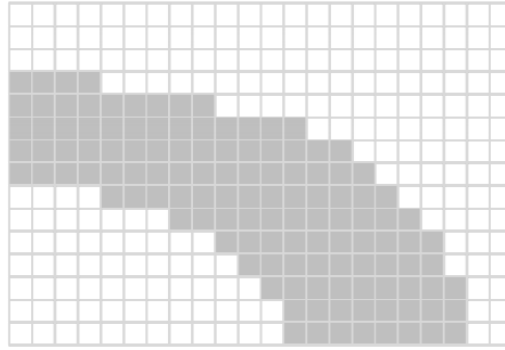
Hình 3.4 minh họa các kết quả thực hiện của giai đoạn này, trong đó Hình 3.4c) là các vùng văn bản ứng viên cuối cùng thỏa mãn các điều kiện được lựa chọn (được đánh dấu bằng hình chữ nhật màu đỏ bao quanh).



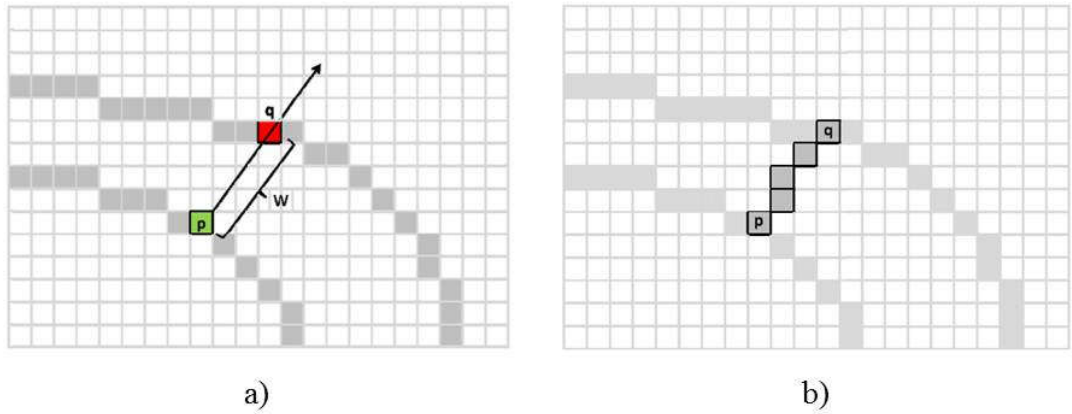
Hình 3.4 a) Kết quả thực hiện toán tử đóng trên ảnh nhị phân; b) Thực hiện phép giãn nở trên ảnh a); c) Các vùng văn bản ứng viên được lựa chọn

3.3.1 Phát sinh các ký tự ứng viên bằng SWT

Trong mỗi vùng văn bản ứng viên phát sinh từ bước xử lý trước đó, chúng tôi sử dụng SWT được đề xuất trong [5] để tạo thành các thành phần liên kết (cũng là các ký tự ứng viên). Esphtein et al. [5] đã chứng tỏ đặc trưng độ rộng nét là đủ mạnh để có thể phân biệt được văn bản với các đối tượng khác. Nét là một thành phần của ảnh gồm các điểm ảnh liên tiếp tạo thành một nhánh có độ rộng hầu như không thay đổi (Hình 3.5). Phép biến đổi độ rộng nét sẽ chuyển đổi từ một ảnh mức xám sang ảnh SWT, trong đó giá trị tại mỗi điểm ảnh chính là độ rộng của nét chứa điểm ảnh đó. Ảnh SWT ban đầu được khởi tạo với các điểm ảnh có giá trị ∞ (thực nghiệm là -1).



Hình 3.5 Minh họa đường nét trong ảnh[5]

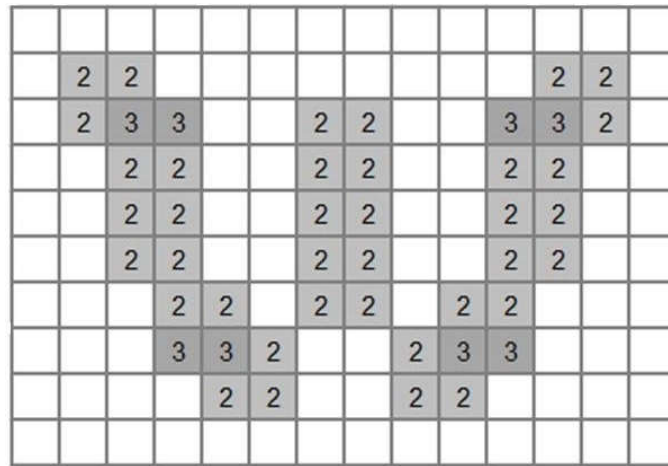


Hình 3.6 Các bước tìm độ rộng nét[5]

🔧 Cách tìm độ rộng nét

Để tìm độ rộng của các nét trong ảnh, đầu tiên ta sử dụng bộ lọc biên cạnh Canny để phát hiện biên (Hình 3.6a). Gọi $d_p \in [-\pi, \pi)$ là hướng của vector gradient tại mỗi điểm ảnh thuộc biên cạnh p . Nếu p nằm trên đường biên nét thì d_p trực giao với hướng của nét. Từ p , dịch chuyển theo hướng d_p , đi qua các điểm ảnh có dạng $r = p + nd_p, n > 0$ cho đến khi tìm được điểm ảnh q cũng thuộc biên cạnh. Nếu hướng của vector gradient d_q tại q ngược với d_p thì khoảng cách từ p đến q , có độ lớn $\|\overrightarrow{p-q}\|$, chính là độ rộng của nét chứa p và q . p và q được gọi là có hướng ngược nhau nếu độ chênh lệch giữa d_p và d_q không vượt quá $\pi/2$, nghĩa là $|d_p - (-d_q)| \leq \pi/2$. Đây là điểm khác biệt so với phương pháp ban đầu từ [5], chúng tôi cho phép ngưỡng chênh lệch lên đến $\pi/2$ thay vì $\pi/6$ như trong [5]. Khi

đó, các điểm ảnh tìm được trong quá trình dịch chuyển từ p đến q , sẽ có giá trị trong ảnh SWT là khoảng cách từ p đến q (Hình 3.6b). Như vậy, thông qua phép biến đổi SWT, ta thu được một ảnh mới có cùng kích thước với ảnh ban đầu, trong đó giá trị tại mỗi điểm ảnh trong ảnh mới là độ rộng của nét chứa điểm ảnh đó. Nếu một điểm ảnh thuộc về nhiều nét, thì giá trị tại điểm ảnh đó là độ rộng của nét có độ rộng nhỏ nhất. Hình 3.7 minh họa giá trị của các điểm ảnh trong ảnh SWT của ký tự “W”.



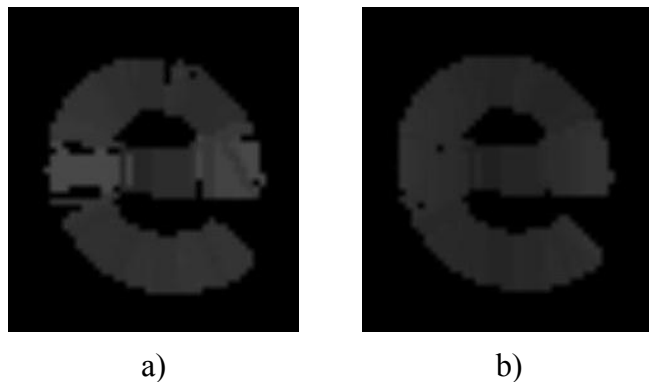
Hình 3.7 Minh họa ảnh SWT cho ký tự W

Một vấn đề đối với phương pháp SWT của Esphtein đó là chi phí tính toán cho SWT là khá lớn khi ảnh chứa nội dung phức tạp. Chi phí cho việc tính toán SWT phụ thuộc vào số lượng điểm ảnh biên cạnh và độ rộng nét tối đa trong ảnh. Vì vậy, khi ảnh có nền phức tạp, số lượng điểm ảnh thuộc biên cạnh sẽ rất lớn dẫn đến thời gian thực hiện SWT cũng rất lớn. Trong mô hình này, chúng tôi chỉ áp dụng SWT trên những vùng văn bản ứng viên đã chọn lọc từ giai đoạn trước. Ngoài ra, kết quả ảnh SWT được tạo ra từ [5] đối với các nét cong hay nét hợp thường xuất hiện những lỗ hổng không mong đợi trong quá trình dịch chuyển các điểm ảnh để tìm độ rộng nét (Hình 3.8a). Để giải quyết vấn đề này, chúng tôi thực hiện bước làm mịn ảnh SWT như sau.

Với mỗi điểm ảnh foreground $I(x, y)$ (có giá trị trong ảnh mức xám > 0) chưa được gán độ rộng nét ($SWT(x, y) = \infty$):

- Gọi $G(x, y)$ là vùng lân cận gồm 8 láng giềng của $I(x, y)$
- $m_{SWT}(x, y)$ là giá trị median của độ rộng nét của các điểm ảnh trong vùng $G(x, y)$
- $m(x, y)$ là giá trị cường độ trung bình của các điểm ảnh trong $G(x, y)$
- Nếu $|I(x, y) - m(x, y)| < 40$, thì giá trị median được gán làm độ rộng nét của điểm ảnh đang xét: $SWT(x, y) = m_{SWT}(x, y)$

Hình 3.8 minh họa các kết quả ảnh SWT của ký tự “e” trước và sau khi làm mịn.



Hình 3.8 a) Ảnh SWT của ký tự “e” trước khi làm mịn; b) Ảnh SWT của ký tự “e” sau khi làm mịn

Gom nhóm các điểm ảnh thành các ký tự

Sau khi thực hiện phép biến đổi độ rộng nét và thu được ảnh SWT, bước tiếp theo của thuật toán là gom nhóm các điểm ảnh lân cận để tạo thành các thành phần kết nối (hay các ký tự ứng viên). Hai điểm ảnh kế tiếp nhau sẽ được gom nhóm nếu tỉ lệ độ rộng nét của chúng không vượt quá 3. Các ký tự ứng viên sau đó được chọn lọc lại dựa vào một số luật heuristic. Một ký tự được chọn nếu thỏa mãn các điều kiện sau:

- Trong cùng một ký tự, giá trị độ rộng nét thường có sự biến đổi không lớn, vì vậy các ký tự có độ lệch chuẩn lớn sẽ bị loại bỏ. Đây là một quy luật quan trọng dựa vào tỉ lệ giữa độ lệch chuẩn (std) và giá trị trung bình (mean) của độ rộng nét trong từng ký tự. Các ký tự cần thỏa mãn: $std/mean$

< 0.5 . Giá trị ngưỡng 0.5 đạt được từ tập huấn luyện của tập dữ liệu ảnh ICDAR 2003.

- Một quy luật khác là kiểm tra hình chữ nhật bao quanh mỗi ký tự ứng viên không chứa nhiều hơn 3 ký tự khác. Quy luật này nhằm loại bỏ các đối tượng như biển hiệu hoặc khung..., là những đối tượng cũng có độ rộng nét thường không đổi.
- Tỷ lệ chiều rộng / chiều cao không vượt quá 5.
- Ngoài ra, các thành phần có kích thước quá lớn (chiều cao lớn hơn 300 hoặc chiều rộng lớn hơn $\frac{1}{2}$ *chiều rộng của ảnh) hoặc quá bé (nhỏ hơn 8) cũng bị loại bỏ.

Các thành phần còn lại chính là các ký tự cuối cùng được chọn lọc. Hình 3.9 minh họa kết quả ảnh SWT và các ký tự được chọn lọc (được đánh dấu bằng các hình chữ nhật màu vàng).



a)



b)

Hình 3.9 a) Ảnh SWT; b) Các ký tự ứng viên được chọn lọc

3.4 Gom nhóm các thành phần liên kết

3.4.1 Nhóm các ký tự thành dòng văn bản

Trong giai đoạn này, các ký tự liên nhau được nhóm lại để tạo thành các dòng văn bản. Hầu hết văn bản thường xuất hiện với hướng ngang hoặc có độ dốc không lớn. Đặc trưng này giúp chúng ta quyết định một đối tượng có phải là ký tự thực

(thuộc về một dòng văn bản) hoặc là đối tượng nhiễu không mong đợi. Đầu tiên, chúng ta phát sinh các cặp ký tự hợp lệ dựa vào các quy luật sau.

- Các ký tự thuộc cùng một dòng thường có độ rộng nét tương tự nhau. Hai ký tự ứng viên được nhóm lại nếu tỉ lệ giá trị trung bình của độ rộng nét của chúng không vượt quá 1.5.
- Tỉ lệ chiều cao giữa các ký tự không vượt quá 2.25. Quy luật này xem xét các ký tự chữ thường và chữ hoa.
- Hai ký tự có khoảng cách không quá xa nhau. Khoảng cách giữa các ký tự không lớn hơn 2.5 lần độ rộng của ký tự có chiều rộng lớn hơn.
- Gọi $C_1(x_1, y_1)$ và $C_2(x_2, y_2)$ là tâm của hai ký tự. Độ lệch giữa y_1 và y_2 không vượt quá 0.5 lần chiều cao của ký tự có chiều cao lớn hơn: $|y_1 - y_2| < 0.5 \times \max(H_1, H_2)$, trong đó, H_1 và H_2 lần lượt là chiều cao của hai ký tự.



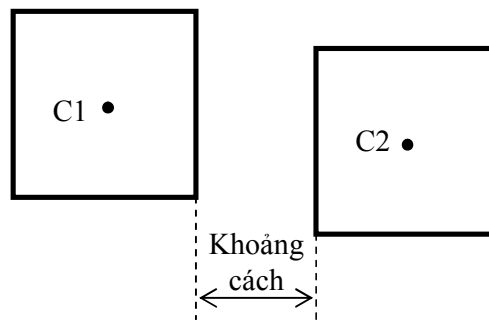
Hình 3.10 Kết quả các dòng văn bản hệ thống phát hiện được

Sau khi phát sinh các cặp ký tự ứng viên, các cặp ký tự này được kết nối lại để tạo thành các chuỗi ký tự. Ban đầu, mỗi chuỗi ký tự chỉ chứa duy nhất một cặp ký tự. Hai chuỗi ký tự sẽ được kết nối lại nếu chúng cùng một ký tự đầu-cuối và có cùng hướng. Quá trình này được lặp lại cho đến khi không còn hai chuỗi nào có thể kết nối được với nhau. Các thành phần liên kết không được gom vào bất kỳ chuỗi

ký tự nào sẽ bị loại bỏ. Giả định các dòng văn bản có ít nhất ba ký tự, các chuỗi ký tự được xem là một dòng văn bản thực nếu số lượng ký tự lớn hơn hoặc bằng 3. Các chuỗi ký tự còn lại chính là các dòng văn bản phát hiện được. Qua giai đoạn này, ta thu được các hình chữ nhật bao quanh các dòng văn bản. Hình 3.10 minh họa các dòng văn bản phát hiện được (được đánh dấu bằng các hình chữ nhật có màu xanh).

3.4.2 Tách dòng văn bản thành các từ

Trong bước xử lý trước, chúng ta đã gom nhóm các ký tự trong cùng một dòng văn bản mà không quan tâm các ký tự có thuộc cùng một từ hay không. Mục đích của giai đoạn này là tách các dòng văn bản thành các từ riêng biệt. Văn bản thường có kích thước tùy ý, độ rộng của các chữ khác nhau, khoảng cách giữa các từ, giữa các ký tự cũng không hoàn toàn bằng nhau. Những yếu tố đó gây khó khăn cho việc tính toán một ngưỡng hợp lý để phân tách chính xác các từ.



Hình 3.11 Khoảng cách giữa các hình chữ nhật bao quanh ký tự

Trong mô hình này, chúng tôi đề xuất phương pháp để tách dòng văn bản thành các từ dựa vào việc tính toán khoảng cách giữa các hình chữ nhật bao quanh các ký tự (Hình 3.11). Chúng tôi tính khoảng cách giữa các ký tự theo hướng ngang trên từng dòng và thực hiện việc phân tách từ dựa vào ngưỡng T được định nghĩa như sau:

$$T(i) = \text{mean}(D(i)) + \beta \times \text{std}(D(i)) \quad (3.1)$$

Trong đó, $D(i)$ là vector chứa khoảng cách giữa hai ký tự liên tiếp nhau trong dòng thứ i . Dựa vào việc thống kê phân bố khoảng cách (giá trị trung bình mean và

độ lệch chuẩn std) trong một dòng, một ngưỡng $T(i)$ được tính từ (3.1) được dùng để tách các từ. Nếu hai ký tự trong cùng dòng có khoảng cách lớn hơn $T(i)$, chúng sẽ được xem là thuộc về hai từ khác nhau. Qua thực nghiệm, chúng tôi tìm được giá trị β cho kết quả tốt nhất là $\beta = 1.5$. Kết quả của giai đoạn này là tập các hình chữ nhật bao quanh các từ ứng viên có trong ảnh. Hình 3.12 minh họa kết quả tách từ từ các dòng văn bản trong Hình 3.10.



Hình 3.12 Các từ ứng viên

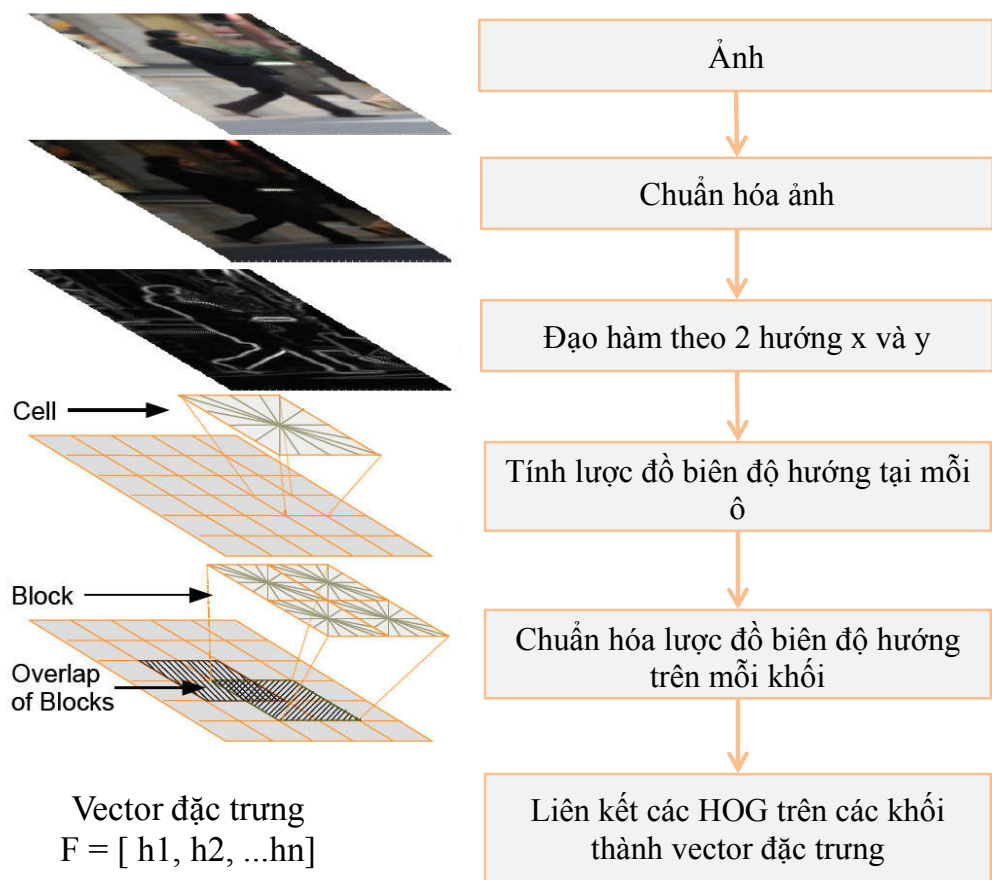
3.5 Tinh lọc các từ ứng viên bằng bộ phân lớp SVM

3.5.1 Đặc trưng HOG

Năm 2005, Dalal đã đề xuất một đặc trưng mới dựa trên Histogram of Oriented Gradient – HOG[3]. Từ đó, HOG được sử dụng một cách rộng rãi trong lĩnh vực thị giác máy tính. Ý tưởng chính của phương pháp là đặc điểm, hình dáng cục bộ của đối tượng có thể được biểu diễn tốt thông qua phân bố cường độ cục bộ của hướng cạnh. Đặc trưng HOG được tính trên cả một vùng vì vậy phù hợp cho bài toán phân lớp đối tượng. Đặc trưng HOG cho thấy hiệu quả cao trong việc đặc tả hình dáng của đối tượng và nó không nhạy cảm với những thay đổi về sự chiếu sáng. Phương pháp tổng quát để rút trích đặc trưng HOG trên một cửa sổ bất kỳ được trình bày trong Bảng 3.2.

Bảng 3.2 Thuật toán rút trích đặc trưng HOG***Thuật toán rút trích đặc trưng HOG***

- Chuẩn hóa ảnh bằng phép biến đổi gamma nhằm làm giảm ảnh hưởng của sự chiếu sáng.
- Với mỗi kênh màu, tính gradient sử dụng kernel $[-1, 0, 1]$ dọc theo hai trục x, y . Tại mỗi điểm ảnh, kênh màu có cường độ gradient lớn nhất được dùng để lấy cường độ và hướng gradient đặc trưng tại điểm ảnh đó.
- Phân hoạch cửa sổ cần tính HOG thành nhiều ô (cell) có kích thước bằng nhau. Mỗi khối (block) gồm nhiều cell, số lượng cell trong mỗi block là bằng nhau.
 - Xác định lược đồ hệ số góc cho từng cell (8x8). Mỗi cell có vector đặc trưng gồm 9 thành phần tương ứng với 9 bin, giá trị tại thành phần thứ i là tổng biên độ của vector gradient của các pixel có hệ số góc thuộc bin i của cell đó.
 - Ghép các vector đặc trưng của các cell trong từng block (2x2 cell) lại với nhau. Như vậy, vector đặc trưng của block có $9 \times 4 = 36$ thành phần.
- Chuẩn hóa theo công thức $L2 - Hys$ hoặc $L1 - sqrt$ cho mỗi block.
- Ghép các vector đặc trưng của tất cả các block trong vùng cửa sổ ảnh thành vector đặc trưng.



Hình 3.13 Quá trình rút trích đặc trưng HOG[3]

3.5.2 Bộ phân lớp SVM

Support Vector Machines – SVM được đề xuất bởi Vapnik [18] và đã cho thấy được hiệu quả tốt trong các bài toán phân lớp nhị phân trong những năm gần đây. Ưu điểm quan trọng của SVM là nó có khả năng huấn luyện một bộ phân lớp phi tuyến trong không gian đặc trưng có số chiều cao với số lượng các mẫu huấn luyện nhỏ. Mục đích của phương pháp SVM là phát sinh ra một mô hình từ tập mẫu học, mô hình này có khả năng dự đoán lớp cho các mẫu thử. SVM tìm ra một hàm quyết định phi tuyến trong tập mẫu học bằng cách ánh xạ hoàn toàn các mẫu học vào một không gian đặc trưng kích thước lớn để có thể phân lớp tuyến tính và phân lớp dữ

liệu trong không gian này bằng cách cực đại khoảng cách lề và cực tiểu lỗi học cùng một lúc.

Cho tập mẫu học được gán nhãn $\{\mathbf{x}_i, y_i\}_{i=1}^n$ trong đó $\mathbf{x}_i \in \mathbf{R}^m$ và $y_i \in \{\pm 1\}$, tồn tại siêu phẳng $\mathbf{w} \cdot \mathbf{x} + b = 0$ có khả năng phân lớp tất cả các mẫu học:

$$\mathbf{w} \cdot \mathbf{x}_i + b \geq +1 \text{ nếu } y_i = +1$$

$$\mathbf{w} \cdot \mathbf{x}_i + b \leq -1 \text{ nếu } y_i = -1$$

Hay:

$$y_i(\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1 \quad \forall i$$

trong đó, \mathbf{w} là vector trọng số của siêu phẳng phân cách. Mục tiêu của phương pháp SVM là tìm một siêu phẳng phân cách sao cho khoảng cách lề (margin) giữa hai lớp đạt cực đại. Khoảng cách lề giữa hai lớp là $\frac{2}{\|\mathbf{w}\|}$. Vấn đề cực đại khoảng

cách lề có thể đạt được bằng cách cực tiểu $\|\mathbf{w}\|^2$ (theo \mathbf{w}, b). Việc huấn luyện SVM chính là giải bài toán tối ưu có ràng buộc:

$$\min_{\mathbf{w}, b} \frac{1}{2} \|\mathbf{w}\|^2$$

sao cho:

$$y_i(\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1, \quad i = 1, \dots, n$$

Trong trường phi tuyến, vấn đề tối ưu trở thành:

$$\min_{\mathbf{w}, b} \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^n \xi_i \quad (3.2)$$

sao cho:

$$y_i(\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1 - \xi_i, \quad \xi_i \geq 0, i = 1, \dots, n$$

Trong đó, C là tham số nói lỏng, ξ_i là lỗi học của mẫu thứ i . Thay vì giải bài toán(3.2), ta giải bài toán đối ngẫu của nó:

$$\max_{\alpha} W(\alpha) = \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j) \quad (3.3)$$

thỏa mãn: $\sum_{i=1}^n \alpha_i y_i = 0$ và $0 \leq \alpha_i \leq C, i = 1, \dots, n$

trong đó, α_i là các nhân tử Lagrange, $K(\mathbf{x}_i, \mathbf{x}_j) = \Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x}_j)$ là hàm nhân (kernel), $\Phi: \mathbf{X} \rightarrow \mathbf{H}$ là một ánh xạ từ không gian đầu vào $\mathbf{X} \subset \mathbf{R}^m$ sang không gian đặc trưng \mathbf{H} có số chiều cao hơn (trong đó bài toán có thể phân lớp tuyến tính). Giải bài toán (3.3), ta thu được trọng số tối ưu:

$$\mathbf{w}^* = \sum_{i=1}^n \alpha_i^* y_i \Phi(\mathbf{x}_i)$$

Khi đó, phân lớp tối ưu có dạng:

$$\sum_{i=1}^n \alpha_i^* y_i \Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x}) + b = \sum_{i=1}^n \alpha_i^* y_i K(\mathbf{x}_i, \mathbf{x}) + b$$

Một mẫu \mathbf{x} được phân lớp theo hàm quyết định:

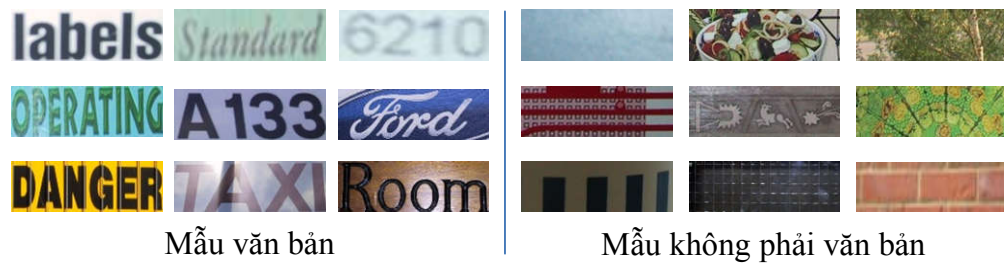
$$f(\mathbf{x}) = \text{sign} \left(\sum_{i=1}^n \alpha_i^* y_i K(\mathbf{x}_i, \mathbf{x}) + b \right)$$

Một số hàm nhân thường được sử dụng:

- Hàm RBF (Radial Basic Function): $K(\mathbf{x}_i, \mathbf{x}_j) = \exp \left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma^2} \right)$
- Đa thức bậc d: $K(\mathbf{x}_i, \mathbf{x}_j) = (1 + \mathbf{x}_i \cdot \mathbf{x}_j)^d$

3.5.3 Huấn luyện và phân lớp từ bằng bộ phân lớp SVM

Trong phương pháp đề xuất, chúng tôi lựa chọn SVM với hàm nhân RBF để phân lớp các từ ứng viên đã tìm được. Bộ phân lớp SVM được huấn luyện trên một tập dữ liệu gồm 1156 vùng ảnh chứa văn bản và 620 vùng ảnh không chứa văn bản được lấy từ tập huấn luyện của tập dữ liệu ICDAR 2003. Tất cả các mẫu huấn luyện được chuẩn hóa về kích thước 48x16 (chiều dài x chiều cao) như Hình 3.14.



Hình 3.14 Một số mẫu từ tập huấn luyện bộ phân lớp

Với mỗi bộ dữ liệu, ta tiến hành rút trích đặc trưng HOG như đã trình bày ở phần trên. Tham số sử dụng trong việc rút trích đặc trưng HOG là:

- Số bin: 9
- Kích thước ô (cell): 4x4 pixels
- Kích thước khối (block): 2x2 cells
- Độ chồng lấp giữa các khối: 4x4 pixels

Sau khi rút trích đặc trưng cho bộ dữ liệu học, tiến hành huấn luyện cho bộ phân lớp SVM. Các mẫu positive (văn bản) được gán nhãn lớp 1 và các mẫu negative (không phải văn bản) được gán nhãn lớp -1.

Với mỗi từ ứng viên đã tìm thấy, chuẩn hóa về kích thước 48x16 pixels và rút trích đặc trưng HOG trên vùng đó. Đặc trưng HOG này được đưa vào bộ phân lớp SVM đã huấn luyện để quyết định đó có phải là một từ thật sự hay không. Các từ được SVM gán nhãn lớp 1 cũng là kết quả cuối cùng của giai đoạn phát hiện và định vị văn bản (Hình 3.15).



Hình 3.15 Kết quả phát hiện văn bản của hệ thống

3.6 Rút trích văn bản

Rút trích và nâng cao chất lượng văn bản là một bước quan trọng trước khi nhận dạng văn bản. Các vùng văn bản sau khi được phát hiện và định vị cần được nhị phân hóa để các phần mềm nhận dạng ký tự có thể nhận dạng tốt hơn và cho kết quả cao hơn. Các vùng văn bản sau khi nhị phân được sử dụng làm đầu vào của bước nhận dạng văn bản dựa vào phần mềm OCR. Có thể nhận thấy việc sử dụng một ngưỡng toàn cục để nhị phân các vùng văn bản không phải là một ý tưởng tốt bởi vì văn bản trong ảnh có màu sắc khác nhau, điều kiện ánh sáng cũng khác nhau. Trong mô hình này, chúng tôi đề xuất phương pháp nhị phân các vùng văn bản dựa vào ảnh SWT để rút trích văn bản từ các ảnh vùng văn bản (từ) thu được ở bước xử lý trước. Quá trình nhị phân các vùng văn bản được thực hiện như sau:

- Với mỗi vùng văn bản R (có dạng hình chữ nhật), gọi $mean(R)$ và $std(R)$ là mức xám trung bình và độ lệch chuẩn của các điểm ảnh $(i, j) \in R$ có $SWT(i, j) > 0$.
- Gọi B_R là vùng ảnh nhị phân tương ứng với R , khi đó:

$$B_R(i, j) = \begin{cases} 0 & \text{if } T_{R,1} \leq gray(i, j) \leq T_{R,2} \\ 255 & \text{other} \end{cases}, \forall (i, j) \in R \quad (3.4)$$

Với:

$$\begin{aligned} T_{R,1} &= mean(R) - k_1 \times std(R) \\ T_{R,2} &= mean(R) + k_2 \times std(R) \end{aligned} \quad (3.5)$$

trong đó, k_1 và k_2 là các tham số từ thực nghiệm. Các điểm ảnh văn bản được gán giá trị màu đen trên nền trắng. Mã giả của thuật toán nhị phân hóa vùng văn bản được trình bày cụ thể trong Bảng 3.3.

Bảng 3.3 Thuật toán nhị phân hóa vùng văn bản**Thuật toán nhị phân hóa vùng văn bản**

Đầu vào: vùng ảnh chứa văn bản $R(x_1, y_1, x_2, y_2)$

Đầu ra: vùng ảnh chứa văn bản được nhị phân $B(x_1, y_1, x_2, y_2)$

```
// Tính cường độ trung bình của các điểm ảnh có SWT > 0
S:=0; n:=0;
for i from  $y_1$  to  $y_2$ 
    for j from  $x_1$  to  $x_2$ 
        if  $SWT(i, j) > 0$ 
             $S := S + gray(i, j)$ 
             $n := n + 1$ 
mean :=  $S / n$ 
// Tính độ lệch chuẩn
std := 0
for i from  $y_1$  to  $y_2$ 
    for j from  $x_1$  to  $x_2$ 
        if  $SWT(i, j) > 0$ 
             $std := std + (gray(i, j) - mean) \times (gray(i, j) - mean)$ 
std :=  $\sqrt{std / n}$ 
// Tạo vùng ảnh nhị phân
 $T_1 := mean - k_1 \times std$ 
 $T_2 := mean + k_2 \times std$ 
for i from  $y_1$  to  $y_2$ 
    for j from  $x_1$  to  $x_2$ 
        if  $T_1 \leq gray(i, j) \leq T_2$ 
             $B(i, j) := 0$ 
        else
             $B(i, j) := 255$ 
```



Hình 3.16 Kết quả nhị phân hóa vùng văn bản

Kết quả nhị phân các từ phát hiện được trong Hình 3.12 được minh họa trong Hình 3.16. Các ảnh nhị phân sẽ được đưa vào phần mềm OCR để nhận dạng. Thực nghiệm cho thấy, giai đoạn phân đoạn ảnh dựa vào ảnh SWT giúp làm tăng đáng kể hiệu quả nhận dạng ký tự của phần mềm OCR. Các tham số sử dụng trong thực nghiệm cho kết quả tốt nhất lần lượt là $k_1 = 1.5$ và $k_2 = 2.0$.

3.7 Hiệu chỉnh kết quả nhận dạng ký tự bằng phần mềm OCR

Như đã trình bày trong phần trước đó, đặc điểm của các phần mềm OCR là được thiết kế để nhận dạng tốt đối với các tài liệu in. Tuy nhiên, khi áp dụng cho văn bản ngoại cảnh trong ảnh thì các phần mềm OCR dễ cho ra kết quả sai (mặc dù văn bản ngoại cảnh đã được rút trích và nhị phân hóa) vì văn bản ngoại cảnh có hướng và kiểu chữ đa dạng. Để giải quyết vấn đề này, cần có phương pháp để hiệu chỉnh các kết quả nhận dạng từ phần mềm OCR để tỉ lệ chính xác cao hơn. Trong phương pháp đề xuất, chúng tôi sử dụng khoảng cách Levenshtein và độ khớp N-gram để đo sự khác biệt giữa hai chuỗi ký tự.

✚ Khoảng cách Levenshtein

Khoảng cách Levenshtein giữa hai chuỗi là số lượng thao tác tối thiểu cần để chuyển chuỗi này thành chuỗi khác, với các thao tác thêm, xóa và thay thế ký tự.

Ví dụ: khoảng cách Levenshtein giữa hai chuỗi “kitten” và “sitting” là 3, vì phải dùng ít nhất ba lần biến đổi sau:

kitten → sitten (thao tác thay thế k cho s)

sitten → sittin (thao tác thay thế **i** cho **e**)

sittin → sitting (thao tác thêm **g** vào cuối)

Để tính khoảng cách Levenshtein, ta sử dụng thuật toán quy hoạch động với độ phức tạp $O(mn)$, trong đó m và n lần lượt là chiều dài của hai chuỗi ký tự. Mã giả của thuật toán tính khoảng cách Levenshtein được trình bày trong Bảng 3.4. Hình 3.17 minh họa quá trình tính toán khoảng cách Levenshtein giữa hai chuỗi ký tự, trong đó ô màu xanh lá là chi phí ban đầu được gán giá trị 0, các ô màu vàng là thứ tự các chuyển đổi giữa hai chuỗi, ô màu xanh dương là kết quả cuối cùng cũng là khoảng cách giữa hai chuỗi ký tự.

Bảng 3.4 Thuật toán tính khoảng cách Levenshtein

```

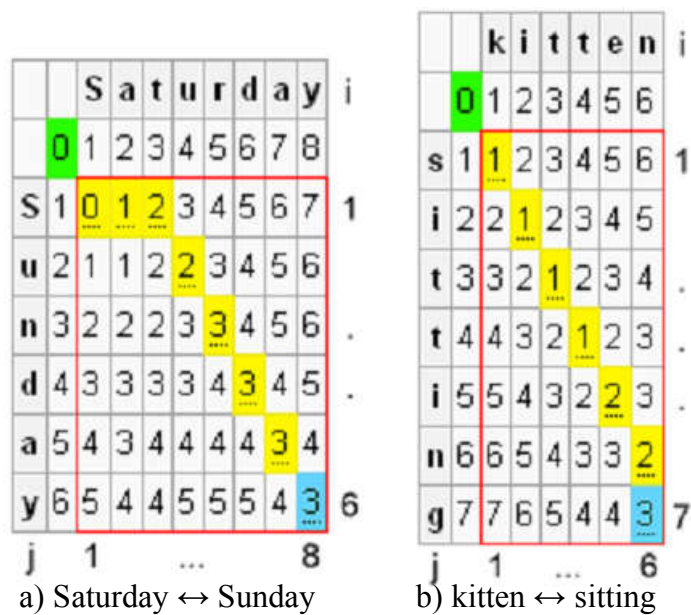
int LevenshteinDistance(char s[1..m], char t[1..n])
    // d là bảng gồm m+1 dòng và n+1 cột
    declare int d[0..m, 0..n]

    for i from 0 to m
        d[i, 0] := i
    for j from 0 to n
        d[0, j] := j

    for i from 1 to m
    for j from 1 to n
        {
            if s[i] = t[j] then cost := 0
            else cost := 1
            d[i, j] := minimum(
                                d[i-1, j] + 1,      // xoá
                                d[i, j-1] + 1,      // thêm
                                d[i-1, j-1] + cost  // thay thế
                                )
        }

    return d[m, n]

```



Hình 3.17 Minh họa các bước tính khoảng cách Levenshtein

🌈 Mô hình ngôn ngữ N-gram

Mô hình ngôn ngữ là một phân bố xác suất trên các tập văn bản. Hay nói cách khác, mô hình ngôn ngữ có thể cho biết xác suất một câu (hoặc cụm từ) thuộc một ngôn ngữ là bao nhiêu.

Cho một câu gồm n từ $S = w_1 w_2 \dots w_n$. Theo lý thuyết xác suất, xác suất của câu S được tính như sau:

$$\begin{aligned}
 P(S) &= P(w_1 w_2 \dots w_n) \\
 &= P(w_1) P(w_2 | w_1) P(w_3 | w_1 w_2) \dots P(w_n | w_1 \dots w_{n-1}) \\
 &= \prod_{k=1}^n P(w_k | w_1 \dots w_{k-1})
 \end{aligned} \tag{3.6}$$

Tuy nhiên, phương pháp này khó được áp dụng trong thực tế, vì việc tính xác suất dãy chuyển như vậy đòi hỏi rất nhiều thông tin huấn luyện, tốn khá nhiều thời gian xử lý và một lượng bộ nhớ vô cùng lớn. Mô hình N-gram được đưa ra nhằm khắc phục khó khăn trên. Một N-gram là một dãy con gồm N phần tử liên tiếp nhau của một dãy phần tử cho trước. Số phần tử trong một N-gram được gọi là bậc của

N-gram. Trong mô hình N-gram, mỗi từ được coi như phụ thuộc xác suất vào N-1 từ trước nó. Theo mô hình N-gram thì:

$$P(w_k | w_1 \dots w_{k-1}) \approx P(w_k | w_{k-N+1} \dots w_{k-2} w_{k-1}) \quad (3.7)$$

Công thức tính xác suất (3.6) được viết lại như sau:

$$P(S) = P(w_1 w_2 \dots w_n) = \prod_{k=1}^n P(w_k | w_{k-N+1} \dots w_{k-2} w_{k-1}) \quad (3.8)$$

Như vậy, để tính xác suất của từ thứ k, thay vì phải xem xét tất cả các từ trước nó trong câu, N-gram chỉ xem xét N-1 từ trước nó. Trong đó, các xác suất về sự phụ thuộc của một từ vào N-1 từ trước đó được thống kê dựa trên một tập dữ liệu. Cách này giúp việc tính toán trở nên đơn giản hơn. N-gram bậc 1 được gọi là unigram, bậc 2 được gọi là bigram, bậc 3 được gọi là trigram, ... Trong phương pháp đề xuất, chúng tôi sử dụng trigram. Thay vì sử dụng mô hình N-gram để tính xác suất của câu, chúng tôi dùng N-gram cho từ. Cho từ $w = a_1 a_2 \dots a_n$, xác suất của từ w được tính theo trigram được viết lại như sau:

$$P(w) = P(a_1 a_2 \dots a_n) = \prod_{k=1}^n P(a_k | a_{k-2} a_{k-1}) \quad (3.9)$$

Độ khớp N-gram giữa các từ

Mỗi từ được tách thành các chuỗi N ký tự liên tiếp nhau tạo thành tập N-gram của từ đó. Ví dụ tập 3-gram của từ “pocket” là {poc, ock, cke, ket}. Độ khớp N-gram giữa hai từ (*term frequency – TF*) được định nghĩa là số lượng N-gram giống nhau giữa hai từ đó.

Xem xét các từ (và các 3-gram tương ứng) sau:

p0ck@t → {p0c, 0ck, ck@, k@t}

pocket → {poc, ock, cke, ket}

rocket → {roc, ock, cke, ket}

Ta có $TF(p0ck@t, pocket) = 0$ và $TF(p0ck@t, rocket) = 0$ vì “pocket” và “rocket” đều không có 3-gram nào chung so với “p0ck@t”. Giả sử trong trường

hợp này, từ thay thế đúng cho “p0ck@t” là “pocket”. Nếu sử dụng cách tách N-gram như trên, ta không chọn được từ thay thế chính xác. Tong et al. [17] đã đề xuất thêm vào phía đầu và phía cuối của mỗi từ các ký tự giả “#” trước khi tách thành các chuỗi N-gram. Trong mô hình đề xuất, chúng tôi lựa chọn thêm vào N-1 ký tự “#”. Xem xét lại các tập 3-gram sau khi đã thêm 2 ký tự “#” vào đầu và cuối của mỗi từ.

$$##p0ck@t## \rightarrow \{##p, #p0, p0c, 0ck, ck@, k@t, @t#, t##\}$$

$$##pocket## \rightarrow \{##p, #po, poc, ock, cke, ket, et#, t##\}$$

$$##rocket## \rightarrow \{##r, #ro, roc, ock, cke, ket, et#, t##\}$$

Khi đó:

$$TF(p0ck@t, pocket) = |\{##p, t##\}| = 2$$

$$TF(p0ck@t, rocket) = |\{t##\}| = 1$$

Trong trường hợp này, chúng ta chọn được từ thay thế chính xác cho “p0ck@t” là “pocket”.

Phương pháp hiệu chỉnh

Đầu tiên, chúng tôi sử dụng khoảng cách Levenshtein để chọn lọc các từ thay thế ứng viên từ một từ điển D (chứa 10000 từ tiếng Anh thông dụng theo Google[22]). Sau đó dựa vào mô hình N-gram (bậc 3) đối với từ để chọn từ thay thế hợp lý nhất. Gọi w_{ocr} là kết quả nhận dạng từ phần mềm OCR. Các bước thực hiện quá trình hiệu chỉnh như sau:

- Với mỗi từ w thuộc từ điển D , tính khoảng cách Levenshtein $L(w_{ocr}, w)$ giữa w_{ocr} và w .
- Các từ có khoảng cách $L(w_{ocr}, w)$ nhỏ nhất được chọn lựa và tạo thành tập các từ thay thế ứng viên CW .
- Với mỗi $w_{cand} = a_1 a_2 a_3 \dots a_n \in CW$, tính $TF(w_{ocr}, w_{cand})$ và $score(w_{cand})$, với $score(w_{cand})$ được định nghĩa trong công thức (3.10).

$$score(w) = \sqrt[n-2]{\prod_{i=1}^{n-2} P(a_{i+2} | a_i a_{i+1})} \quad (3.10)$$

- Từ được chọn để thay thế là từ có tổng $score(w_{cand}) + TF(w_{ocr}, w_{cand})$ lớn nhất: $w_o = \arg \max_{w_{cand} \in CW} (score(w_{cand}) + TF(w_{ocr}, w_{cand}))$.

Chương 4 Mô hình truy vấn ảnh dựa vào văn bản ngoại cảnh

Trong chương này, chúng tôi trình bày mô hình truy vấn ảnh với hai vấn đề chính là tổ chức dữ liệu và cách thức truy vấn nhằm đạt được các mục tiêu đã đề ra.

4.1 Mô hình tổ chức dữ liệu



Hình 4.1 Sơ đồ tổ chức dữ liệu ảnh

Tổ chức dữ liệu ảnh gồm các giai đoạn chính sau:

- Phát hiện, rút trích và nhận dạng văn bản trong ảnh
- Gom nhóm văn bản
- Trích chọn phần tử đại diện nhóm văn bản

4.1.1 *Phát hiện, rút trích và nhận dạng văn bản*

Từ tập dữ liệu ảnh ban đầu sử dụng mô hình phát hiện và rút trích văn bản để thu được tập các vùng ảnh chứa văn bản đã nhị phân. Tập các vùng ảnh này được đưa vào phần mềm OCR Tesseract để nhận dạng văn bản và được hiệu chỉnh bằng phương pháp đã đề xuất. Kết quả sau bước này, mỗi ảnh trong tập dữ liệu ảnh ban đầu được thay thế bằng chuỗi ký tự đã rút trích và nhận dạng được. Chúng tôi đã chuyển từ tập dữ liệu ảnh ban đầu thành tập chuỗi ký tự, tập chuỗi ký tự này chứa đựng một phần nào ngữ nghĩa của tập ảnh. Tập chuỗi ký tự được tổ chức lưu trữ dưới dạng tập tin XML.

4.1.2 *Gom nhóm văn bản*

Để tổ chức dữ liệu được gọn nhẹ và hiệu quả hơn, chúng tôi tiến hành phân lớp tập các chuỗi ký tự thu được ở bước trước đó nhằm rút trích được các chuỗi có cùng sự tương đồng. Các chuỗi ký tự được gom nhóm bằng giải thuật gom nhóm phân cấp HAC. Độ đo dị biệt giữa các nhóm là khoảng cách Levenshtein giữa các chuỗi ký tự.

Giải thuật gom nhóm cây phân cấp (HAC)

Chúng tôi sử dụng giải thuật gom nhóm phân cấp HAC vì ưu điểm so với các thuật toán gom nhóm khác là: số lớp cơ sở không cần xác định trước vì vậy ta có thể phân lớp từ thô đến mịn. Cây gom nhóm phân cấp có thể được tạo hoặc từ dưới lên (bắt đầu với mỗi đối tượng là một nhóm và gom nhóm các đối tượng thành một nhóm) hoặc từ trên xuống (bắt đầu với tất cả các đối tượng thuộc cùng một nhóm, sau đó tiến hành chia thành các nhóm nhỏ hơn).

Thuật toán từ dưới lên thực hiện với các bước như sau:

- (i) Ban đầu mỗi đối tượng được phân thành một nhóm. Khởi tạo ma trận khoảng cách giữa các nhóm.
- (ii) Xác định hai đối tượng có sự tương đồng cao nhất và gom thành một nhóm mới. Cập nhật ma trận khoảng cách giữa nhóm mới tạo và các nhóm cũ.
- (iii) Lặp lại cho đến khi chỉ còn một nhóm lớn chứa tất cả các đối tượng.

Giải thuật gom nhóm phân cấp từ dưới lên được trình bày cụ thể trong Bảng 4.1.

Bảng 4.1 Giải thuật gom nhóm phân cấp từ dưới lên

Giải thuật gom nhóm phân cấp từ dưới lên

Cho tập đối tượng $X = \{x_1 \dots x_N\}$

Giai đoạn khởi tạo

Bước 1: Tạo phân cấp $R_0 = \{x_i, i = 1, \dots, N\}$, với các lớp $C_i = \{x_i\}$

Bước 2: Đặt $P_0 = P(X)$, trong đó:

$P(X)$ là ma trận kích thước $N \times N$, với $P(i, j) = d(x_i, x_j)$

d là độ đo khác biệt giữa 2 đối tượng

Bước 3: $t=0$

Giai đoạn phân lớp

Bước 1: $t=t+1$

Bước 2: Chọn cặp C_i, C_j sao cho:

$$d(C_i, C_j) = \min d(C_r, C_s), r, s = 1, \dots, N, r \neq s$$

Bước 3: Kết nạp C_i, C_j vào lớp C_q để tạo phân cấp R_t

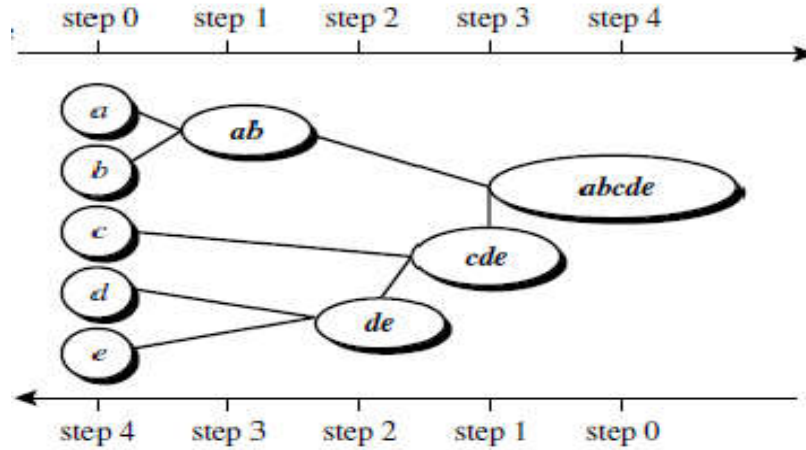
$$R_t = (R_{t-1} - \{C_i, C_j\}) \cup \{C_q\}$$

Bước 4: Cập nhật P_t từ P_{t-1} bằng 2 bước:

- Xóa các dòng và cột tương ứng với hai lớp vừa kết nạp

- Thêm dòng mới và cột mới chứa khoảng cách giữa lớp mới tạo với các lớp cũ

Bước 5: Lặp lại bước 1 cho đến khi nào các x_i chưa thuộc cùng một lớp

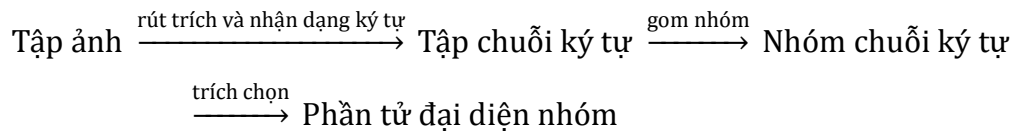


Hình 4.2 Minh họa các bước gom nhóm bằng thuật toán HAC

4.1.3 Trích chọn phần tử đại diện nhóm văn bản

Trong mỗi nhóm chuỗi ký tự đã gom nhóm ở bước 2, trích chọn phần tử đại diện cho nhóm để thuận tiện cho việc truy vấn. Phần tử đại diện nhóm là chuỗi ký tự có khoảng cách nhỏ nhất đến tất cả các chuỗi ký tự còn lại trong cùng nhóm.

Sơ đồ tổ chức dữ liệu ảnh như sau:



Tập dữ liệu ảnh ban đầu $SI = \{I_i, i \in [1..n_{SI}]\}$.

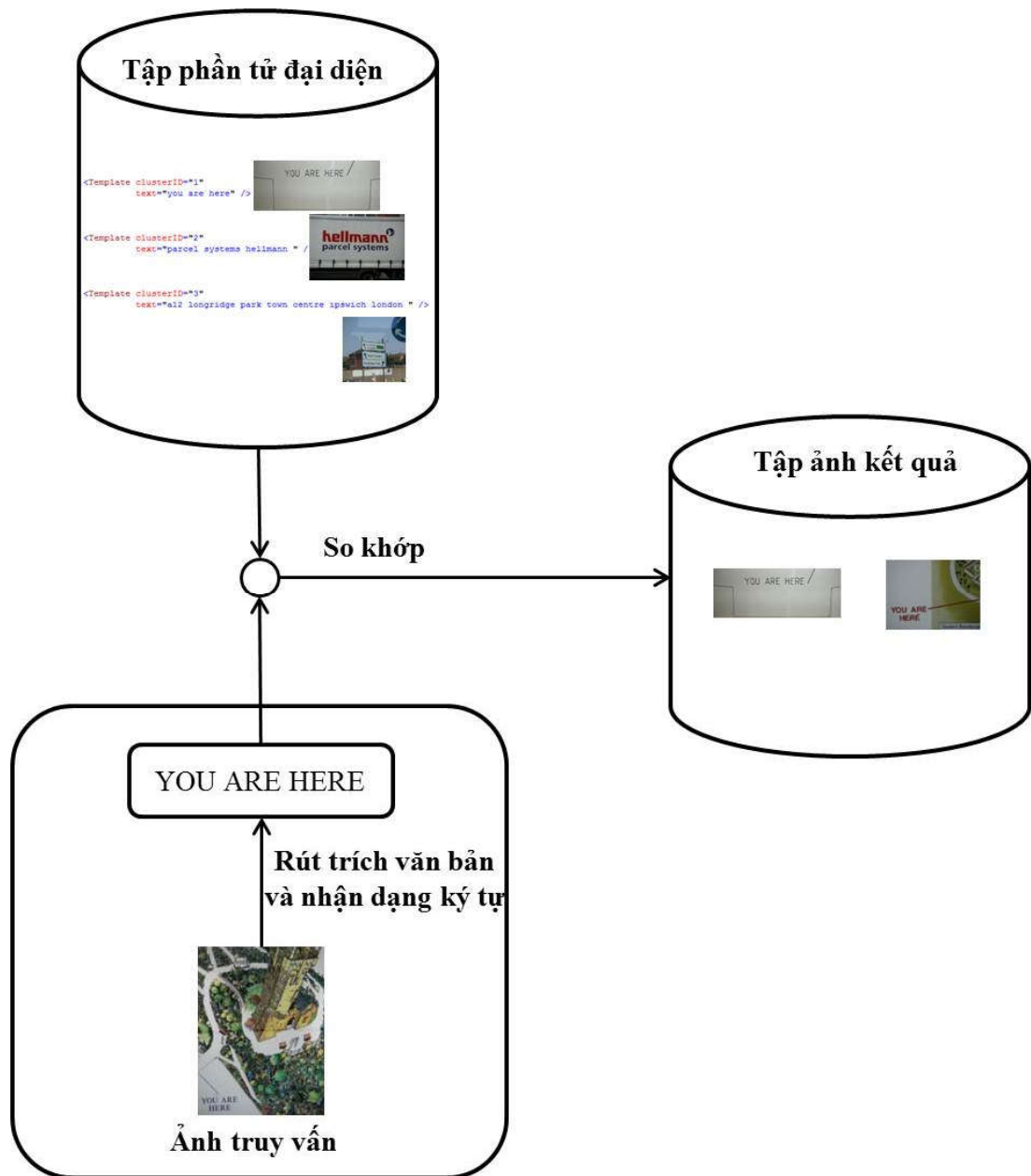
Từ tập SI , qua rút trích và nhận dạng ký tự, ta có tập $SS = \{s_i, i \in [1..n_{SI}]\}$.

Từ tập SS , qua gom nhóm chuỗi ký tự, ta có tập $SC = \{c_k, k = 1..M\}$.

Trong mỗi nhóm $c_k = \{s_{kl}, k \in [1..M], l \in [1..n_{s_k}]\}$, chọn phần tử đại diện t_k thỏa $\sum_{l=1}^{n_{s_k}} d(t_k, s_{kl}) = \min_{l=1..n_{s_k}} \sum_{j=1}^{n_{s_k}} d(s_{kl}, s_{kj})$.

Từ tập SC , trích chọn phần tử đại diện, ta có tập $ST = \{t_k, k = 1..M\}$.

4.2 Mô hình truy vấn ảnh dựa vào văn bản ngoại cảnh



Hình 4.3 Sơ đồ truy vấn ảnh

Trong mô hình truy vấn, chúng tôi xem xét ba yếu tố quan trọng sau: cách thức nhập liệu, chiến lược truy vấn và hiệu quả truy vấn.

🌈 Cách thức nhập liệu

Trong mô hình này, chúng tôi hỗ trợ hai cách thức truy vấn: bằng từ khóa và bằng ảnh.

- Truy vấn bằng từ khóa:

Câu truy vấn có dạng $Q = w_1 w_2 \dots w_n$, trong đó w_1, w_2, \dots, w_n là các từ khóa.

- Truy vấn bằng ảnh hỗ trợ trong trường hợp vì lý do nào đó người dùng không thể nhập được từ khóa cần truy vấn (ví dụ người dùng không biết ngôn ngữ của từ khóa):

Cho trước ảnh truy vấn QI , ảnh truy vấn được rút trích các vùng văn bản và nhận dạng ký tự để thu được chuỗi truy vấn $Q = w_1 w_2 \dots w_n$.

Chiến lược truy vấn

Định nghĩa độ đo dị biệt: khoảng cách giữa chuỗi ký tự $Q = w_1 w_2 \dots w_n$ và chuỗi ký tự $s = v_1 v_2 \dots v_m$ sử dụng để so khớp trong phần truy vấn được định nghĩa trong công thức (4.1).

$$d(Q, s) = \frac{\sum_{i=1}^n \text{dmin}(w_i, s)}{n} \quad (4.1)$$

Trong đó, $\text{dmin}(w_i, s) = \min \{L(w_i, v_j) \mid v_j \in s\}$ và $L(w, v)$ là khoảng cách Levenshtein giữa hai từ.

Chúng tôi hỗ trợ hai chế độ tìm kiếm: tìm chính xác và tìm gần đúng. Đối với chế độ tìm chính xác, một ảnh được xem là tìm được đúng nếu ảnh đó chứa ít nhất chính xác một từ thuộc chuỗi truy vấn $w_i \in Q$. Khi đó, độ dị biệt $\xi = 0$. Trong chế độ tìm gần đúng, một ảnh được xem là tìm được đúng nếu độ dị biệt theo công thức (4.1) giữa chuỗi truy vấn và chuỗi ký tự xuất hiện trong ảnh không vượt quá ngưỡng ξ do người dùng xác định.

Quá trình truy vấn như sau:

- So khớp chuỗi ký tự truy vấn Q với các chuỗi ký tự đại diện. Với mỗi phần tử đại diện $t_k \in ST$, tính $d(Q, t_k)$.

- Tập các phần tử đại diện được chọn thỏa mãn điều kiện sau:

$$ST_q = \left\{ t_k \mid t_k \in ST \wedge \left(IsSubstring(Q, t_k) \vee d(Q, t_k) \leq \xi \right) \right\}$$

Trong đó, $IsSubstring(Q, t_k)$ trả về giá trị true nếu t_k có chứa một từ $w_i \in Q$.

- Gọi SC_q là tập các nhóm có phần tử đại diện được chọn:

$$SC_q = \left\{ c_k \mid c_k \in SC \wedge t_k \in ST_q \right\}$$

- Với mỗi chuỗi ký tự s_{kl} trong từng nhóm $c_k \in SC_q$, tính khoảng cách với chuỗi truy vấn $d(Q, s_{kl})$. Tập chuỗi ký tự được chọn thỏa mãn điều kiện:

$$SS_q = \left\{ s_{kl} \mid s_{kl} \in c_k \wedge c_k \in SC_q \wedge \left(IsSubstring(Q, s_{kl}) \vee d(Q, s_{kl}) \leq \xi \right) \right\}.$$

- Tập ảnh kết quả là:

$$SRI_q = \left\{ I_i \mid s_i \in SS \wedge s_{kl} \in SS_q \wedge s_i = s_{kl}, i \in [1..n_{SI}] \right\}.$$

Hiệu quả truy vấn

Sắp hạng kết quả truy vấn: có khả năng sắp hạng dựa vào độ đo $d(Q, s_{kl})$.

Đánh giá hiệu quả của kết quả truy vấn

Chúng tôi dùng hai đại lượng là độ chính xác và độ phủ để đánh giá hiệu quả tìm kiếm của hệ thống.

Độ chính xác được xác định:

$$precision = \frac{N_{correct}}{N_{found}} \quad (4.2)$$

Độ phủ được xác định:

$$recall = \frac{N_{correct}}{N_{manual}} \quad (4.3)$$

Trong đó:

$N_{correct}$: số ảnh tìm được đúng với độ dị biệt ξ .

N_{found} : số ảnh tìm được với độ dị biệt ξ .

N_{manual} : số ảnh đúng thực có với độ dị biệt ξ .

Chương 5 Kết quả thực nghiệm

Trong chương này, chúng tôi trình bày các kết quả đạt được từ mô hình phát hiện và rút trích văn bản, hệ thống truy vấn ảnh dựa vào văn bản ngoại cảnh đã đề xuất.

5.1 Kết quả phát hiện và rút trích văn bản

5.1.1 Tập dữ liệu thử nghiệm và phương pháp đánh giá

Mô hình phát hiện và rút trích văn bản ngoại cảnh được đánh giá trên tập dữ liệu ICDAR 2003[8]. Đây là tập dữ liệu được sử dụng các trong cuộc thi phát hiện và định vị văn bản trong ảnh năm 2003, 2005 và 2011. Tập dữ liệu này bao gồm 251 ảnh thuộc tập TrialTrain và 249 ảnh thuộc tập TrialTest được chụp cả trong nhà và ngoài trời, kích thước ảnh thay đổi từ 307x93 pixels đến 1600x1200 pixels. Các ảnh trong tập TrialTrain được sử dụng để huấn luyện bộ phân lớp SVM. Hiệu quả của phương pháp phát hiện văn bản ngoại cảnh trong ảnh được đánh giá trên tập ảnh TrialTest.

Chúng tôi đánh giá hiệu quả của hệ thống phát hiện văn bản với độ chính xác (*precision*) và độ phủ (*recall*) theo chuẩn ICDAR 2003[12] được mô tả như sau. Đầu ra của mỗi thuật toán là tập các hình chữ nhật bao quanh các từ phát hiện được.

Gọi E là tập các hình chữ nhật phát hiện được và T là tập các hình chữ nhật thực (groundtruth).

Độ khớp $m_p(r_1, r_2)$ giữa hai hình chữ nhật r_1 và r_2 được định nghĩa là tỉ số giữa r_2 diện tích vùng giao nhau của r_1 và r_2 với diện tích của vùng nhỏ nhất chứa cả r_1 và r_2 được thể hiện trong công thức (5.1), với $a(r)$ là diện tích của vùng r . Độ khớp giữa hai hình chữ nhật sẽ có giá trị 1 nếu chúng trùng nhau, ngược lại độ khớp có giá trị 0 nếu hai hình chữ nhật không giao nhau.

$$m_p(r_1, r_2) = \frac{2a(r_1 \cap r_2)}{a(r_1) + a(r_2)} \quad (5.1)$$

Độ khớp tốt nhất của một hình chữ nhật r đối với tập các hình chữ nhật R được định nghĩa như sau:

$$m(r, R) = \max\{m_p(r, r') \mid r' \in R\} \quad (5.2)$$

Khi đó, độ chính xác (Precision) và độ phủ (Recall) được xác định như sau:

$$\text{Precision} = \frac{\sum_{r_e \in E} m(r_e, T)}{|E|} \quad (5.3)$$

$$\text{Recall} = \frac{\sum_{r_t \in T} m(r_t, E)}{|T|} \quad (5.4)$$

Độ đo chuẩn f được sử dụng để kết hợp độ chính xác và độ phủ thành một độ đo duy nhất định nghĩa như sau: $f = \frac{1}{\alpha / P + (1 - \alpha) / R}$. Để cân bằng trọng số giữa độ chính xác và độ phủ, chúng tôi chọn $\alpha = 0.5$.

5.1.2 Kết quả thực nghiệm

Hiệu quả của hệ thống phát hiện văn bản ngoại cảnh đề xuất trên tập dữ liệu học TrialTrain gồm 251 ảnh được trình bày trong Bảng 5.1.

Bảng 5.1 Hiệu quả phát hiện văn bản trong tập dữ liệu học của phương pháp đề xuất

Phương pháp	Precision	Recall	f
Phương pháp đề xuất	0.81	0.63	0.71

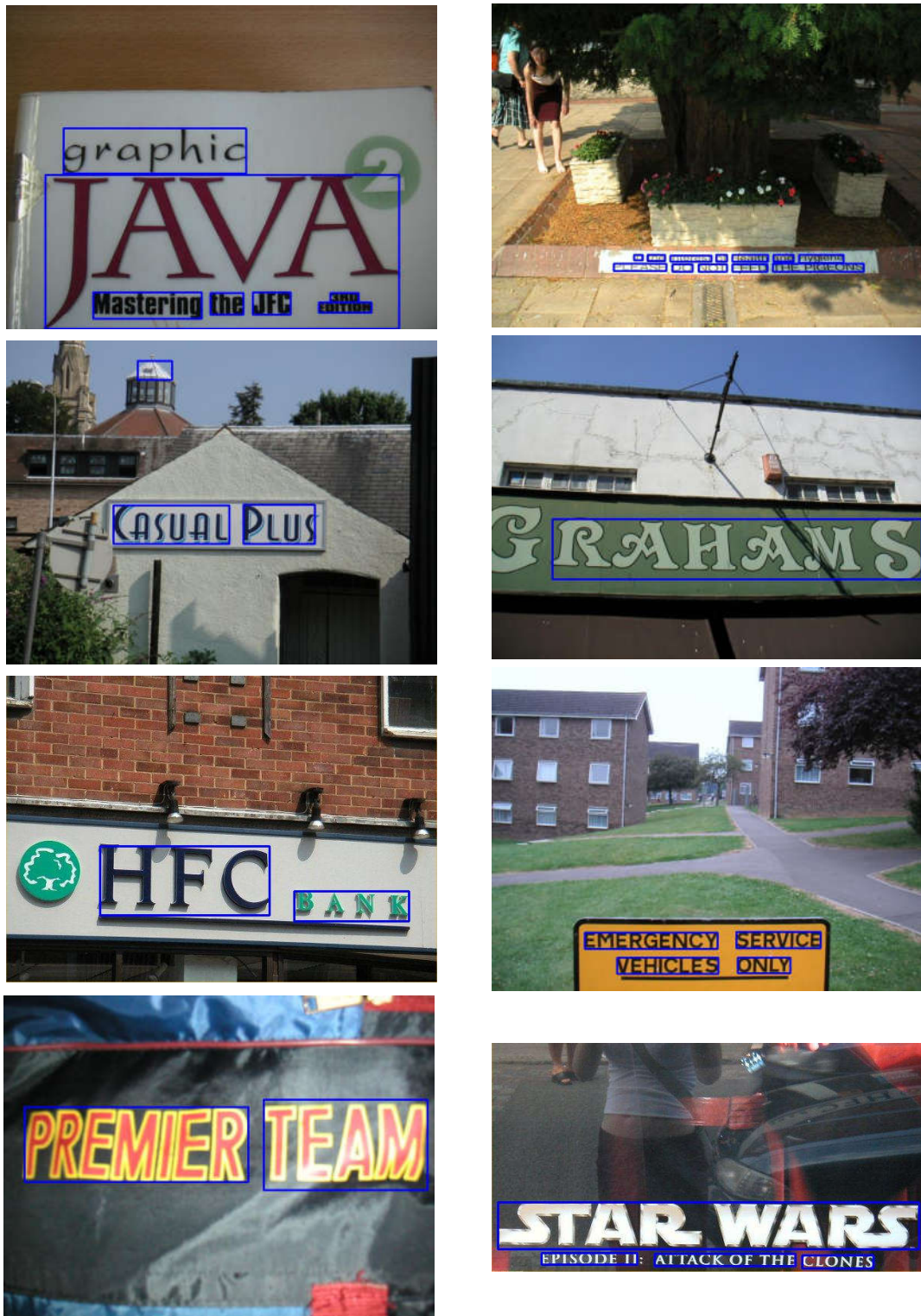
Hiệu quả trên tập dữ liệu thử nghiệm TrialTest gồm 249 ảnh của phương pháp đề xuất và các phương pháp khác trong ICDAR 2003 [12], ICDAR 2005 [13], và ICDAR 2011 [15] được trình bày trong Bảng 5.2.

Bảng 5.2 Hiệu quả của các phương pháp phát hiện văn bản khác nhau

Phương pháp	Precision	Recall	f
Kim's method[15]	0.83	0.62	0.71

Phương pháp đề xuất	0.78	0.62	0.69
Epshtein [5]	0.73	0.60	0.66
Yi [15]	0.67	0.58	0.62
TH-TextLoc [15]	0.67	0.58	0.62
Hinnerk Becker[13]	0.62	0.67	0.62
Neumann [15]	0.69	0.53	0.60
Alex Chen[13]	0.60	0.60	0.58
Ashida[12]	0.55	0.46	0.50
HWDavid[12]	0.44	0.46	0.45

Phương pháp đề xuất đạt độ phủ tương tự với phương pháp của Kim – phương pháp tốt nhất tại ICDAR 2011 với 62%, độ chính xác của phương pháp đề xuất không cao bằng phương pháp của Kim với 78%. Hệ thống của Kim chỉ được công bố về mặt hiệu quả, nhưng phương pháp đến nay vẫn chưa được công bố. So sánh với phương pháp của Epshtein[5] cũng như các phương pháp còn lại trong Bảng 5.2, phương pháp đề xuất đã có những cải tiến đáng kể về hiệu quả. Giai đoạn tiền xử lý đã cho thấy hiệu quả trong việc loại bỏ các đối tượng nhiễu đồng thời làm nổi bật các vùng văn bản trong ảnh. Kết quả thực nghiệm cũng chứng tỏ đặc trưng HOG có khả năng đặc tả tốt cấu trúc và các đặc trưng của vùng văn bản. Một số kết quả phát hiện văn bản ngoại cảnh trong trên tập dữ liệu ICDAR được minh họa trong Hình 5.1. Mô hình phát hiện văn bản đề xuất gặp thất bại trong các trường hợp như: (a) kích thước văn bản quá nhỏ (bé hơn 8); (b) ánh sáng quá mạnh, (c) chiều dài chuỗi ký tự nhỏ hơn 3; (d) màu văn bản giống màu nền; (e) màu chữ trong suốt; (f) bị che khuất bởi các đối tượng khác (như lưới) (xem Hình 5.2).



Hình 5.1 Minh họa một số kết quả phát hiện văn bản ngoại cảnh trong ảnh



(a)



(b)



(c)



(d)



(e)



(f)

Hình 5.2 Minh họa một số trường hợp thất bại

5.2 Đánh giá hiệu quả phương pháp hiệu chỉnh kết quả nhận dạng bằng phần mềm OCR

Để nhận dạng ký tự trong vùng ảnh văn bản đã được nhị phân, chúng tôi sử dụng phần mềm mở Tesseract của Google. Kết quả này sau đó sẽ được hiệu chỉnh bằng phương pháp đề xuất. Tập ngữ liệu được dùng để tính các xác suất trong mô hình N-gram được lấy từ groundtruth (tập các từ thực có trong tập ảnh). Kết quả nhận dạng văn bản được đánh giá với độ chính xác và độ phủ được định nghĩa như sau:

$$\text{Độ chính xác} = \frac{\text{Số từ nhận dạng đúng}}{\text{Số từ nhận dạng được}} \quad (5.5)$$

$$\text{Độ phủ} = \frac{\text{Số từ nhận dạng đúng}}{\text{Tổng số từ thực có}} \quad (5.6)$$

Các thông số sử dụng trong công thức (5.5) và (5.6) được thống kê dựa trên các từ phát hiện chính xác của toàn bộ tập dữ liệu thử nghiệm bao gồm tập TrialTrain và TrialTest. Hiệu quả nhận dạng ký tự trước và sau khi áp dụng phương pháp hiệu chỉnh đề xuất được trình bày trong Bảng 5.3. Bảng 5.4 trình bày một số kết quả nhận dạng văn bản trước và sau khi hiệu chỉnh.


Bảng 5.3 Hiệu quả nhận dạng văn bản trước và sau khi hiệu chỉnh

	OCR	OCR + hiệu chỉnh
Số từ nhận dạng đúng	1036	1148
Số từ nhận dạng được	1674	1674
Tổng số từ thực có	1701	1701
Độ chính xác	61.89%	68.58%
Độ phủ	60.91%	67.49%

Kết quả thực nghiệm đã chứng tỏ hiệu quả của phương pháp nhị phân vùng ảnh văn bản và phương pháp hiệu chỉnh kết quả nhận dạng đề xuất. Phương pháp

hiệu chỉnh kết quả nhận dạng đề xuất làm tăng đáng kể hiệu quả nhận dạng từ phần mềm OCR, từ đó góp phần vượt qua một phần các thách thức khi nhận dạng văn bản ngoại cảnh trên ảnh tự nhiên.

Bảng 5.4 Một số kết quả nhận dạng văn bản trước và sau khi hiệu chỉnh

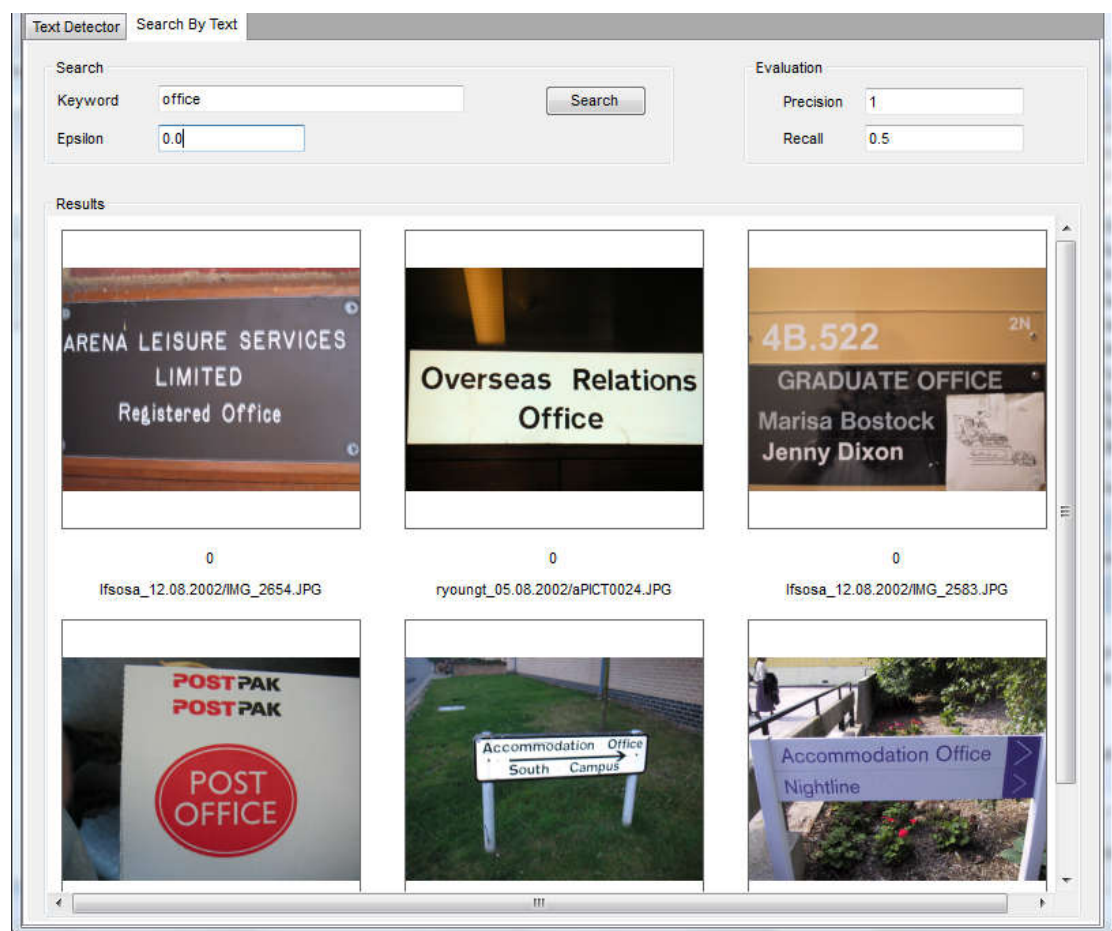
Ảnh	Kết quả nhận dạng không hiệu chỉnh	Kết quả nhận dạng sau khi hiệu chỉnh
	SNOUT7	SNOUT
	INAL iearance EURTDN	FINAL Clearance BURTON
	SPJIDLERMAN	SPIDER MAN
	.rf.cks FAMOUS SUPPLIES EST 1946	JACKS FAMOUS SUPPLIES EST 1946

5.3 Kết quả truy vấn ảnh

Chúng tôi đã thử nghiệm hiệu quả của hệ thống truy vấn ảnh dựa vào văn bản ngoại cảnh trên cùng tập dữ liệu ICDAR gồm tổng cộng 500 ảnh.

5.3.1 Kết quả truy vấn ảnh bằng từ khóa

Để đánh giá hiệu quả truy vấn bằng từ khóa, chúng tôi sử dụng 50 từ xuất hiện trong tập dữ liệu ảnh làm các từ khóa truy vấn. Hình 5.3 và Hình 5.4 minh họa một số kết quả truy vấn ảnh bằng cách dùng từ khóa với độ dị biệt $\xi = 0.0$.











Hình 5.3 Kết quả truy vấn ảnh dùng từ khóa “office”

Text Detector Search By Text

Search
Keyword: Search
Epsilon:

Evaluation
Precision:
Recall:

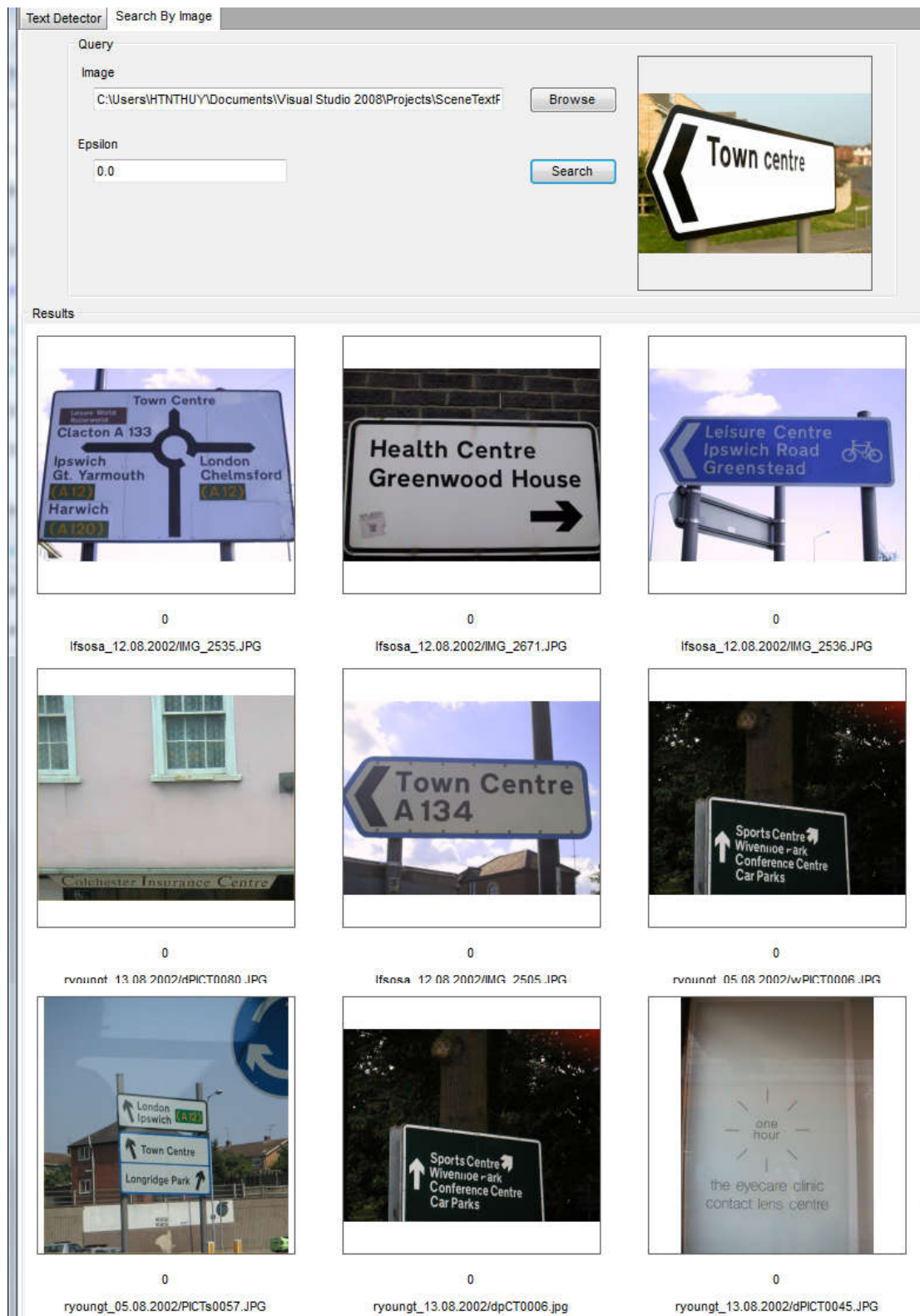
Results

		
0 ryoungt_05.08.2002/wPCT0006.JPG	0 ryoungt_05.08.2002/Pict0021.jpg	0 ryoungt_05.08.2002/PICTs0057.JPG
		
0 ryoungt_13.08.2002/dpCT0006.jpg	0 Ifsosa_12.08.2002/IMG_2471.JPG	0 ryoungt_05.08.2002/qPCT0013.JPG
		
0 Ifsosa_12.08.2002/IMG_2510.JPG	0 ryoungt_13.08.2002/fdPCT0011.JPG	

Hình 5.4 Kết quả truy vấn ảnh dùng từ khóa (“car park”)

5.3.2 Kết quả truy vấn ảnh bằng ảnh chứa văn bản tự nhiên

Để đánh giá hiệu quả truy vấn bằng ảnh, chúng tôi sử dụng 25 ảnh không thuộc tập dữ liệu làm các ảnh truy vấn. Hình 5.5 và Hình 5.6 minh họa các kết quả truy vấn ảnh từ một ảnh truy vấn có chứa các từ khóa mong muốn với độ dị biệt $\xi = 0.0$.



Hình 5.5 Kết quả truy vấn bằng ảnh

Text Detector Search By Image










Query

Image
C:\Users\HTNTHUY\Documents\Visual Studio 2008\Projects\SceneTextF Browse

Epsilon
0.0 Search

University of Essex

Results

		
0 ryoungt_13.08.2002/dPICT0093.JPG	0 ryoungt_05.08.2002/qPICT0010.JPG	0 ryoungt_05.08.2002/qPICT0013.JPG
		
0 ifsosa_12.08.2002/IMG_2591.JPG	0 ifsosa_12.08.2002/IMG_2498.JPG	0 ryoungt_05.08.2002/PICT0016.JPG
		
0 ryoungt_05.08.2002/aPICT0034.JPG	0 ryoungt_13.08.2002/dPICT0030.JPG	0 ifsosa_12.08.2002/IMG_2469.JPG

Hình 5.6 Kết quả truy vấn bằng ảnh

Hiệu quả của hệ thống được đánh giá với độ chính xác và độ phủ được định nghĩa trong các công thức (4.2) và (4.3). Hiệu quả của cả hai cách thức truy vấn với độ dị biệt $\xi = 0.0$ được trình bày trong Bảng 5.5.

Bảng 5.5 Hiệu quả truy vấn ảnh với độ dị biệt $\xi = 0.0$

	Truy vấn bằng từ khóa	Truy vấn bằng ảnh
Số lần truy vấn	50	25
Tổng số ảnh tìm được đúng	160	249
Tổng số ảnh tìm được	163	273
Tổng số ảnh đúng thực có	215	412
Độ chính xác	98.36%	91.20%
Độ phủ	74.41%	60.43%

Hiệu quả của hệ thống truy vấn ảnh dựa vào văn bản ngoại cảnh phụ thuộc nhiều vào hiệu quả của mô hình phát hiện và nhận dạng văn bản ngoại cảnh trong ảnh. Hiệu quả truy vấn bằng ảnh có giảm so với truy vấn bằng từ khóa do ảnh hưởng của phương pháp phát hiện và nhận dạng văn bản ngoại cảnh. Tuy nhiên, truy vấn bằng ảnh vẫn cần thiết để hỗ trợ người dùng trong các trường hợp đặc biệt đã nêu. Thực nghiệm cho thấy hệ thống truy vấn ảnh dựa vào văn bản ngoại cảnh bước đầu đạt được kết quả tương đối khả quan và có nhiều triển vọng trong tương lai. Độ chính xác của hệ thống đạt được khá cao vì quá trình truy vấn chủ yếu dựa vào việc so khớp giữa từ khóa truy vấn và các từ khóa đã rút trích được từ tập dữ liệu ảnh. Mô hình truy vấn ảnh dựa vào văn bản ngoại cảnh chưa được đề xuất và công bố bởi bất kỳ nghiên cứu nào trước đây, nên chúng tôi không thể so sánh hiệu quả của mô hình đề xuất với các mô hình khác.

Chương 6 Kết luận và hướng phát triển

6.1 Kết luận

Trong luận văn này, chúng tôi đã trình bày về việc xây dựng hệ thống truy vấn ảnh dựa vào văn bản ngoại cảnh bao gồm hai mô hình chính là mô hình phát hiện và rút trích văn bản ngoại cảnh trong ảnh và mô hình truy vấn ảnh.

Trong mô hình phát hiện và rút trích văn bản ngoại cảnh trong ảnh, chúng tôi đề xuất phương pháp theo mô hình định vị/tinh lọc văn bản. Phương pháp đề xuất góp phần vượt qua một số thách thức như độ phân giải thấp, nền nhiễu loạn, không biết trước về màu sắc, font chữ, cỡ chữ, bố cục và vị trí của văn bản trong ảnh. Trong đó, chúng tôi đã sử dụng phép *reconstruction* cho giai đoạn tiền xử lý để loại bỏ các đối tượng nền xung quanh văn bản. Chúng tôi cũng sử dụng các toán tử hình thái học để phát sinh các vùng văn bản ứng viên. Các từ ứng viên được hình thành dựa vào SWT. Một bộ phân lớp SVM sử dụng đặc trưng HOG được huấn luyện để phân lớp các từ ứng viên. Đối với giai đoạn rút trích văn bản, chúng tôi đã đề xuất phương pháp nhị phân hóa các vùng ảnh chứa văn bản giúp làm tăng đáng kể hiệu quả của giai đoạn nhận dạng văn bản bằng phần mềm OCR. Đối với giai đoạn nhận dạng văn bản, vì nội dung này không nằm trong phạm vi nghiên cứu của luận văn nên chúng tôi chỉ trình bày phương pháp hiệu chỉnh kết quả OCR nhằm nâng cao hiệu quả nhận dạng văn bản ngoại cảnh bằng phần mềm. Kết quả thực nghiệm cho thấy phương pháp phát hiện văn bản đề xuất có những cải tiến so với các phương pháp hiện tại và có nhiều triển vọng trong tương lai.

Trong mô hình truy vấn ảnh, chúng tôi đã xây dựng mô hình tổ chức dữ liệu và cách thức truy vấn ảnh. Từ tập dữ liệu ảnh ban đầu, chúng tôi thay thế mỗi ảnh bằng chuỗi ký tự rút trích và nhận dạng được từ ảnh. Từ đó, chúng tôi gom nhóm văn bản và rút trích các phần tử đại diện dựa vào giải thuật gom nhóm phân cấp HAC. Luận văn đã đề xuất và thử nghiệm hệ thống truy vấn ảnh nhằm phục vụ cho nhu cầu tìm kiếm các ảnh có chứa các từ khóa mong muốn. Đây là mô hình truy vấn mới góp phần vượt qua một phần vấn đề về lỗ hổng ngữ nghĩa giữa dữ liệu lưu trữ

và thông tin truy vấn. Đồng thời, với việc cho phép người dùng truy vấn bằng cách sử dụng ảnh có chứa văn bản ngoại cảnh, mô hình cũng đã góp phần vượt qua các trở ngại trong trường hợp người dùng không biết (hoặc không thể nhập) ngôn ngữ của từ khóa cần truy vấn. Thực nghiệm chứng tỏ hệ thống truy vấn ảnh của chúng tôi bước đầu đạt được kết quả tương đối khả quan.

6.2 Hướng phát triển

Đối với mô hình phát hiện văn bản ngoại cảnh, để nâng cao hiệu quả hơn nữa, chúng tôi sẽ tìm kiếm và kết hợp với các đặc trưng khác có khả năng phân biệt tốt hơn văn bản và các đối tượng khác. Ngoài ra, để tạo thành một mô hình phát hiện và nhận dạng văn bản ngoại cảnh trong ảnh một cách hoàn thiện, chúng tôi sẽ nghiên cứu và tích hợp các phương pháp nhận dạng văn bản ngoại cảnh vào hệ thống thay vì sử dụng các phần mềm OCR như hiện nay.

Hệ thống truy vấn ảnh dựa vào văn bản ngoại cảnh hiện tại chỉ được thử nghiệm trên tập dữ liệu với một số lượng ảnh tương đối nhỏ. Trong tương lai, chúng tôi sẽ thử nghiệm hiệu quả của hệ thống truy vấn ảnh trên một tập ảnh với số lượng lớn. Bên cạnh đó, việc kết hợp mô hình truy vấn ảnh dựa vào văn bản ngoại cảnh vào các hệ thống truy vấn ảnh dựa vào đặc trưng thị giác và ngữ nghĩa hiện có cũng là một hướng nghiên cứu triển vọng. Đồng thời, cùng với việc phát triển của các thiết bị di động, trong tương lai những kết quả từ luận văn có thể áp dụng để xây dựng một hệ thống truy vấn thông tin trên các thiết bị di động. Ngoài ra, một hướng phát triển rất có ý nghĩa là mở rộng mô hình hiện nay để có thể xử lý trên chữ Việt.

Truy vấn dữ liệu ảnh luôn là một bài toán quan trọng và có ý nghĩa thiết thực trong cuộc sống. Tất cả các vấn đề mở này là những hướng phát triển đáng suy nghĩ nhằm xây dựng một hệ thống truy vấn ảnh hoàn thiện và hiệu quả trong tương lai.

Tài liệu tham khảo

- [1] D. Chen, J. M. Odobez, and H. Bourlard, Text detection and recognition in images and video frames, *Pattern Recognition*, 37(3): 595-608, 2004.
- [2] X. Chen and A. L. Yuille, Detecting and reading text in natural scenes, in *CVPR*, 2004, Vol.2, pp.II-366–II-373.
- [3] N. Dalal, *Finding People in Images and Videos*, 2006.
- [4] R. Datta, D. Joshi, J. Li, and J. Z. Wang, Image retrieval: Ideas, influences, and trends of the new age, *ACM Computing Surveys*, 40(2), 2008.
- [5] B. Epshtein, E. Ofek, and Y. Wexler, Detecting text in natural scenes with stroke width transform, *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2963-2970, 2010.
- [6] N. Ezaki, M. Bulacu, L. Schomaker, Text detection from natural scene images: Towards a system for visually impaired persons, *International Conference on Pattern Recognition*, 2004, pp. 683–686.
- [7] S. M. Hanif, and L. Prevost, Text detection and localization in complex scene images using constrained adaboost algorithm, *ICDAR*, 2009.
- [8] ICDAR 2003 Robust reading and text locating competition image database
<http://algoval.essex.ac.uk/icdar/Datasets.html>
- [9] K. Jung, K. Kim, A. K. Jain, Text information extraction in images and video: A survey, *Pattern Recognition*, 37(5): 977 – 997, 2004.
- [10] K. I. Kim, K. Jung and J. H. Kim, Texture-Based Approach for Text Detection in Images using Support Vector Machines and Continuous Adaptive Mean Shift Algorithm, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 25, No. 12, December 2003.
- [11] Z. Liu and S. Sarkar, Robust outdoor text detection using text intensity and shape features, *ICPR*, 2008.
- [12] S. M. Lucas et al., ICDAR2003 robust reading competitions: entries, results and future directions, In *IJDAR*, Vol. 7, pp.105 – 122, 2005.

- [13] S. M. Lucas, Text Locating Competition Results, ICDAR, Eighth International Conference on Document Analysis and Recognition (ICDAR'05), pp. 80-85, 2005.
- [14] K. Robinson and P. F. Whelan, Efficient Morphological Reconstruction: A Downhill Filter, Pattern Recognition Letters, Volume 25, Issue 15, November 2004, Pages 1759–1767.
- [15] A. Shahab, F. Shafait, and A. Dengel, ICDAR 2011 robust reading competition challenge 2: Reading text in scene images In ICDAR 2011, pp. 1491–1496, 2011.
- [16] P. Soille, Morphological Image Analysis: Principles and Applications, Springer, 2003, pp. 182–198 .
- [17] X. Tong, D. A. Evans, A Statistical Approach to Automatic OCR Error Correction in Context, In Proceedings of the Four Workshop on Very Large Corpora, 1996, pp. 88 – 100.
- [18] V. N. Vapnik, The Nature of Statistical Learning Theory, Springer, 1995.
- [19] L. Vincent, Morphological grayscale reconstruction in image analysis: applications and efficient algorithms, IEEE Transactions on Image Processing, Vol. 2, No. 2, pp. 176-201, April 1993.
- [20] Q. Ye, W. Gao, W. Wang, and W. Zeng, A robust text detection algorithm in images and video frames, Joint Conference of Fourth International Conference on Information Communications and Signal Processing and Pacific-Rim Conference on Multimedia, Singapore 2003.
- [21] Q. Ye, Q. Huang, W. Gao, and D. Zhao, Fast and robust text detection in images and video frames, Image Vision Comput., vol.23, pp. 565–576, 2005.
- [22] <https://github.com/eurekaoverdrive/google-10000-english>