

BỘ GIÁO DỤC VÀ ĐÀO TẠO
TRƯỜNG ĐẠI HỌC CẦN THƠ
TRƯỜNG CÔNG NGHỆ THÔNG TIN & TRUYỀN THÔNG



**LUẬN VĂN TỐT NGHIỆP ĐẠI HỌC
NGÀNH CÔNG NGHỆ THÔNG TIN**

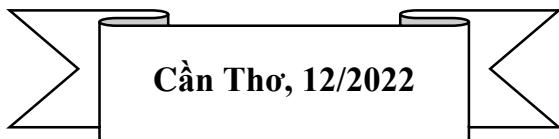
Đề tài

**ỨNG DỤNG OPTICAL CHARACTER RECOGNITION
VÀO HỆ THỐNG QUẢN LÝ CÔNG VĂN**

Sinh viên: Trần Văn Hòa

Mã số: B1809127

Khóa: K44



BỘ GIÁO DỤC VÀ ĐÀO TẠO
TRƯỜNG ĐẠI HỌC CẦN THƠ
TRƯỜNG CÔNG NGHỆ THÔNG TIN & TRUYỀN THÔNG
KHOA TRUYỀN THÔNG ĐA PHƯƠNG TIỆN



LUẬN VĂN TỐT NGHIỆP ĐẠI HỌC
NGÀNH CÔNG NGHỆ THÔNG TIN
CHUYÊN NGÀNH TIN HỌC ÚNG DỤNG

Đề tài
**ÚNG DỤNG OPTICAL CHARACTER RECOGNITION
VÀO HỆ THỐNG QUẢN LÝ CÔNG VĂN**

Người hướng dẫn

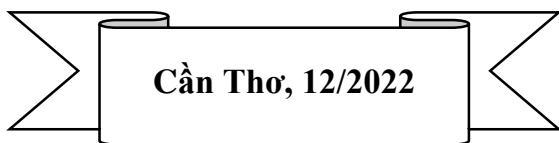
Ths Võ Hải Đăng

Sinh viên thực hiện

Trần Văn Hòa

Mã số: B1809127

Khóa: K44



Ứng dụng optical character recognition vào hệ thống quản lý công văn

NHẬN XÉT CỦA CÁN BỘ HUỐNG DẪN

୪୫

Cần Thơ, ngày tháng năm 2022

Cán bộ hướng dẫn

Ths Võ Hải Đăng

Ứng dụng optical character recognition vào hệ thống quản lý công văn

NHẬN XÉT CỦA CÁN BỘ PHẢN BIỆN

୪୫

Cần Thơ, ngày tháng năm 2022

Cán bộ phản biện

Ứng dụng optical character recognition vào hệ thống quản lý công văn

LỜI CẢM ƠN

Trong suốt quá trình học tập và thực hiện luân văn tốt nghiệp em luôn được sự quan tâm, hướng dẫn và giúp đỡ tận tình của các thầy, cô giáo trong Trường Công nghệ thông tin & Truyền thông cùng với sự động viên giúp đỡ của bạn bè.

Lời đầu tiên em xin được bày tỏ lòng biết ơn sâu sắc đến thầy, cô Trường Đại học Cần Thơ nói chung và thầy cô Trường Công nghệ thông tin & Truyền thông nói riêng đã tận tình giúp đỡ cho em suốt thời gian học tập tại trường.

Đặc biệt em xin bày tỏ lòng biết ơn chân thành sâu sắc tới thầy **Võ Hải Đăng** đã trực tiếp giúp đỡ, hướng dẫn em hoàn thành luận văn này.

Em cũng xin bày tỏ lòng biết ơn sâu sắc đến gia đình, người thân và bạn bè đã giúp đỡ động viên em hoàn thiện luận văn tốt nghiệp này.

Em xin trân trọng cảm ơn!

Cần Thơ, ngày 04 tháng 12 năm 2022

Sinh viên

Trần Văn Hòa

Ứng dụng optical character recognition vào hệ thống quản lý công văn

MỤC LỤC

MỤC LỤC.....	i
DANH MỤC HÌNH ẢNH	v
DANH MỤC BẢNG.....	viii
DANH MỤC BIỂU MÃU	ix
DANH MỤC SƠ ĐỒ	x
DANH MỤC KÝ HIỆU CHỮ VIẾT TẮT	xii
TÓM TẮT	xiii
ABSTRACT	xiv
PHẦN GIỚI THIỆU	1
I. ĐẶT VÂN ĐỀ	2
II. LỊCH SỬ GIẢI QUYẾT VÂN ĐỀ	2
III. MỤC TIÊU ĐỀ TÀI.....	3
IV. ĐỐI TƯỢNG VÀ PHẠM VI NGHIÊN CỨU.....	3
V. NỘI DUNG NGHIÊN CỨU	4
VI. BỐ CỤC CỦA QUYỀN LUẬN VĂN	4
PHẦN NỘI DUNG	5
CHƯƠNG 1 MÔ TẢ BÀI TOÁN.....	6
1.1. MÔ TẢ CHI TIẾT BÀI TOÁN	6
1.2. TIẾP CẬN VÂN ĐỀ VÀ LỰA CHỌN GIẢI PHÁP	7
1.2.1. Nhận dạng ký tự quan học (Optical Character Recognition).....	7
1.2.1.1. Giới thiệu.....	7
1.2.1.2. Lịch sử của OCR.....	8
1.2.1.3. Phương pháp nhận dạng ký tự.....	8
1.2.2. Tesseract [4], [5]	11
1.2.3. Giải thuật cây quyết định (Decision tree)	12
1.2.3.1. Giải thuật xây dựng cây quyết định [6].....	12
1.2.3.2. Nghi thức kiểm tra giải thuật [11].....	13

Ứng dụng optical character recognition vào hệ thống quản lý công văn

1.2.3.3. Đo độ hiệu quả của giải thuật [11].....	13
1.2.4. Văn bản hành chính [12]	14
1.2.5. Quy trình quản lý công văn.....	16
1.2.6. Các công nghệ sử dụng	19
CHƯƠNG 2 THIẾT KẾ VÀ CÀI ĐẶT THUẬT TOÁN.....	22
2.1. PHƯƠNG PHÁP THỰC HIỆN.....	22
2.1.1. Xử lý và trích xuất các thành phần từ ảnh văn bản.....	22
2.1.1.1. Chuẩn hóa kích thước ảnh.....	22
2.1.1.2. Chuyển đổi thành ảnh xám.....	23
2.1.1.3. Làm mịn ảnh	24
2.1.1.4. Xử lý hình thái ảnh.....	24
2.1.1.5. Cắt ngưỡng ảnh và tìm biên	26
2.1.2. Phương pháp xử lý văn bản.....	26
2.1.2.1. Tiền xử lý	27
2.1.2.2. Nhận dạng các thành phần	27
2.1.2.3. Nhận dạng ký tự	28
2.1.2.4. Hậu xử lý.....	28
2.2. PHÂN TÍCH VÀ THIẾT KẾ MÔ HÌNH	29
2.2.1. Sơ đồ Use Case	29
2.2.1.1. Sơ đồ Use Case quản trị viên	29
2.2.1.2. Sơ đồ Use Case cán bộ	31
2.2.1.3. Sơ đồ Use Case văn thư	32
2.2.1.4. Sơ đồ Use Case lãnh đạo đơn vị	33
2.2.2. Sơ đồ ER	34
2.2.3. Sơ đồ LDM.....	34
2.2.4. Sơ đồ LDM của MongoDB	35
2.2.5. Sơ đồ phân rã chức năng	36
2.2.6. Sơ đồ hoạt động.....	37

Ứng dụng optical character recognition vào hệ thống quản lý công văn

2.2.6.1. Sơ đồ hoạt động chức năng đăng nhập	37
2.2.6.2. Sơ đồ hoạt động tìm công văn đến.....	38
2.2.6.3. Sơ đồ hoạt động chức năng thêm công văn đến.....	39
2.2.6.4. Sơ đồ hoạt động chức năng sửa công văn đến	40
2.2.6.5. Sơ đồ hoạt động chức năng phê duyệt công văn đến	41
2.2.6.6. Sơ đồ hoạt động chức năng xử lý công văn đến	41
2.2.6.7. Sơ đồ hoạt động chức năng xóa công văn đến.....	42
2.2.6.8. Sơ đồ hoạt động thông kê công văn đến	42
2.2.6.9. Sơ đồ hoạt động thêm cán bộ	43
2.2.6.10. Sơ đồ hoạt động sửa cán bộ	44
2.2.6.11. Sơ đồ hoạt động xóa cán bộ	45
2.2.7. Sơ đồ tuần tự	46
2.2.7.1. Sơ đồ tuần tự chức năng đăng nhập	46
2.2.7.2. Sơ đồ tuần tự chức năng tìm công văn đến.....	47
2.2.7.3. Sơ đồ tuần tự chức năng thêm công văn đến	48
2.2.7.4. Sơ đồ tuần tự chức năng sửa công văn đến.....	49
2.2.7.5. Sơ đồ tuần tự chức năng phê duyệt công văn đến.....	50
2.2.7.6. Sơ đồ tuần tự chức năng xử lý công văn đến	50
2.2.7.7. Sơ đồ tuần tự chức năng xóa công văn đến	51
2.2.7.8. Sơ đồ tuần tự thông kê công văn đến	51
2.2.7.9. Sơ đồ tuần tự thêm cán bộ.....	52
2.2.7.10. Sơ đồ tuần tự sửa cán bộ	53
2.2.7.11. Sơ đồ tuần tự xóa cán bộ	53
2.3. GIAO DIỆN ỦNG DỤNG.....	54
2.3.1. Giao diện đăng nhập.....	54
2.3.2. Giao diện xem danh sách công văn đến.....	56
2.3.3. Xem chi tiết công văn đến.....	57
2.3.4. Giao diện thêm công văn đến.....	58

Ứng dụng optical character recognition vào hệ thống quản lý công văn

2.3.5. Giao diện sửa công văn đến	61
2.3.6. Giao diện phê duyệt công văn đến	61
2.3.7. Giao diện xử lý công văn đến	62
2.3.8. Giao diện xóa công văn đến	63
2.3.9. Giao diện thống kê công văn đến	64
CHƯƠNG 3 KIỂM THỬ VÀ ĐÁNH GIÁ	66
3.1. ĐÁNH GIÁ MÔ HÌNH.....	66
3.1.1. Mô tả tập dữ liệu	66
3.1.2. Đánh giá mô hình	68
3.2. KIỂM THỬ CHỨC NĂNG PHÂN TÍCH VĂN BẢN.....	71
3.2.1. Môi trường kiểm thử	71
3.2.2. Kết quả kiểm thử	72
PHẦN KẾT LUẬN	74
I. KẾT QUẢ ĐẠT ĐƯỢC	75
I.1. Kiến thức đạt được.....	75
I.2. Kinh nghiệm thực tiễn	75
I.3. Hệ thống.....	75
II. HẠN CHẾ	75
III. HƯỚNG PHÁT TRIỂN	76
TÀI LIỆU THAM KHẢO.....	77
PHỤ LỤC I BIỂU MÃU THƯỜNG GẶP.....	79
PHỤ LỤC II CÁC BẢNG DỮ LIỆU	88
PHỤ LỤC III SƠ ĐỒ HOẠT ĐỘNG CON.....	95
PHỤ LỤC IV SƠ ĐỒ TUẦN TỰ CON	103

DANH MỤC HÌNH ẢNH

Hình 1.1 Các loại của nhận dạng ký tự	7
Hình 1.2 OCR-A (Trên), OCR-B (Dưới)	8
Hình 1.3 Các thành phần của hệ thống OCR	9
Hình 1.4 Các vân đề trong việc chọn ngưỡng: Trên: hình ảnh mức xám gốc, Giữa: hình ảnh phân ngưỡng không tốt, Dưới: hình ảnh phân ngưỡng tốt.....	10
Hình 1.5 Các ký tự bị lỗi	10
Hình 1.6 Chuẩn hóa dữ liệu trong OCR.....	11
Hình 1.7 Các giải thuật được sử dụng nhiều nhất [10]	12
Hình 1.8 Sơ đồ bố trí các thành phần thẻ thức văn bản hành chính.....	15
Hình 1.9 Quy trình nhận công văn đến từ hệ thống	17
Hình 1.10 Quy trình nhận công văn đến từ ngoài hệ thống	17
Hình 1.11 Quy trình phát hành công văn	18
Hình 2.1 Các bước xử lý ảnh	22
Hình 2.2 Ảnh trước khi chuẩn hóa kích thước.....	23
Hình 2.3 Ảnh sau khi chuẩn hóa kích thước	23
Hình 2.4 Kết quả sau khi làm mịn ảnh: Trái – Trước khi làm mịn; Phải – Sau khi làm mịn.....	24
Hình 2.5 Kết quả kéo dãn văn bản	25
Hình 2.6 Kết quả khi dùng phép dãn ảnh.....	25
Hình 2.7 Kết quả sau khi kéo dãn và dùng phép đóng ảnh.....	25
Hình 2.8 Kết quả khi dùng phép co ảnh.....	26
Hình 2.9 Kết quả tìm biên	26
Hình 2.10 Các bước xử lý văn bản.....	27
Hình 2.11 Cấu trúc tập tin ARFF	27
Hình 2.12 Sơ đồ Use Case quản trị viên	30
Hình 2.13 Sơ đồ Use Case cán bộ	31
Hình 2.14 Sơ đồ Use Case văn thư	32

Ứng dụng optical character recognition vào hệ thống quản lý công văn

Hình 2.15 Sơ đồ Use Case lãnh đạo đơn vị	33
Hình 2.16 Sơ đồ ER	34
Hình 2.17 Sơ đồ LDM.....	34
Hình 2.18 Sơ đồ LDM của MongoDB	35
Hình 2.19 Sơ đồ phân rã chức năng	36
Hình 2.20 Sơ đồ hoạt động chức năng đăng nhập	37
Hình 2.21 Sơ đồ hoạt động tìm công văn đến.....	38
Hình 2.22 Sơ đồ hoạt động thêm công văn đến	39
Hình 2.23 Sơ đồ hoạt động chức năng sửa công văn đến	40
Hình 2.24 Sơ đồ hoạt động chức năng phê duyệt công văn đến	41
Hình 2.25 Sơ đồ hoạt động chức năng xử lý công văn đến	41
Hình 2.26 Sơ đồ hoạt động chức năng xóa công văn đến.....	42
Hình 2.27 Sơ đồ hoạt động thêm cán bộ	43
Hình 2.28 Sơ đồ hoạt động sửa cán bộ	44
Hình 2.29 Sơ đồ hoạt động xóa cán bộ	45
Hình 2.30 Sơ đồ tuần tự chức năng đăng nhập	46
Hình 2.31 Sơ đồ tuần tự chức năng tìm công văn đến.....	47
Hình 2.32 Sơ đồ tuần tự chức năng thêm công văn đến	48
Hình 2.33 Sơ đồ tuần tự chức năng sửa công văn đến.....	49
Hình 2.34 Sơ đồ tuần tự chức năng phê duyệt công văn đến.....	50
Hình 2.35 Sơ đồ tuần tự chức năng xử lý công văn đến	50
Hình 2.36 Sơ đồ tuần tự chức năng xóa công văn đến.....	51
Hình 2.37 Sơ đồ tuần tự thêm cán bộ	52
Hình 2.38 Sơ đồ tuần tự sửa cán bộ	53
Hình 2.39 Sơ đồ tuần tự xóa cán bộ	53
Hình 2.40 Giao diện đăng nhập.....	54
Hình 2.41 Thông báo chưa nhập đủ thông tin đăng nhập	54
Hình 2.42 Thông báo không tìm thấy người dùng	55

Ứng dụng optical character recognition vào hệ thống quản lý công văn

Hình 2.43 Thông báo sai mật khẩu	55
Hình 2.44 Thông báo tài khoản bị khóa vì đăng nhập sai nhiều lần.....	55
Hình 2.45 Thông báo tài khoản bị khóa.....	56
Hình 2.46 Giao diện danh sách công văn đến.....	56
Hình 2.47 Giao diện tìm kiếm nâng cao	57
Hình 2.48 Giao diện chi tiết công văn đến.....	57
Hình 2.49 Giao diện nhập file đính kèm.....	58
Hình 2.50 Giao diện kết quả xử lý công văn.....	59
Hình 2.51 Giao diện thêm công văn	59
Hình 2.52 Thông báo lỗi khi không nhập các trường bắt buộc của công văn đi	60
Hình 2.53 Thông báo thêm công văn đến thành công	60
Hình 2.54 Giao diện phê duyệt công văn đến	61
Hình 2.55 Hộp thoại chọn cán bộ xử lý công văn.....	61
Hình 2.56 Thông báo duyệt thành công	62
Hình 2.57 Hộp thoại từ chối công văn	62
Hình 2.58 Giao diện xử lý công văn	62
Hình 3.1 Biểu diễn tập dữ liệu	68
Hình 3.2 Biểu đồ đường thống kê tần số phân lớp đúng của 10 lần thực nghiệm	69
Hình 3.3 Kết quả cây quyết định với k=14	70

DANH MỤC BẢNG

Bảng 1.1 Ma trận confusion 2x2 hay bảng contingency.....	13
Bảng 1.2 Quy trình tiếp nhận công văn đến và đi trường Đại học Cần Thơ [14].....	18
Bảng 3.1 Bảng thống kê loại văn bản hành chính trong tập dữ liệu	66
Bảng 3.2 Bảng mô tả thuộc tính của tập dữ liệu	67
Bảng 3.3 Thống kê các lần chạy thực nghiệm bằng cách k-fold	68
Bảng 3.4 Ma trận Confusion	70
Bảng 3.5 Bảng thống số kết quả chạy giải thuật.....	71
Bảng 3.6 Kết quả kiểm thử chức năng phân tích văn bản đến.....	72

Ứng dụng optical character recognition vào hệ thống quản lý công văn

DANH MỤC BIỂU MẪU

Mẫu 1 Nghị quyết (cá biệt)	79
Mẫu 2 Quyết định (cá biệt) quy định trực tiếp.....	80
Mẫu 3 Văn bản có tên loại	81
Mẫu 4 Công văn	82
Mẫu 5 Công điện	83
Mẫu 6 Giấy mời	84
Mẫu 7 Giấy giới thiệu	85
Mẫu 8 Biên bản	86
Mẫu 9 Giấy nghỉ phép.....	87

DANH MỤC SƠ ĐỒ

Sơ đồ III.1 Sơ đồ hoạt động lấy mã số cán bộ mới	95
Sơ đồ III.2 Sơ đồ hoạt động nhập họ và tên lót cán bộ	95
Sơ đồ III.3 Sơ đồ hoạt động nhập tên cán bộ	96
Sơ đồ III.4 Sơ đồ hoạt động nhập địa chỉ email.....	96
Sơ đồ III.5 Sơ đồ hoạt động nhập số điện thoại	97
Sơ đồ III.6 Sơ đồ hoạt động nhập chức vụ.....	97
Sơ đồ III.7 Sơ đồ hoạt động chọn quyền.....	98
Sơ đồ III.8 Sơ đồ hoạt động chọn tổ chức.....	98
Sơ đồ III.9 Sơ đồ hoạt động chọn trạng thái cán bộ.....	98
Sơ đồ III.10 Sơ đồ hoạt động nhập mã số đăng nhập	99
Sơ đồ III.11 Sơ đồ hoạt động nhập mật khẩu.....	99
Sơ đồ III.12 Sơ đồ hoạt động chọn loại văn bản.....	99
Sơ đồ III.13 Sơ đồ hoạt động chọn trạng thái	100
Sơ đồ III.14 Sơ đồ hoạt động chọn cán bộ	100
Sơ đồ III.15 Sơ đồ hoạt động chọn ngày.....	100
Sơ đồ III.16 Sơ đồ hoạt động nhập mã công văn	101
Sơ đồ III.17 Sơ đồ hoạt động chọn ngôn ngữ	101
Sơ đồ III.18 Sơ đồ hoạt động chọn độ khẩn.....	102
Sơ đồ III.1 Sơ đồ tuần tự lấy mã số mới	103
Sơ đồ III.2 Sơ đồ tuần tự nhập họ và tên lót	103
Sơ đồ III.3 Sơ đồ tuần tự nhập tên	104
Sơ đồ III.4 Sơ đồ tuần tự nhập địa chỉ email	104
Sơ đồ III.5 Sơ đồ tuần tự nhập số điện thoại.....	105
Sơ đồ III.6 Sơ đồ tuần tự nhập chức vụ	105
Sơ đồ III.7 Sơ đồ tuần tự chọn quyền	106
Sơ đồ III.8 Sơ đồ tuần tự chọn tổ chức	106
Sơ đồ III.9 Sơ đồ tuần tự chọn trạng thái	107

Ứng dụng optical character recognition vào hệ thống quản lý công văn

Sơ đồ III.10 Sơ đồ tuần tự nhập mã số đăng nhập	107
Sơ đồ III.11 Sơ đồ tuần tự nhập mật khẩu	107

Ứng dụng optical character recognition vào hệ thống quản lý công văn

DANH MỤC KÝ HIỆU CHỮ VIẾT TẮT

STT	Từ khóa	Diễn giải
1	OCR	Optical Character Recognition (Nhận dạng ký tự quang học)

TÓM TẮT

Ngày nay với sự phát triển mạnh mẽ của việc ứng dụng công nghệ thông tin trong việc quản lý các cơ quan, doanh nghiệp, kèm theo việc ưu tiên sự chính xác trong việc quản lý giấy tờ, sổ sách. Từ đó nhu cầu ứng dụng công nghệ thông tin trong việc quản lý các loại tài liệu được tăng cao. Cùng với sự phổ biến của công nghệ OCR và lĩnh vực khai khoán dữ liệu. Từ đó đề tài “**Ứng dụng optical character recognition vào hệ thống quản lý công văn**” được đề xuất để giải quyết vấn đề trên.

Ứng dụng sử dụng cơ sở dữ liệu MongoDB và được chia thành hai thành phần là client và server. Ở client sử dụng ReactJS framework và CoreUI để xây dựng giao diện, đối với server sử dụng NodeJS với Express framework. Và sử dụng công cụ Weka để xây dựng mô hình dự đoán bằng thuật toán cây quyết định J48. Kết hợp công cụ Tesseract để nhận dạng và nhập liệu tự động các trường trong văn bản.

Ứng dụng sau khi hoàn thành sẽ hỗ trợ việc quản lý, xử lý, lưu trữ các loại công văn tài liệu một cách dễ dàng hơn. Giúp giảm sự sai sót cũng như giảm công sức cho văn thư trong quản lý ứng dụng.

Từ khóa: *NodeJS, Express, ReactJS, nhận dạng ký tự quang học, cây quyết định, công văn.*

ABSTRACT

Today with the strong development of the application of information technology in the management of agencies and businesses, together with the priority of accuracy in the management of papers and books. Since then, the demand for information technology applications in the management of documents has increased. Along with the popularity of OCR technology and the field of data mining. Understanding that problem, the topic "**An Application of Optical Character Recognition to the Official Dispatch Management System**" is proposed to solve that current problem.

The application uses the MongoDB database and is divided into two components, the client and the server. In the client use ReactJS framework and Reactstrap to build the interface, for the server use NodeJS with Express framework. And use Weka to build a model using the J48 decision tree algorithm. Combines the Tesseract engine for optical character recognition and automatic input of fields in the text.

When being completed, the application will support the management, handling and storage of all kinds of documents more easily. Helps to reduce errors as well as reduce the effort for clerical in application management.

Keyword: NodeJS, Express, ReactJS, optical character recognition, decision tree, documentary.

Ứng dụng optical character recognition vào hệ thống quản lý công văn

PHẦN GIỚI THIỆU

Ứng dụng optical character recognition vào hệ thống quản lý công văn

I. ĐẶT VẤN ĐỀ

Trong bất kì lĩnh vực nào thì công tác quản lý luôn giữ một vai trò vô cùng quan trọng, trong đó có công tác quản lý công văn. Việc quản lý công văn trong các đơn vị hành chính và doanh nghiệp luôn là vấn đề cấp thiết, đòi hỏi có sự đầu tư cả về nhân lực và trang thiết bị.

Trước đây, việc quản lý công văn, giấy tờ thường theo quy cách truyền thống là lưu trữ trên giấy tờ, trong các cặp, tủ hồ sơ nên gây ra không ít phiền phức trong việc tra cứu, lưu trữ và bảo quản.

Bên cạnh đó Chính phủ cũng ban hành nhiều văn bản nhằm khuyến khích việc chuyển đổi số trong công tác quản lý công văn như: Nghị định số 30/2020/NĐ-CP ngày 05/3/2020 của Chính phủ về công tác văn thư có quy định về hệ thống quản lý tài liệu điện tử; Nghị quyết số 26/NQ-CP ngày 15/4/2015 của Chính phủ ban hành chương trình hành động của Chính phủ thực hiện nghị quyết số 36-NQ/TW ngày 01/7/2014 của Bộ Chính trị Ban Chấp hành Trung ương Đảng về đẩy mạnh ứng dụng, phát triển công nghiệp thông tin đáp ứng yêu cầu phát triển bền vững và hội nhập quốc tế;

Thời gian gần đây việc áp dụng các hệ thống quản lý vào việc quản lý công văn cũng đã được sử dụng nhưng đa phần đều phải nhập liệu truyền thống. Do đó cần có một biện pháp quản lý tối ưu hơn trong việc quản lý công văn trong các đơn vị hành chính và doanh nghiệp.

Cùng với sự phổ biến của lĩnh vực Data Mining và công nghệ OCR trong thời gian gần đây. Nên em quyết định nghiên cứu xây dựng một hệ thống quản lý bằng công nghệ OCR kết hợp với các thuật toán trong Data Mining để giúp việc quản lý công văn diễn ra nhanh chóng, dễ dàng và ít sai sót hơn.

Vì vậy trong luận văn sẽ tập trung tìm hiểu các kỹ thuật, công nghệ cần thiết để xây dựng hệ thống **Ứng dụng Optical Character Recognition vào hệ thống quản lý công văn** làm đề tài luận văn tốt nghiệp của mình.

II. LỊCH SỬ GIẢI QUYẾT VẤN ĐỀ

Xây dựng một hệ thống quản lý công văn, nhận dạng phân loại văn bản hay nghiên cứu về các giải thuật Data Mining là những đề tài thu hút nhiều sự quan tâm của mọi người.

Trong khoa có các đề tài nghiên cứu như: **Tìm kiếm chuyên gia với giải thuật máy học cây quyết định C4.4** luận văn tốt nghiệp thạc sĩ của tác giả Văn Thị Xuân

Ứng dụng optical character recognition vào hệ thống quản lý công văn

Hồng năm 2010, bài báo khoa học **Phân loại văn bản với máy học vector hỗ trợ và cây quyết định** của hai tác giả Trần Cao Đệ và Phạm Nguyên Khang năm 2012, **Nghiên cứu nhận dạng chữ in trong ảnh scan, ứng dụng vào trích lọc thông tin trích yếu của văn bản hành chính** luận văn thạc sĩ của tác giả Tạ Đoàn Hiệp năm 2013.

Hay các hệ thống quản lý công văn của các cơ quan như: Hệ thống văn phòng điện tử của trường Đại học Cần thơ, phần mềm VNU-OFFICE của Đại học Quốc gia Hà Nội, phần mềm E_OFFICE của trường Đại học Mỏ - Địa chất.

Ngoài ra còn các phần mềm thương mại như: VietOCR, VNPT iOffice, DOCEYE, Team Drive.

III. MỤC TIÊU ĐỀ TÀI

Hệ thống được xây dựng với mục tiêu cung cấp đầy đủ các tính năng cho quy trình nhận/gửi công văn của một hệ thống quản lý công văn thông thường tích hợp thêm việc trích xuất và nhận dạng các thành phần của công văn:

- Văn thư dễ dàng tìm kiếm công văn theo các tiêu chí như, ngày ban hành, số ban hành, người ký, ... Dễ dàng quản lý trạng thái xử lý công văn cũng như phân loại công văn.
- Hỗ trợ việc nhận dạng và trích xuất các thành phần trên công văn giúp cho việc nhập liệu nhanh chóng và ít sai sót hơn.
- Hỗ trợ việc báo cáo hoạt động quản lý một cách trực quan, rõ ràng, nhanh chóng.
- Hỗ trợ quản trị viên quản lý hệ thống tốt hơn, dễ dàng thiết lập các dữ liệu và việc phân quyền được thực hiện linh hoạt.
- Giao diện hài hòa, thân thiện và hiện đại giúp người dùng dễ dàng sử dụng.

IV. ĐỐI TƯỢNG VÀ PHẠM VI NGHIÊN CỨU

Đối tượng nghiên cứu là một hệ thống quản lý công văn tích hợp các thuật toán Data Mining và công nghệ OCR.

Phạm vi nghiên cứu là các tính năng thêm mới công văn, xử lý các nghiệp vụ quản lý như: phê duyệt, phân công xử lý, xử lý, ban hành, Đặc biệt là chức năng trích xuất và nhận dạng các thành phần của công văn.

V. NỘI DUNG NGHIÊN CỨU

Nghiên cứu về quy trình quản lý công văn, nghiên cứu về công nghệ OCR, xử lý ảnh, các thuật toán về Data Mining.

Về lý thuyết:

- Tìm hiểu về Optical Character Recognition.
- Tìm hiểu về các phương pháp xử lý ảnh.
- Tìm hiểu về các thuật toán Data Mining.
- Tìm hiểu về quy trình quản lý công văn theo quy định của nhà nước.

Về kỹ thuật:

- Kỹ thuật lập trình web, kỹ thuật RESTful API.
- Sử dụng ReactJS, Redux, React Router kết hợp thư viện CoreUI để xây dựng trang web giao diện người dùng
- Sử dụng NodeJS và framework Express để xây dựng web server.
- Sử dụng MongoDB để lưu trữ dữ liệu
- Sử dụng Weka và thuật toán cây quyết định J48 để xây dựng mô hình dự đoán.
- Sử dụng công cụ Tesseract để nhận dạng ký tự.
- Kết hợp Weka và Tesseract vào hệ server.

VI. BỐ CỤC CỦA QUYỀN LUẬN VĂN

Bố cục của luận văn bao gồm các thành phần như sau:

Phần giới thiệu: Giới thiệu tổng quan về luận văn, đặt vấn đề, lịch sử giải quyết vấn đề, mục tiêu đề tài, đối tượng và phạm vi nghiên cứu, nội dung nghiên cứu.

Phần nội dung: Nội dung gồm 3 chương chính:

Chương 1: Mô tả bài toán

Chương 2: Thiết kế và cài đặt thuật toán

Chương 3: Kiểm thử và đánh giá

Phần kết luận: Tổng hợp kết quả đạt được và đề xuất hướng phát triển trong tương lai.

Ứng dụng optical character recognition vào hệ thống quản lý công văn

PHẦN NỘI DUNG

CHƯƠNG 1

MÔ TẢ BÀI TOÁN

1.1. MÔ TẢ CHI TIẾT BÀI TOÁN

Người dùng gồm có quản trị viên lãnh đạo đơn vị, văn thư và cán bộ. Mỗi người dùng gồm có các thông tin sau: mã số, chức vụ, tên địa chỉ email, số điện thoại. Mỗi người dùng sẽ thuộc một đơn vị trong cơ quan và chỉ thực hiện các chức năng dành cho cơ quan mình thuộc về. Thông tin của mỗi người dùng sẽ được quản lý tạo sẵn với tài khoản và mật khẩu mặc định sau đó mỗi người dùng sẽ tự thay đổi lại. Mỗi người dùng khi muốn sử dụng phải đăng nhập vào hệ thống mới có thể sử dụng.

Tương ứng với từng người dùng sẽ có các quyền sử dụng riêng cho phép người dùng thực hiện các chức năng khác nhau. Quyền sử dụng sẽ được quản trị hệ thống tạo riêng cho từng nhóm đối tượng. Quyền sử dụng bao gồm các thuộc tính: tên quyền, quyền đọc công văn, quyền thêm mới công văn, quyền cập nhật công văn, quyền đọc danh sách cán bộ, quyền thêm mới thông tin cán bộ, quyền cập nhật thông tin cán bộ, quyền xóa thông tin cán bộ, quyền đọc danh sách quyền, quyền tạo mới quyền, quyền cập nhật quyền, quyền xóa quyền, phạm vi quyền (0: phạm vi hệ thống, 1: phạm vi cơ quan).

Các chức năng của quản trị viên gồm: Cập nhật thông tin của cán bộ, khóa hoặc mở khóa tài khoản của cán bộ, cập nhật thông tin của đơn vị, cập nhật thông tin của loại công văn, tạo quyền quản lý cho cán bộ.

Cán bộ thường có các chức năng sau: Xem và tìm kiếm công văn đi/đến trong đơn vị của mình, xử lý các công văn được phân công.

Cán bộ văn thư đơn vị cũng chính là cán bộ thường vì vậy có các chức năng của cán bộ thường kèm theo các chức năng: Duyệt công văn đến, chuyển công văn đã phê duyệt cho cán bộ xử lý, cập nhật thông tin công văn đến (thêm mới, sửa, xóa), cập nhật thông tin công văn đi (thêm mới, sửa, xóa) và nhắc nhở xử lý công văn.

Tương tự như văn thư, cán bộ lãnh đạo cũng có các chức năng của cán bộ thường và phê duyệt công văn đến, chuyển công văn đã phê duyệt cho cán bộ xử lý, nhắc nhở xử lý công văn, phân quyền văn thư, xem tình hình xử lý công văn.

Đối với chức năng xem danh sách công văn: Mỗi cán bộ chỉ có thể xem được công văn của đơn vị thuộc về, những công văn có độ mật là không, những công văn do cán bộ đó nhập, duyệt hoặc xử lý.

Ứng dụng optical character recognition vào hệ thống quản lý công văn

Đối với chức năng tìm kiếm công văn sẽ có 2 loại là tìm kiếm cơ bản theo ký tự có trong một trường nào đó của công văn và tìm kiếm nâng cao theo từng trường của công văn (tìm theo ngày, theo tiêu đề, ...).

Mỗi đơn vị bao gồm các thông tin sau: tên đơn vị, mã đơn vị, địa chỉ email đơn vị, đơn vị đó có phải ở bên trong cơ quan không, cơ quan thuộc về.

Công văn gồm các thông tin cơ bản sau: Số công văn, ngày ban hành, trích dẫn, số trang, ý kiến bút phê, tên người ký, chức vụ người ký, hạn giải quyết, danh sách tập tin đính kèm. Đối với công văn đi có thêm số lượng ban hành. Đối với công văn đến kèm theo ngày đến và số đến.

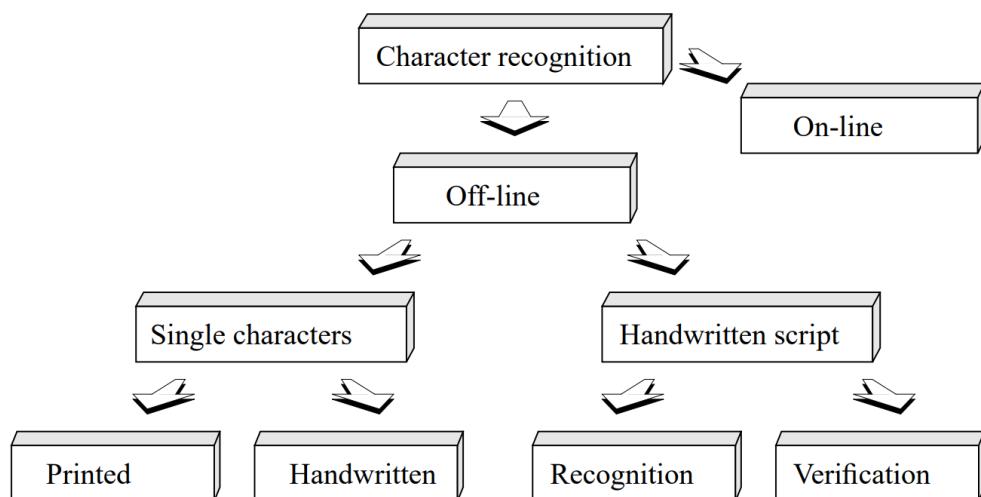
1.2. TIẾP CẬN VĂN ĐỀ VÀ LỰA CHỌN GIẢI PHÁP

1.2.1. Nhận dạng ký tự quan học (Optical Character Recognition)

1.2.1.1. Giới thiệu

Optical Character Recognition (OCR) là một hệ thống chuyển đổi hình ảnh của các văn bản đầu vào (chữ viết tay hoặc chữ đánh máy) thành mã máy. OCR được hình thành từ một lĩnh vực nghiên cứu về nhận dạng mẫu, trí tuệ nhân tạo và machine vision. [1], [2]

Optical Character Recognition giải quyết vấn đề nhận dạng các ký tự được xử lý quang học. Nhận dạng quang học được thực hiện ngoại tuyến sau khi viết hoặc in xong, trái ngược với nhận dạng trực tuyến khi máy tính nhận dạng các ký tự khi chúng được vẽ. Có thể nhận dạng được cả ký tự viết tay và ký tự in nhưng hiệu suất phụ thuộc vào chất lượng của dữ liệu đầu vào. [3]



Hình 1.1 Các loại của nhận dạng ký tự

Ứng dụng optical character recognition vào hệ thống quản lý công văn

1.2.1.2. Lịch sử của OCR [4]

Năm 1870, C.R. Carey ở Boston Massachusetts đã phát minh ra máy quét võng mạc, hệ thống truyền hình ảnh sử dụng các tế bào quang điện.

Năm 1940, phiên bản hiện đại hơn của OCR ra đời.

Năm 1950, Máy OCR đầu tiên ra đời.

Từ 1960 đến 1965, các hệ thống OCR thương mại đầu tiên xuất hiện.

Từ 1965 đến 1975, OCR thế hệ thứ hai với hiệu xuất cao và chi phí thấp hơn ra đời. Font chữ OCR-A và OCR-B được tạo ra để hỗ trợ việc nhận dạng dễ dàng hơn.

A	B	C	D	E	F	G	H	I	J	K	L
M	N	O	P	Q	R	S	T	U	V	W	X
Y	Z	1	2	3	4	5	6	7	8	9	0
A	B	C	D	E	F	G	H	I	J	K	L
M	N	O	P	Q	R	S	T	U	V	W	X
Y	Z	1	2	3	4	5	6	7	8	9	0

Hình 1.2 OCR-A (Trên), OCR-B (Dưới)

Từ năm 1975 đến 1985, thế hệ thứ ba với các thiết bị OCR đơn giản hơn nhờ sự xuất hiện của máy tính cá nhân và máy in laser.

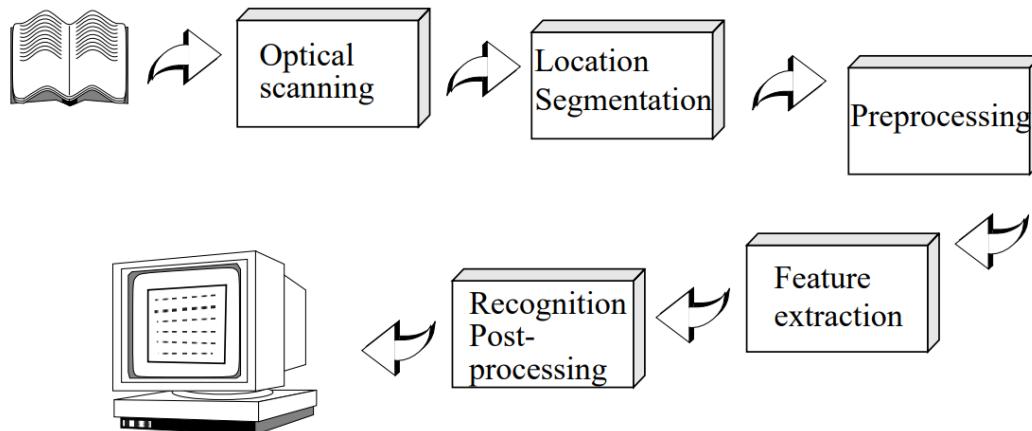
Ngày nay khi phần cứng ngày càng rẻ hơn và các hệ thống OCR đã có sẵn dưới dạng các gói phần mềm.

1.2.1.3. Phương pháp nhận dạng ký tự [4]

1.2.1.3.1. Các thành phần của hệ thống OCR

Một hệ thống OCR thông thường sẽ bao gồm một số thành phần như Hình 1.3. Bước đầu tiên là số hóa tài liệu bằng máy quét quang học. Khi các vùng chứa văn bản được định vị, mỗi biểu tượng được trích xuất thông qua quá trình phân đoạn. Các biểu tượng được trích xuất sau đó có thể được tiền xử lý, loại bỏ nhiễu, để tạo thuận lợi cho việc trích xuất các trong bước tiếp theo.

Ứng dụng optical character recognition vào hệ thống quản lý công văn



Hình 1.3 Các thành phần của hệ thống OCR

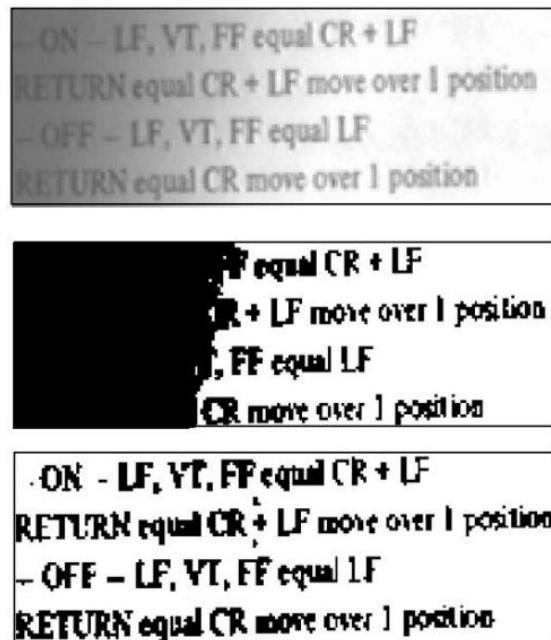
Các biểu tượng được tìm thấy bằng cách so sánh các đặc điểm được trích xuất với các mô tả về các lớp biểu tượng thu được qua giai đoạn học tập trước đó. Cuối cùng, thông tin theo ngữ cảnh được sử dụng để tái tạo lại các từ và số của văn bản gốc.

1.2.1.3.2. Optical scanning (quét quang học)

Hình ảnh được chụp bằng máy quét, sau đó sẽ chuyển đổi cường độ sáng thành các mức xám. Tài liệu in thường bao gồm chữ đen trên nền trắng. Do đó, thông thường là chuyển đổi hình ảnh đa cấp độ thành hình ảnh hai cấp độ đen trắng. Quá trình này được gọi là phân ngưỡng.

Quá trình tạo ngưỡng rất quan trọng vì kết quả nhận dạng hoàn toàn phụ thuộc vào chất lượng của hình ảnh đã phân ngưỡng.

Ứng dụng optical character recognition vào hệ thống quản lý công văn

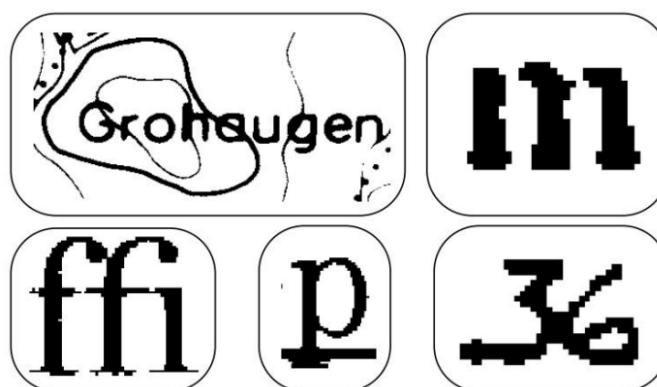


Hình 1.4 Các vấn đề trong việc chọn ngưỡng: Trên: hình ảnh mức xám gốc, Giữa: hình ảnh phân ngưỡng không tốt, Dưới: hình ảnh phân ngưỡng tốt.

1.2.1.3.3. Location and segmentation (định vị và phân đoạn)

Phân đoạn là quá trình xác định các thành phần của một hình ảnh. Cần xác định vị trí các vùng của tài liệu nơi dữ liệu đã được in ra và phân biệt chúng với các hình và đồ họa.

Áp dụng trong văn bản, phân đoạn là sự cô lập của các ký tự hoặc từ. Kỹ thuật này dễ thực hiện, nhưng sẽ gặp hạn chế nếu các ký tự chạm vào nhau hoặc các ký tự bị phân mảnh và bao gồm nhiều phần.



Hình 1.5 Các ký tự bị lỗi

1.2.1.3.4. Preprocessing (tiền xử lý)

Hình ảnh thu được từ quá trình quét có thể chứa một lượng nhiễu nhất định tùy thuộc vào máy quét. Các ký tự có thể bị nhòe hoặc bị hỏng. Những lỗi này làm cho việc nhận dạng bị kém đi, có thể được cải thiện bằng cách sử dụng bộ tiền xử lý để làm mịn các ký tự số hóa.



Hình 1.6 Chuẩn hóa dữ liệu trong OCR

1.2.1.3.5. Feature extraction (trích xuất đặc trưng)

Sử dụng các kỹ thuật như Template-matching and correlation (đối sánh và tương quan mẫu), Feature based (dựa trên tính năng) để tiến hành trích xuất các đặc trưng.

1.2.1.3.6. Recognition and Post-processing (nhận dạng và hậu xử lý)

Nhận dạng bằng cách phân loại từng ký tự vào lớp từng lớp chính xác. Có hai cách phân loại được sử dụng là Decision-theoretic methods (phương pháp lý thuyết quyết định) và Structural Methods (phương pháp cấu trúc).

Hậu xử lý bằng cách gom nhóm các từ lại thành một chuỗi để thể hiện đủ thông tin (Grouping) và phát hiện lỗi và sửa lỗi (Error-detection and correction).

1.2.2. Tesseract [5], [6]

Tesseract là một gói mã nguồn mở được phát triển bởi HP từ những năm 1984 đến 1994.

Tesseract ban đầu là một luận án nghiên cứu tiến sĩ trong phòng thí nghiệm HP, Bristol. Kết quả của công việc nghiên cứu là xây dựng một phần mềm hoặc phần cứng add on hỗ trợ cho máy quét HP phẳng.

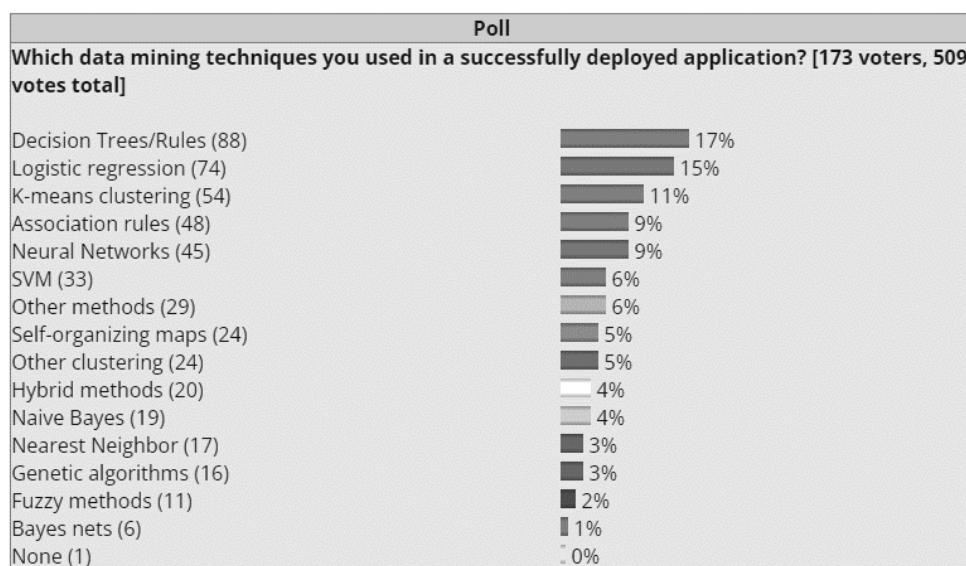
Sau dự án hợp tác giữa HP Labs, Bristol và bộ phận quét HP tại Colorado, độ chính xác của Tesseract tăng lên đáng kể theo các công cụ thương mại, nhưng không trở thành sản phẩm thương mại.

Ứng dụng optical character recognition vào hệ thống quản lý công văn

Giai đoạn phát triển tiếp theo của Tesseract chủ yếu được thực hiện ở HP Labs, Briston, tập trung nhiều vào việc cải thiện độ chính xác và dự án này kết thúc vào cuối năm 1994. Cuối năm 2005, HP công bố Tesseract là mã nguồn mở. Hiện nay đã có tại <http://code.google.com/p/tesseract-ocr>.

1.2.3. Giải thuật cây quyết định (Decision tree)

Cây quyết định là giải thuật khai khoáng dữ liệu rất phổ biến và hiệu quả [7], [8]. Cây quyết định có cấu trúc là một cây nhị phân, mỗi nút trong tương ứng với việc phân hoạch dựa trên một thuộc tính nào đó, nhánh đại diện cho một quy tắc quyết định. Việc xây dựng cây quyết định phụ thuộc vào việc lựa chọn thuộc tính để phân hoạch. [9], [10]



Hình 1.7 Các giải thuật được sử dụng nhiều nhất [11]

1.2.3.1. Giải thuật xây dựng cây quyết định [7]

Giải thuật học cây quyết định gồm 2 bước lớn: xây dựng cây (Top-down), cắt nhánh (Bottom-up) để tránh học vẹt. Quá trình xây dựng được làm như sau:

- Bắt đầu nút gốc, tất cả các dữ liệu học ở nút gốc.
- Nếu dữ liệu tại 1 nút có cùng lớp thì nút được cho là nút lá và nhãn của nút lá là nhãn của các phần tử trong nút lá (hay luận bình chọn số đông nếu nút lá có chứa các phần tử có lớp khác nhau).
- Nếu dữ liệu ở lớp có chứa các phần tử có lớp rất khác nhau (không thuần nhất) thì nút được cho là nút trong, tiến hành phân hoạch dữ liệu một cách đệ quy bằng việc chọn một thuộc tính để thực hiện phân hoạch tốt nhất có thể.

1.2.3.2. Nghi thức kiểm tra giải thuật [12]

Nghi thức **k-fold**: chia tập dữ liệu thành **k** phần (fold) bằng nhau, lặp lại **k** lần, mỗi lần sử dụng **k-1** folds để học và **1** fold để kiểm tra, sau đó tính trung bình của **k** lần kiểm tra. Khi tập dữ liệu có nhiều phần tử (hơn 300) thường chọn **k = 10**, ngược lại chọn **k = số phần tử**.

Nghi thức **hold-out**: chia tập dữ liệu thành 2 phần (thường là **2/3** tập dữ liệu và **1/3** tập dữ liệu) 1 phần dùng để học phần còn lại dùng để kiểm tra, có thể lặp lại **k** lần rồi tính trung bình.

1.2.3.3. Đo độ hiệu quả của giải thuật [12]

Hiệu quả của giải thuật có thể quan sát dựa vào dự đoán nhãn của giải thuật trên tập dữ liệu. Ma trận confusion cung cấp thông tin dự đoán nhãn của giải thuật, phần tử $[i, j]$ của ma trận trình bày số phần tử lớp **I** được dự đoán là lớp **j**.

Bảng 1.1 Ma trận confusion 2x2 hay bảng contingency

Dự đoán		Dương	Âm
Thực tế	Dương		
Dương	TP	FN	
Âm	FP	TN	

- **TP** tổng phần tử lớp **dương** được dự đoán là lớp **dương**.
- **FN** tổng phần tử lớp **dương** được dự đoán là lớp **âm**.
- **TN** tổng phần tử lớp **âm** được dự đoán là lớp **âm**.
- **FP** tổng phần tử lớp **âm** được dự đoán là lớp **dương**.

Các độ đo thường dùng là:

- Tỉ lệ dương tính thật.

$$TP\ Rate = \frac{TP}{TP + FN}$$

- Tỉ lệ dương tính giả.

$$FP\ Rate = \frac{FP}{FP + TN}$$

- **Độ chính xác**: Tỉ lệ số mẫu dự đoán dương đúng trong tổng số những mẫu được dự đoán là dương.

Ứng dụng optical character recognition vào hệ thống quản lý công văn

$$precision = \frac{TP}{TP + FP}$$

- **Độ tái hiện:** Tỉ lệ số mẫu dự đoán dương đúng trong số những mẫu dương.

$$recall = \frac{TP}{TP + FN}$$

- Tỉ lệ số mẫu dự đoán đúng trong tổng số mẫu.

$$accuracy = \frac{TP + TN}{TP + FN + TN + FP}$$

- **Độ đo F1:** Tỉ số dung hòa giữa recall và precision.

$$F1 = \frac{2 \times precision \times recall}{precision + recall}$$

1.2.4. Văn bản hành chính [13]

“**Văn bản**” là thông tin thành văn được truyền đạt bằng ngôn ngữ hoặc ký hiệu, hình thành trong hoạt động của các cơ quan, tổ chức và được trình bày đúng thể thức, kỹ thuật theo quy định.

“**Văn bản hành chính**” là văn bản hình thành trong quá trình chỉ đạo, điều hành, giải quyết công việc của các cơ quan, tổ chức.

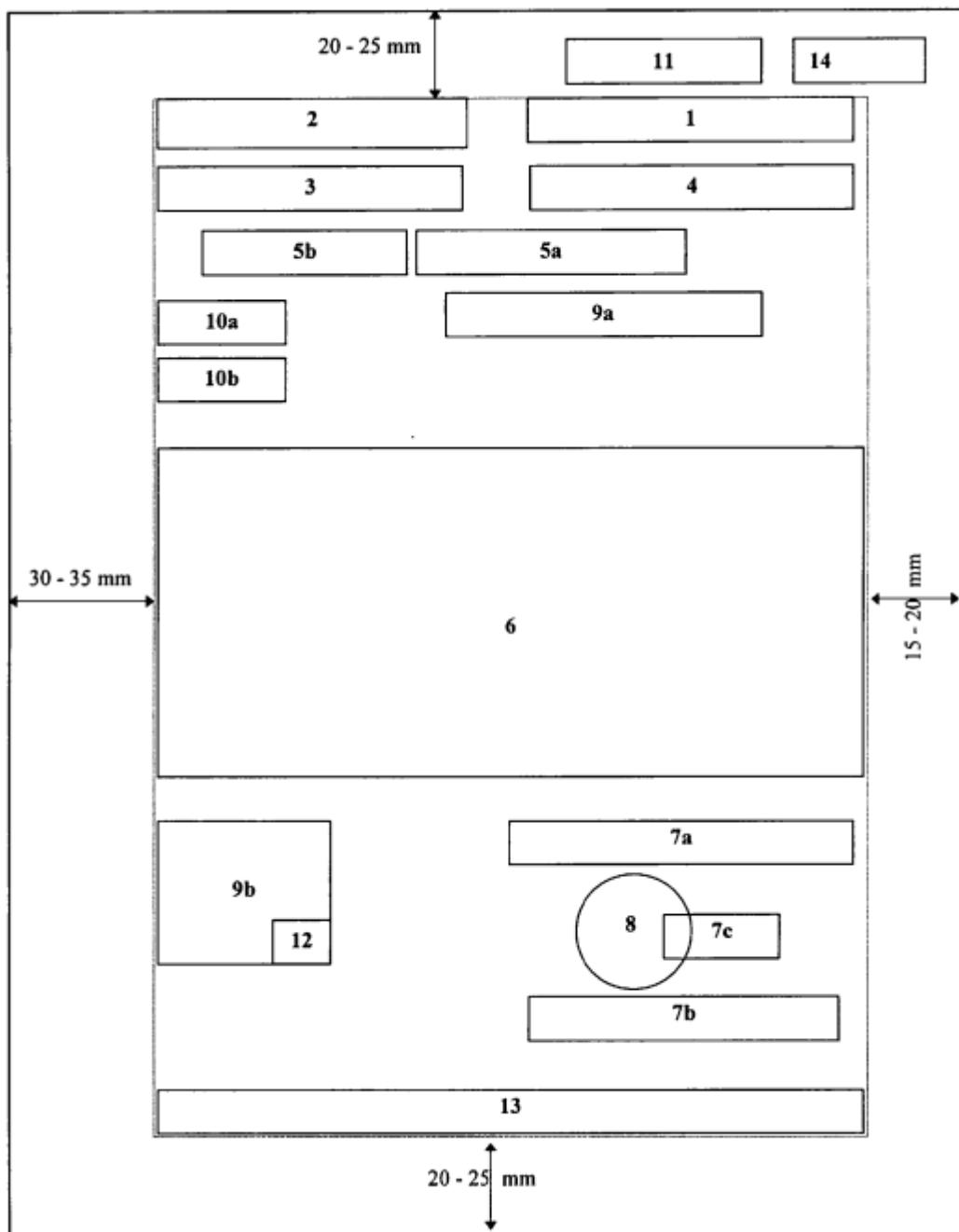
“**Văn bản điện tử**” là văn bản dưới dạng thông điệp dữ liệu được tạo lập hoặc được số hóa từ văn bản giấy và trình bày đúng thể thức, kỹ thuật, định dạng theo quy định.

“**Văn bản đi**” là tất cả các loại văn bản do cơ quan, tổ chức ban hành.

“**Văn bản đến**” là tất cả các loại văn bản do cơ quan, tổ chức nhận được từ cơ quan, tổ chức, cá nhân khác gửi đến.

Trình bày theo chiều dài của khổ A4 (210 mm x 297 mm). Sơ đồ bố trí các thành phần thể thức văn bản hành chính theo hướng dẫn của **Phụ lục I thể thức kỹ thuật trình bày văn bản hành chính và bản sao văn bản (Kèm theo nghị định số 30/2020/NĐ-CP ngày 05 tháng 3 năm 2020 của Chính phủ)**

Ứng dụng optical character recognition vào hệ thống quản lý công văn



Hình 1.8 Sơ đồ bố trí các thành phần thẻ thức văn bản hành chính

Vị trí trình bày các thành phần thẻ thức:

Ô số : **Thành phần thẻ thức văn bản**

- 1 : Quốc hiệu và Tiêu ngữ
- 2 : Tên cơ quan, tổ chức ban hành văn bản
- 3 : Số, ký hiệu của văn bản

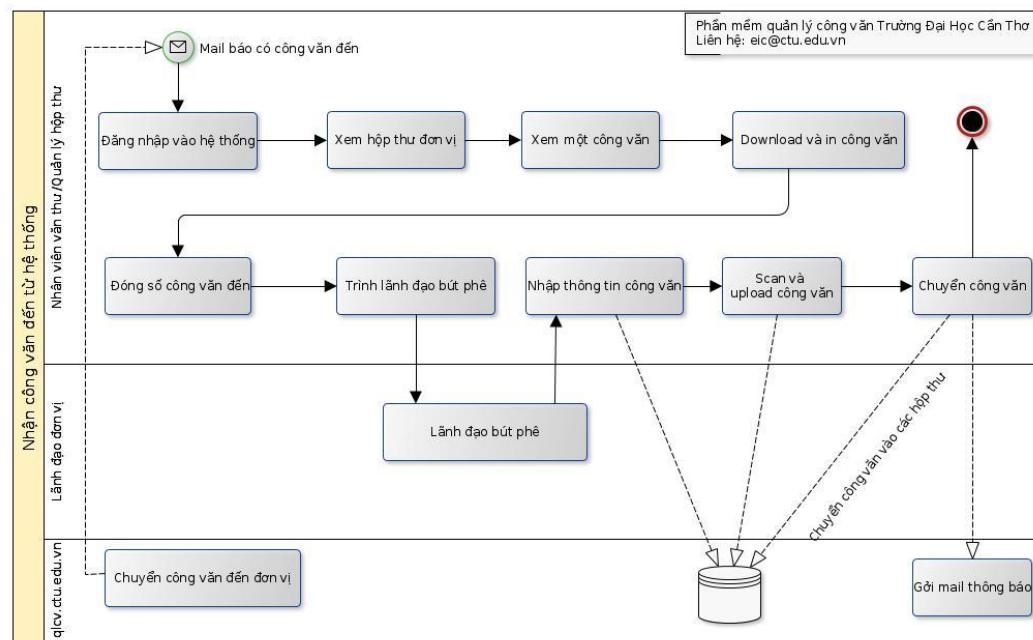
Ứng dụng optical character recognition vào hệ thống quản lý công văn

Ô số	: Thành phần thẻ thực văn bản
4	: Địa danh và thời gian ban hành văn bản
5a	: Tên loại và trích yếu nội dung văn bản
5b	: Trích yếu nội dung công văn
6	: Trích yếu nội dung công văn
7a, 7b, 7c	: Chức vụ, họ tên và chữ ký của người có thẩm quyền
8	: Dấu, Chữ ký số của cơ quan, tổ chức
9a, 9b	: Nơi nhận
10a	: Dấu chỉ độ mật
10b	: Dấu chỉ mức độ khẩn
11	: Chỉ dẫn về phạm vi lưu hành
12	: Ký hiệu người soạn thảo văn bản và số lượng bản phát hành
13	: Địa chỉ cơ quan, tổ chức; thư điện tử; trang thông tin điện tử; số điện thoại; số Fax
14	: Chữ ký số của cơ quan, tổ chức cho bản sao văn bản sang định dạng điện tử

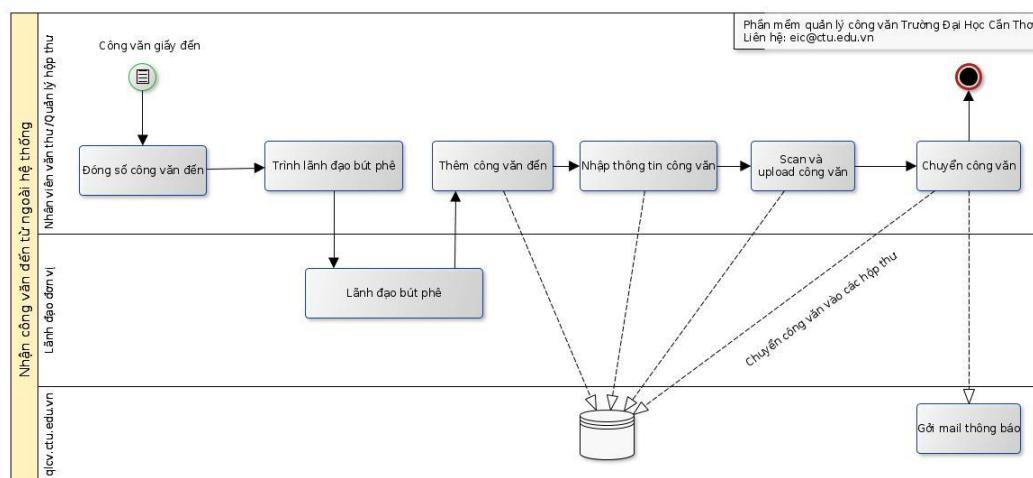
1.2.5. Quy trình quản lý công văn

Quy trình phát hành công văn đến và đi được tham khảo từ Hệ thống Quản lý công văn của Trường Đại học Cần Thơ [14]. Bao gồm 3 quy trình: Quy trình nhận công văn đến từ hệ thống, quy trình nhận công văn đến ngoài hệ thống, quy trình phát hành công văn.

Ứng dụng optical character recognition vào hệ thống quản lý công văn

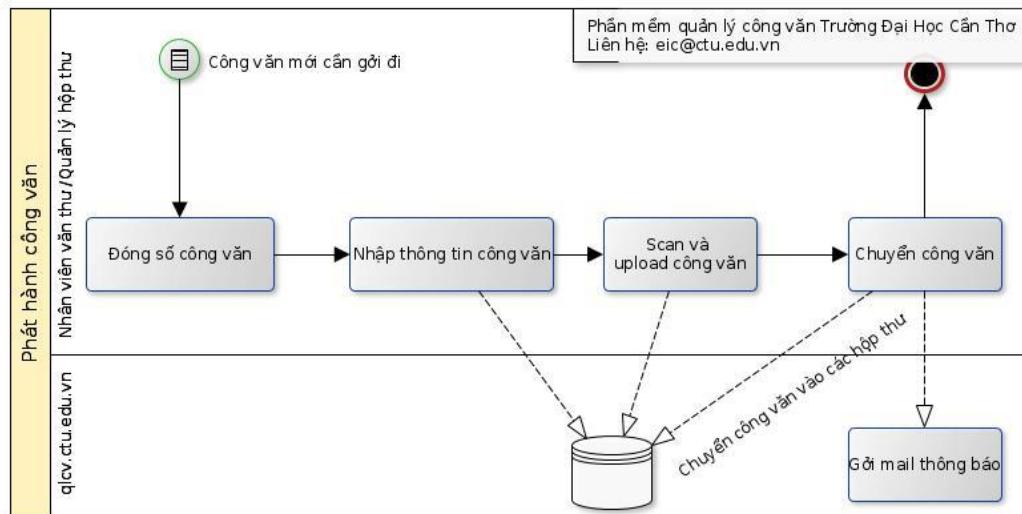


Hình 1.9 Quy trình nhận công văn đến từ hệ thống



Hình 1.10 Quy trình nhận công văn đến từ ngoài hệ thống

Ứng dụng optical character recognition vào hệ thống quản lý công văn



Hình 1.11 Quy trình phát hành công văn

Bảng 1.2 Quy trình tiếp nhận công văn đến và đi trường Đại học Cần Thơ [15]

Bước	Lưu đồ	Nội dung công việc	Người thực hiện	Thời gian thực hiện
1	<div style="border: 1px solid black; padding: 5px; text-align: center;"> Tiếp nhận văn bản </div>	<ul style="list-style-type: none"> – Tiếp nhận VB từ: Bưu điện, hộp thư điện tử, Fax, người gửi trực tiếp – Rà soát thể thức (nếu là VB trình BGH ký gửi đi) – Đóng dấu VB đến, ghi số và ngày đến, vào sổ VB đến 	<ul style="list-style-type: none"> – Nhân viên bưu điện, người gửi – Văn thư 	Trong các ngày làm việc
2	<div style="border: 1px solid black; padding: 5px; text-align: center;"> Phân loại văn bản </div>	Phân loại văn bản để trình các thành viên trong BGH	<ul style="list-style-type: none"> – Trưởng phòng KHTH – Thư ký BGH 	Trong các ngày làm việc
3	<div style="border: 1px solid black; padding: 5px; text-align: center;"> Phê duyệt </div>	BGH phê duyệt hoặc có ý kiến chỉ đạo để đối với văn bản	BGH	Trong ngày

Ứng dụng optical character recognition vào hệ thống quản lý công văn

Bước	Lưu đồ	Nội dung công việc	Người thực hiện	Thời gian thực hiện
4	<pre> graph TD A[] --> B[Xử lý văn bản] B --> C[Lưu trữ] </pre>	<p>Tiếp nhận ý kiến chỉ đạo, bút phê của BGH để triển khai thực hiện</p> <p>Lấy số, ngày VB, đóng dấu</p> <p>Ghi bút phê xử lý của BGH vào sổ VB đến</p> <p>Điện tử hóa văn bản (hoặc photocopy văn bản)</p> <p>Gởi đơn vị, cá nhân liên quan (gởi thông qua phần mềm quản lý, hoặc VB giấy)</p>	<ul style="list-style-type: none"> – Trưởng phòng KHTH – Văn thư phòng KHTH 	Trong các ngày làm việc
5	<pre> graph TD B[Lưu trữ] --> C[] </pre>	Phân loại và đưa vào kho lưu trữ	Văn thư phòng KHTH	Hàng tuần

1.2.6. Các công nghệ sử dụng

HTML (Hyper Text Markup Language): HTML là ngôn ngữ đánh dấu siêu văn bản. Cho phép tạo các trang web phối hợp hài hòa văn bản thông thường với hình ảnh âm thanh, video và mối liên kết đến các siêu văn bản khác. [16]

CSS (Cascading Style Sheets): CSS là một dạng tài liệu dùng để chứa thông tin về các mẫu định dạng mà tại liệu thông tin này có thể được nhiều trang web sử dụng. Các mẫu này dùng để định nghĩa phương thức hiển thị (đường kẻ, khung, khoảng cách giữa các dòng v.v.) và định dạng (màu chữ kiểu chữ màu nền v.v.) phần nội dung của trang web. [16]

JavaScript: JavaScript là ngôn ngữ dưới dạng script có thể gắn với các tập tin HTML được trình duyệt diễn dịch. Trình duyệt sẽ đọc JavaScript dưới dạng mã nguồn. [16]

Bootstrap: Bootstrap là một front-end framework miễn phí để phát triển web nhanh hơn và dễ hơn. Bootstrap bao gồm các mẫu thiết kế dựa trên HTML và CSS cho typography, forms, buttons, tables, navigation, modals, image carousels, ... cũng như các plugin JavaScript tùy chọn. Bootstrap cũng có khả năng tạo ra các trang web tương thích đa thiết bị.

CoreUI: CoreUI là một mẫu giao diện quản trị (Admin Template) dựa trên Bootstrap 4 và cung cấp 5 phiên bản: HTML5, AJAX, AngularJS, Angular4 và React. Với mã nguồn thuần túy, rõ ràng không có các thành phần dự phòng, vì vậy ứng dụng đủ nhẹ để cung cấp trải nghiệm người dùng cuối. API Layout CoreUI cho phép bạn tùy chỉnh dự án của mình cho hầu hết mọi thiết bị - có thể là thiết bị di động, web hoặc web. [17]

MongoDb: MongoDB là kho lưu trữ dữ liệu hướng tài liệu (document-oriented), mạnh mẽ, linh hoạt và dễ mở rộng. Kết hợp khả năng mở rộng quy mô với nhiều tính năng của một cơ sở dữ liệu quan hệ như là chỉ mục phụ, truy vấn phạm vi và sắp xếp. MongoDB cũng có rất nhiều tính năng hữu ích như hỗ trợ tích hợp cho các chỉ mục không gian địa lý và tổng hợp kiểu MapReduce. [18], [19]

React: React là một thư viện phổ biến được sử dụng để tạo giao diện người dùng. Được xây dựng tại Facebook để giải quyết một số vấn đề liên quan đến các trang web có dữ liệu và quy mô lớn. [20]

Nodejs: Node.js là công nghệ phía máy chủ (server-side) dựa trên công cụ JavaScript V8 của Google. Đó là một hệ thống có khả năng mở rộng cao sử dụng I/O (input/output) không đồng bộ, hướng sự kiện, thay vì các luồng hoặc các quy trình riêng biệt. Nó phù hợp cho các ứng dụng web được truy cập thường xuyên nhưng tính toán đơn giản. [21]

Express: Express được mô tả là “một khung ứng dụng web node.js nhỏ gọn và linh hoạt, cung cấp một bộ tính năng mạnh mẽ để xây dựng các ứng dụng web đơn, đa trang và kết hợp”. [22]

Weka: Weka còn có tên đầy đủ là Waikato Environment for Knowledge Analysis. Đây là bộ phần mềm mã nguồn mở được sử dụng miễn phí để khai thác dữ liệu thuộc các dự án nghiên cứu của đại học Waikato, New Zealand. Weka là tập hợp các thuật toán học máy cho các nhiệm vụ khai thác dữ liệu. Nó chứa các công cụ để chuẩn bị dữ liệu, phân loại, hồi quy, phân cụm, khai thác quy tắc kết hợp và trực quan hóa. [23]

Ứng dụng optical character recognition vào hệ thống quản lý công văn

OpenCV: OpenCV là một thư viện thị giác máy tính mã nguồn mở. Được viết bằng C và C++ và chạy trên Linux, Windows và Mac OS X. Phát triển trên các giao diện dành cho Python, Java, MATLAB và các ngôn ngữ khác, bao gồm chuyển thư viện sang Android và iOS cho các ứng dụng di động. OpenCV được thiết kế cho hiệu quả tính toán và tập trung mạnh vào các ứng dụng thời gian thực. Thư viện OpenCV chứa hơn 500 chức năng bao gồm nhiều lĩnh vực trong thị giác máy tính, bao gồm kiểm tra sản phẩm tại nhà máy, hình ảnh y tế, bảo mật, giao diện người dùng, hiệu chỉnh máy ảnh, tầm nhìn âm thanh nổi và người máy. [24]

CHƯƠNG 2

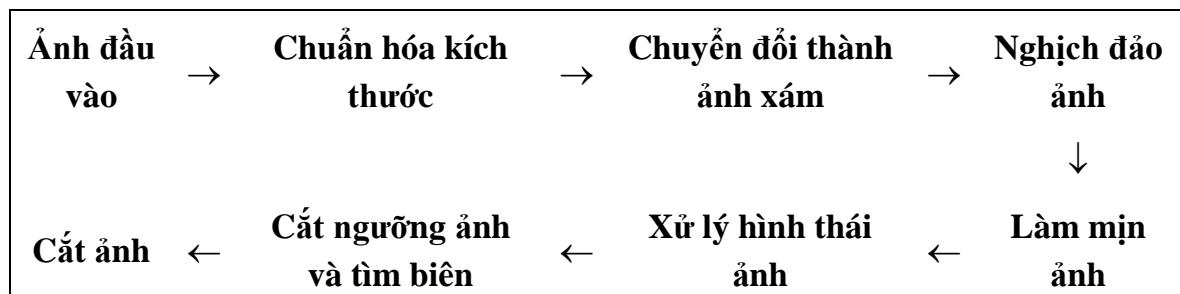
THIẾT KẾ VÀ CÀI ĐẶT THUẬT TOÁN

2.1. PHƯƠNG PHÁP THỰC HIỆN

2.1.1. Xử lý và trích xuất các thành phần từ ảnh văn bản

Hình ảnh đầu vào sẽ được chuẩn hóa kích thước để dữ liệu ảnh đầu vào được đồng bộ. Tiếp đến sẽ chuyển đổi không gian màu từ ảnh màu thành ảnh xám. Sau đó tiến hành nghịch đảo ảnh bằng cách đảo ngược từng bit của ảnh. Sau khi nghịch đảo ta tiến hành làm mịn ảnh và kéo giãn ảnh theo chiều cao. Tiếp đó ta mở rộng ảnh và tiếp tục kéo giãn ảnh theo chiều cao thêm một lần nữa. Tiếp sau đó đóng ảnh và điều chỉnh kích thước trở lại kích thước chuẩn để thực hiện bước tiếp theo. Sau cùng là làm xói mòn ảnh.

Sau khi thực hiện xong các bước trên ta tiến hành cắt ngưỡng ảnh để tìm biên. Từ các biên tìm được ta sẽ cắt ảnh thành các thành phần trong văn bản.



2.1.1.1. Chuẩn hóa kích thước ảnh

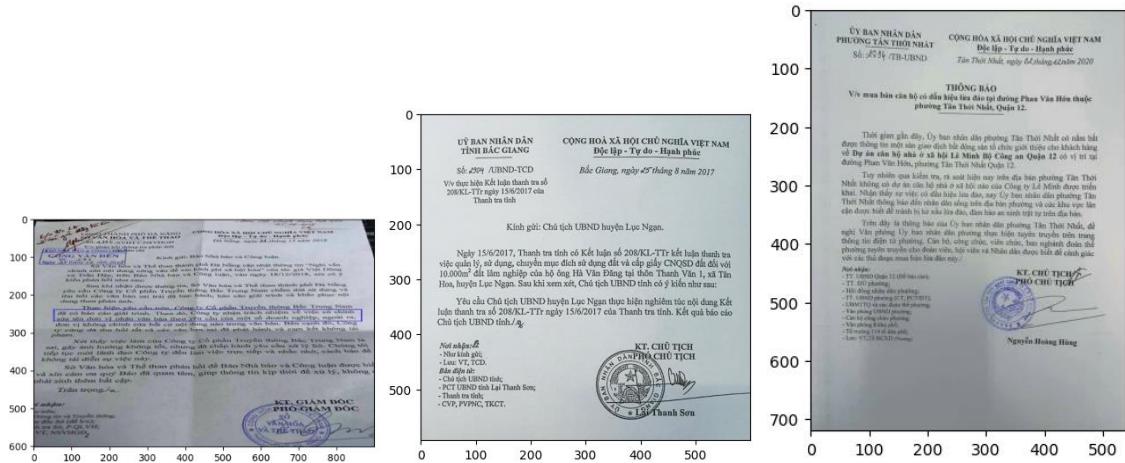
Tùy theo chất lượng và loại văn bản mà ảnh đầu vào có nhiều kích thước khác nhau. Vì vậy trước tiên cần chuẩn hóa kích thước ảnh đầu vào. Ta đưa tất cả các ảnh về kích thước của khổ giấy A4 (210 mm x 297 mm) với DPI là 300. [13]

Công thức chuyển đổi mm sang pixels:

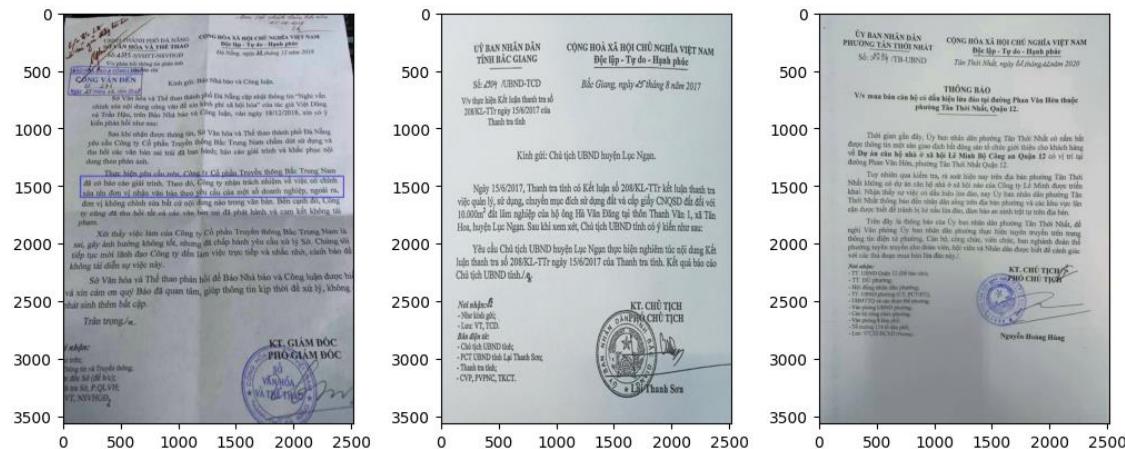
$$pixels = \frac{dpi \times mm}{1 \text{ inch}} = \frac{300 * mm}{25.4} = 12 * mm$$

Vậy kích thước khổ giấy A4 khi chuyển sang pixels là 2520 pixels x 3564 pixels.

Ứng dụng optical character recognition vào hệ thống quản lý công văn



Hình 2.2 Ảnh trước khi chuẩn hóa kích thước



Hình 2.3 Ảnh sau khi chuẩn hóa kích thước

2.1.1.2. Chuyển đổi thành ảnh xám

Ảnh màu thực chất chỉ là tập hợp của những ma trận số có cùng kích thước. Khi muốn xử lý thông tin trên ảnh, sẽ dễ dàng hơn nếu ta chỉ xử lý dữ liệu trên một ma trận số thay vì nhiều ma trận số.

Công thức chuyển từ ảnh màu RGB sang ảnh xám trong OpenCV [25]:

$$Y = 0.299 \times R + 0.578 \times G + 0.114 \times B$$

Trong đó:

- Y: ma trận xám cần tìm.
- R: ma trận đỏ.
- G: ma trận lục.
- B: ma trận lam.

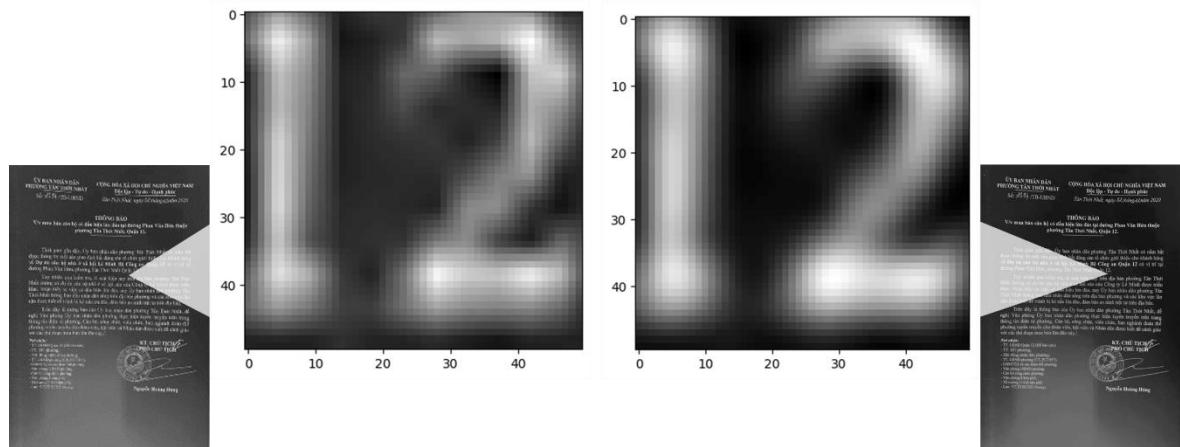
2.1.1.3. Làm mịn ảnh

Lọc ảnh (làm mịn ảnh, làm mượt ảnh) là một bước quan trọng trong xử lý ảnh. Có tác dụng như loại bỏ nhiễu, tìm biên đồi tượng. Trong openCV có nhiều phương pháp lọc như: Averaging (lọc trung bình), Gaussian Blurring (lọc Gauss), Median Blurring (lọc trung vị),

Trong khuôn khổ luận văn này sẽ chọn phương pháp lọc trung bình với ma trận kích thước 11x11.

Ma trận lọc của lọc trung bình có dạng:

$$K = \frac{1}{121} \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix}$$



Hình 2.4 Kết quả sau khi làm mịn ảnh: Trái – Trước khi làm mịn; Phải – Sau khi làm mịn

2.1.1.4. Xử lý hình thái ảnh

Xử lý hình thái ảnh có các tác dụng như: giúp loại bỏ nhiễu, làm đẹp cấu trúc và hình thức của ảnh nhị phân. Từ đó giúp ta dễ dàng tìm kiếm được các đối tượng hay cụ thể là các vùng trong văn bản. Các phép toán xử lý hình thái bao gồm: phép dãn (Dilation), phép co (Erosion), phép đóng (Closing) và phép mở (Opening).

Ứng dụng optical character recognition vào hệ thống quản lý công văn

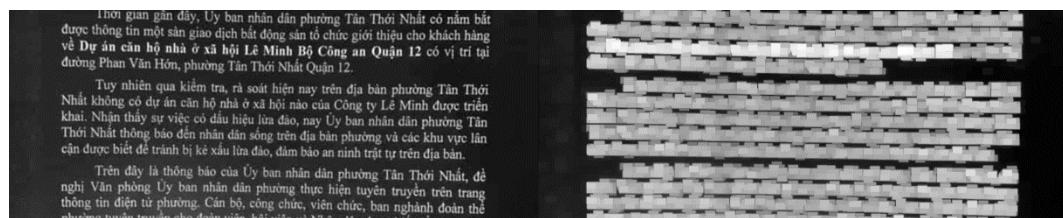
Trong khuôn khổ luận văn này sẽ kết hợp phép dãn, phép đóng và phép co để tìm kiếm các vùng trong văn bản

Đầu tiên kéo dãn ảnh theo chiều cao để phân tách các đoạn văn trong văn bản.



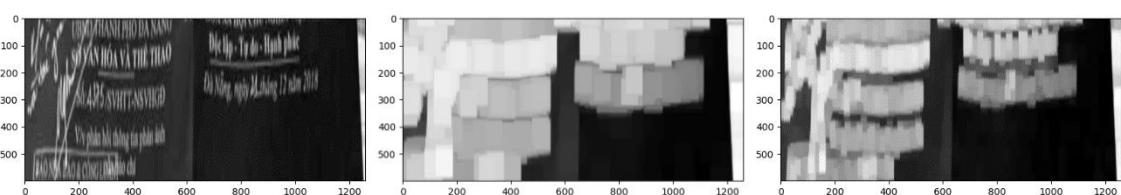
Hình 2.5 Kết quả kéo dãn văn bản

Tiếp theo dùng phép dãn để hợp nhất các từ trong câu và các câu trong đoạn văn.



Hình 2.6 Kết quả khi dùng phép dãn ảnh

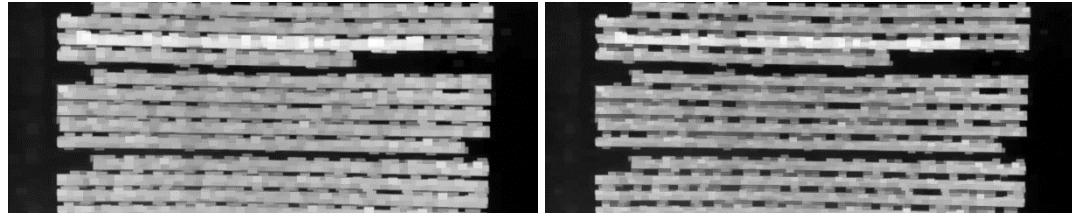
Sau đó ta tiếp tục kéo dãn và dùng phép đóng ảnh. Thao tác này giúp tách phần ngày tháng trong văn bản khỏi phần Quốc hiệu - Tiêu ngữ.



Hình 2.7 Kết quả sau khi kéo dãn và dùng phép đóng ảnh

Cuối cùng ta sử dụng phép co ảnh để hạn chế các đoạn văn bị dính vào nhau.

Ứng dụng optical character recognition vào hệ thống quản lý công văn

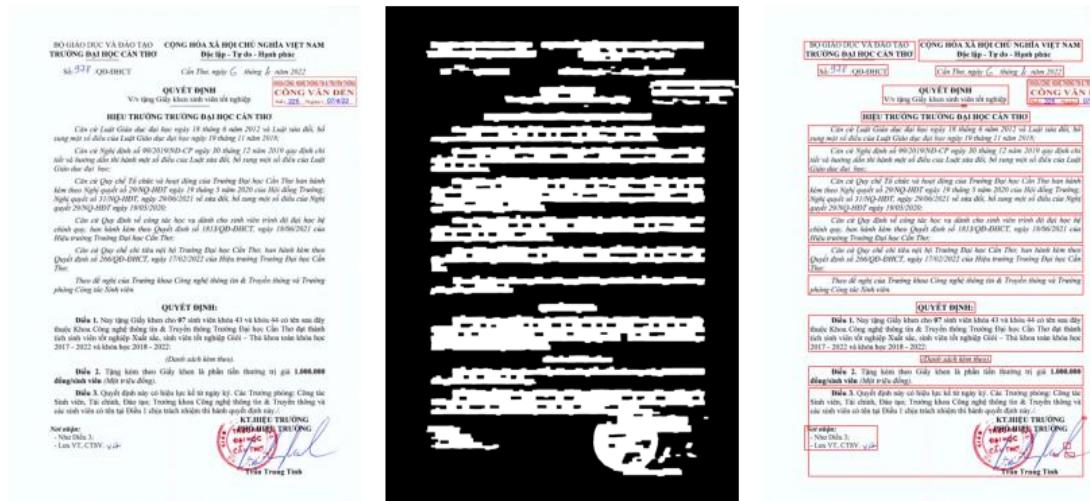


Hình 2.8 Kết quả khi dùng phép co ảnh

2.1.1.5. Cắt ngưỡng ảnh và tìm biên

Sau các bước xử lý ở trên ta tiến hành bước cuối cùng là cắt ngưỡng để tìm biên của các đối tượng trên văn bản.

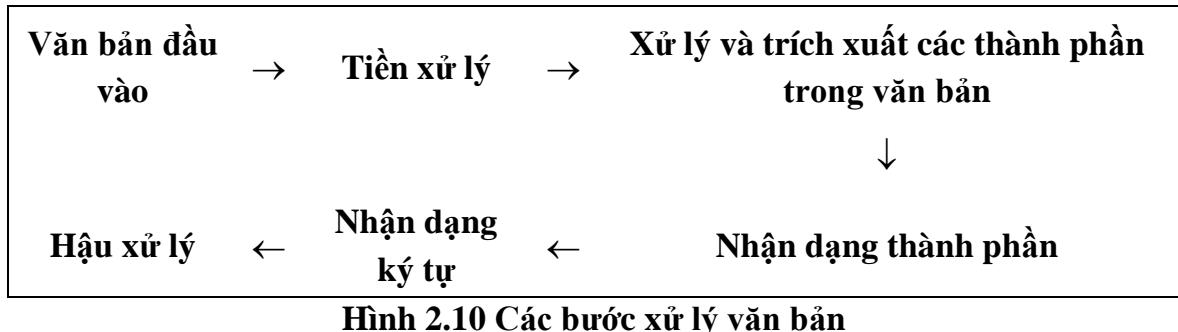
Phân ngưỡng là phương pháp mà chúng ta sẽ sử dụng khi muốn chuyển đổi từ ảnh xám thành ảnh nhị phân – giá trị các điểm ảnh chỉ bao gồm hai giá trị 0 hoặc 1. Sau khi dùng kỹ thuật phân ngưỡng ta thu được một ảnh nhị phân thể hiện rõ các đối tượng có trong văn bản.



Hình 2.9 Kết quả tìm biên

2.1.2. Phương pháp xử lý văn bản

Văn bản đầu vào là một tập tin PDF hoặc một tập tin hình ảnh sẽ được tiền xử lý để thu được hình ảnh văn bản cho quá trình sau. Tiếp đến là quá trình xử lý ảnh văn bản đã được trình bày trong phần Xử lý và trích xuất các thành phần từ ảnh văn bản. Sau quá trình đó ta dùng Weka và mô hình đã đào tạo trước đó để nhận dạng các thành phần trong văn bản. Khi thu được các thành phần trong văn bản, ta tiến hành nhận dạng ký tự bằng Tesseract. Cuối cùng hậu xử lý bằng cách sử dụng biểu thức chính quy để trích xuất các thành phần cần thiết.



2.1.2.1. Tiền xử lý

Ở bước này ta sẽ sử dụng công cụ ImageMagick để chuyển đổi tập tin đầu vào thành tập tin hình ảnh. Do ImageMagick chỉ nhận tập tin đầu vào là PDF và các tập tin hình ảnh, nên ta cần kiểm tra trước tập tin đầu vào có phù hợp không.

Câu lệnh chuyển đổi của ImageMagick:

```
magick convert -density 150 input.pdf -quality 90 output.jpg
```

-density: mật độ được sử dụng để chỉ định DPI của hình ảnh đầu ra.

input.pdf: tập tin đầu vào

-quality: chỉ định chất lượng cho hình ảnh được tạo.

output.jpg: tập tin hình ảnh đầu ra

2.1.2.2. Nhận dạng các thành phần

Sau khi có được các thành phần ta tiến hành nhận dạng bằng các bước sau

Bước 1: Tạo tập tin ARFF từ mỗi thành phần. Tập tin ARFF có dạng như sau:

```

@attribute x numeric
@attribute y numeric
@attribute w numeric
@attribute h numeric
@attribute page {0,1}
@attribute multi {0,1}
@attribute od {0,1}
@attribute type {1,2,3,4,5,6,7,8,9}

@data
x,y,w,h,page,multi,od,?
  
```

Hình 2.11 Cấu trúc tập tin ARFF

Ứng dụng optical character recognition vào hệ thống quản lý công văn

Trong đó:

x,y,w,h: lấy từ các thành phần.

page, od: do người dùng nhập.

mutil: do hệ thống phát hiện.

? : phần mà Weka sẽ dự đoán

Bước 2: Dùng câu lệnh của Weka để dự đoán. Câu lệnh có dạng:

```
java -cp weka.jar weka.classifiers.trees.J48 -p 0 -l  
file.model -T file.arff
```

Trong đó:

-p: cho ra kết quả dự đoán

-l: tập tin mô hình đã được huấn luyện

-T: tập tin ARFF dùng để dự đoán

Sau khi dự đoán ta có kết quả dự đoán có dạng như sau:

inst#	actual	predicted	error	prediction
1	1 : ?	2 : 2		0 . 985

Trong đó 2 giá trị predicted và prediction lần lượt là giá trị dự đoán và độ chính xác. Trong ví dụ này giá trị dự đoán là 2 (Tên cơ quan tổ chức ban hành) và độ chính xác là 98,5%.

2.1.2.3. Nhận dạng ký tự

Sau khi nhận dạng các thành phần ta sẽ loại các thành phần được dự đoán là 1 (Quốc hiệu Tiêu ngữ) và 6 (phần nội dung hoặc lỗi) không nhận dạng ký tự. Vì ta không dùng đến hai thành phần đó trong lưu trữ. Sau bước loại bỏ các phần là 1 và 6 ta dùng thư viện tesseract.js để tiến hành nhận dạng ký tự.

2.1.2.4. Hậu xử lý

Sau bước nhận dạng ta sẽ thu được văn bản trong các thành phần. Tuy nhiên các văn bản thu được cần phải trích xuất thêm một lần nữa bằng các kỹ thuật xử lý chuỗi như biểu thức chính quy để thu được kết quả tốt nhất.

2.2. PHÂN TÍCH VÀ THIẾT KẾ MÔ HÌNH

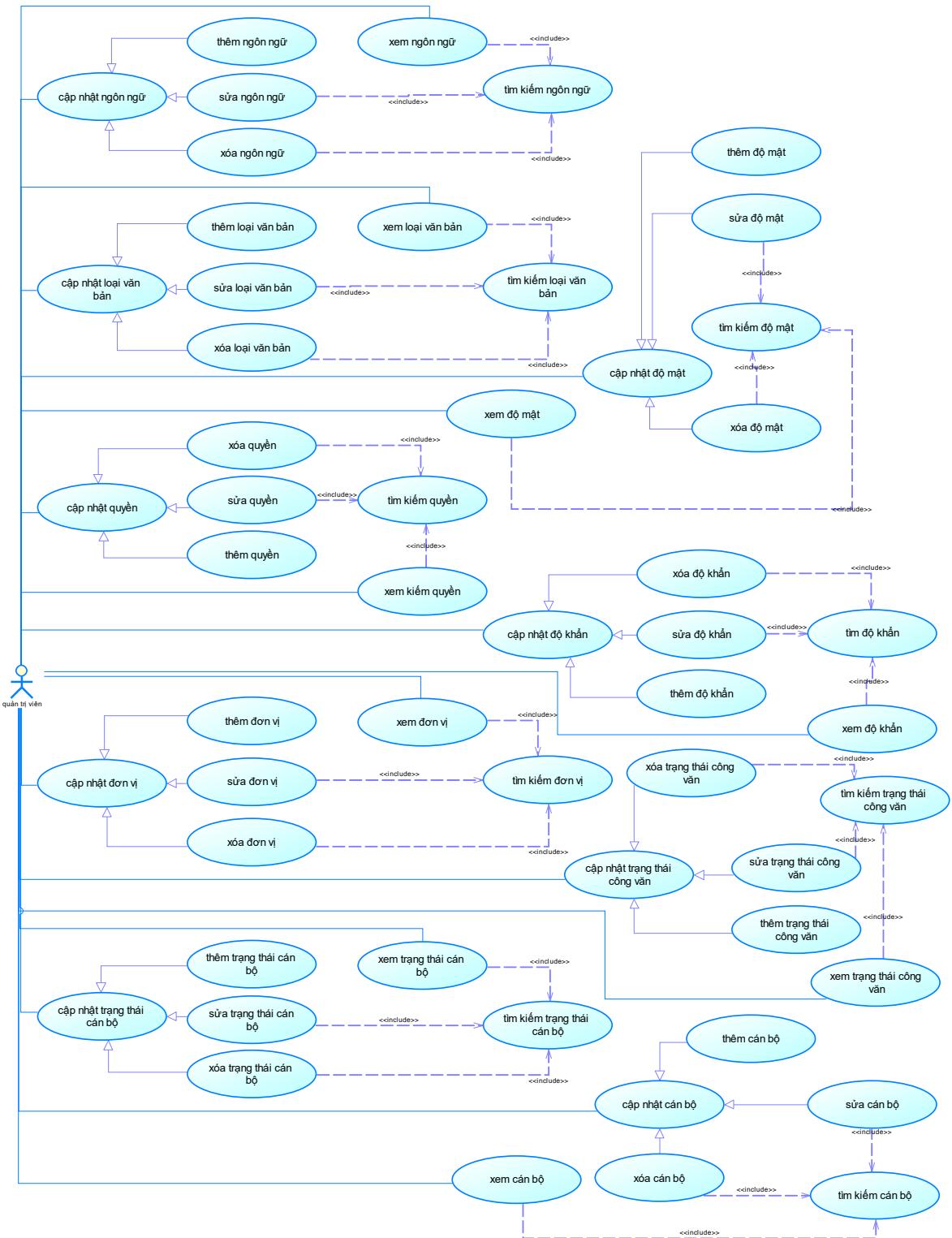
2.2.1. Sơ đồ Use Case

Tất cả các tác nhân đều phải đăng nhập để có thể sử dụng được các trường hợp sử dụng.

2.2.1.1. Sơ đồ Use Case quản trị viên

Quản trị viên sau khi đăng nhập vào hệ thống sẽ có các chức năng quản lý các thông tin về ngôn ngữ, loại văn bản, độ mật độ khẩn, quyền, đơn vị, trạng thái công văn, trạng thái cán bộ và cán bộ.

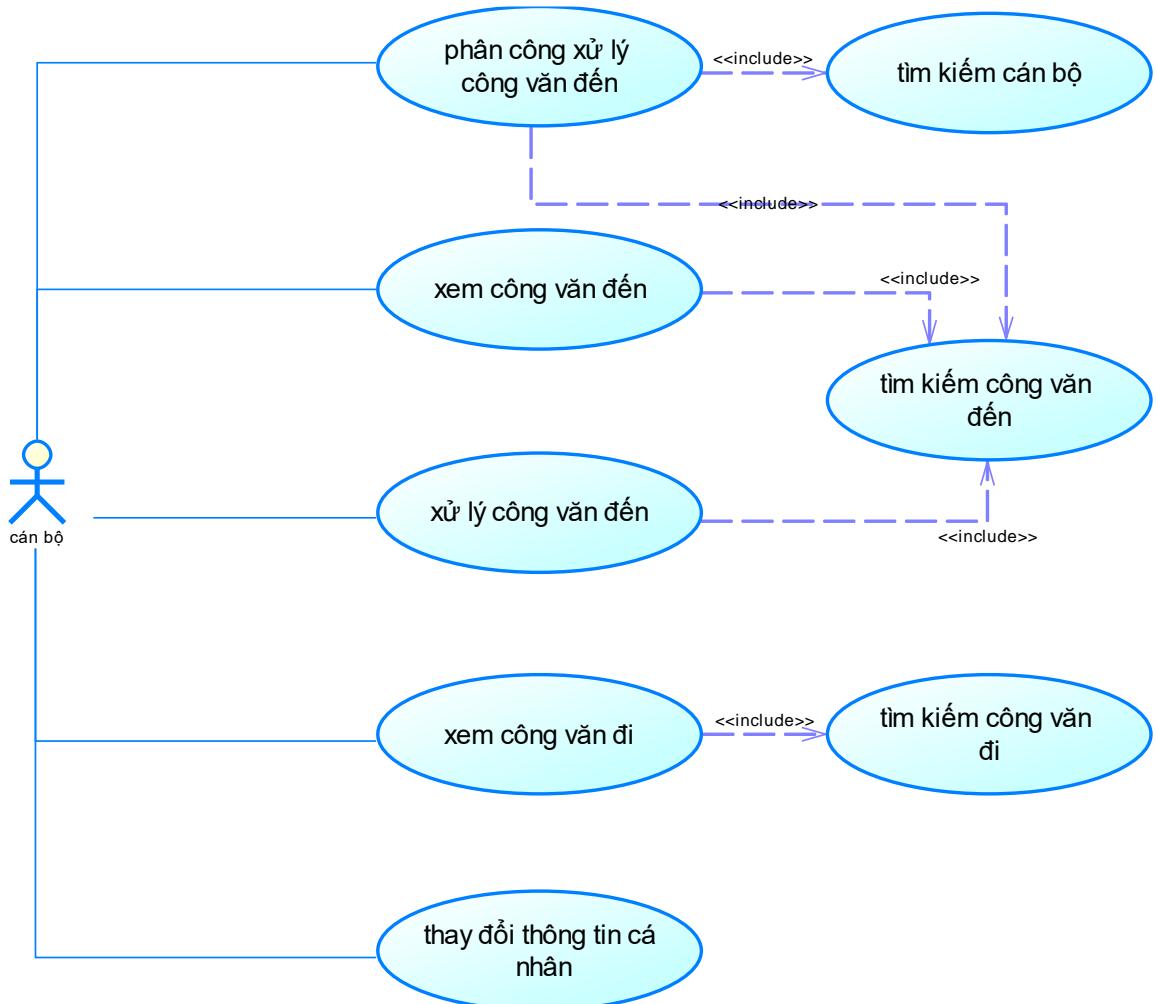
Ứng dụng optical character recognition vào hệ thống quản lý công văn



Hình 2.12 Sơ đồ Use Case quản trị viên

2.2.1.2. Sơ đồ Use Case cán bộ

Cán bộ sau khi đăng nhập vào hệ thống sẽ có các chức năng thay đổi thông tin cá nhân xem công văn đi, xem công văn đến, xử lý công văn đến được phân công và phân công xử lý cho cán bộ khác.

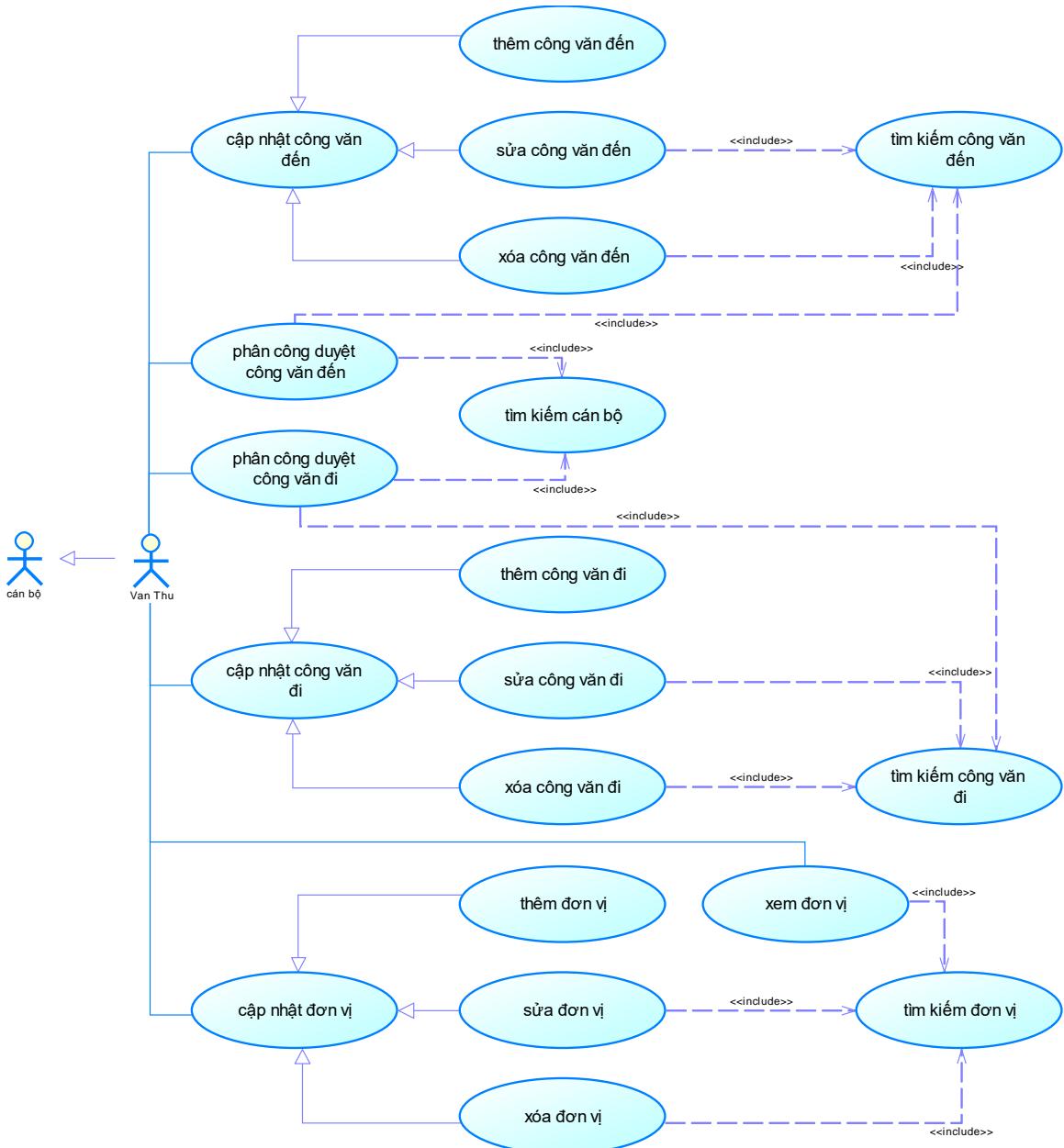


Hình 2.13 Sơ đồ Use Case cán bộ

Ứng dụng optical character recognition vào hệ thống quản lý công văn

2.2.1.3. Sơ đồ Use Case văn thư

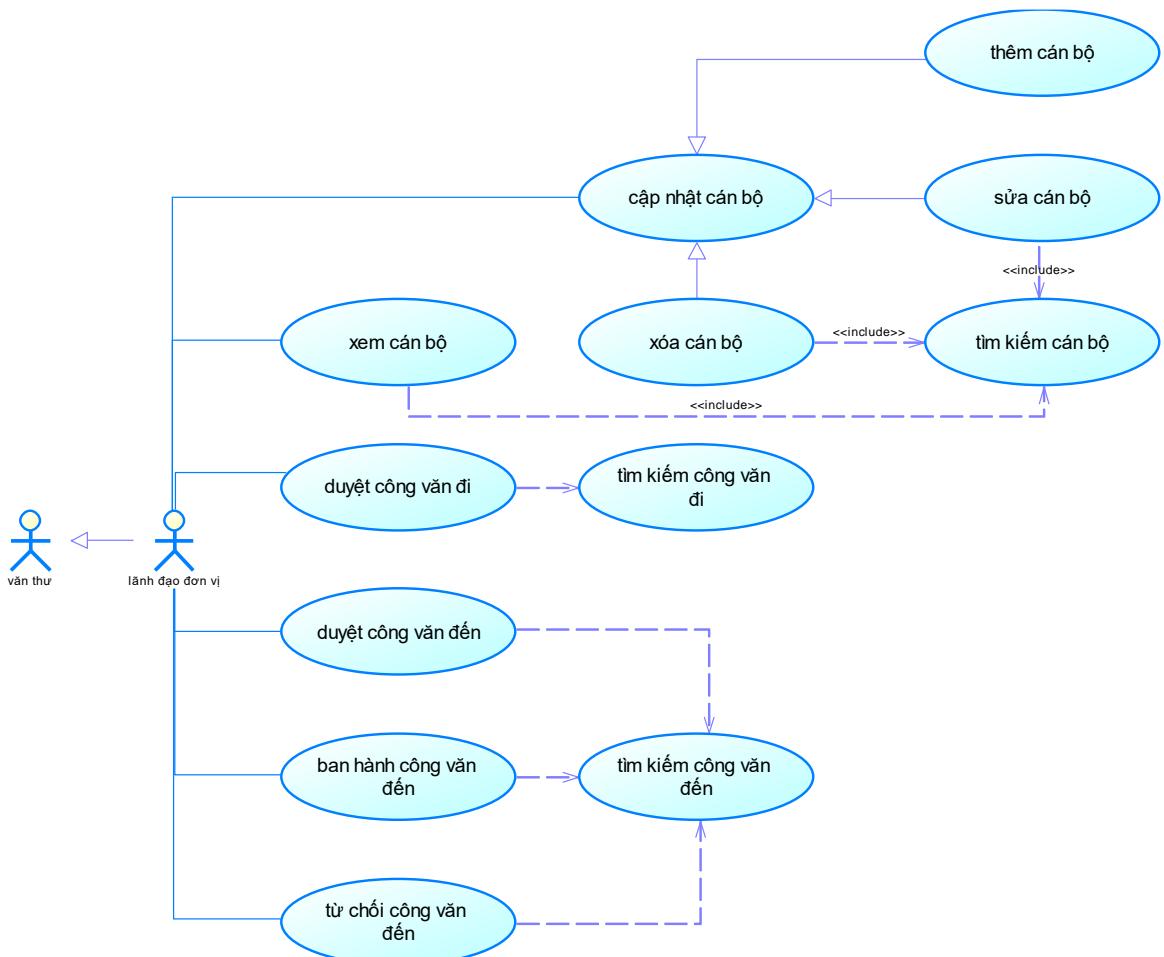
Văn thư sau khi đăng nhập vào hệ thống sẽ có tất cả các chức năng của cán bộ. Ngoài ra còn có các chức năng quản lý công văn đi, công văn đến, quản lý đơn vị con của đơn vị mình thuộc về và phân công duyệt công văn đi, công văn đến.



Hình 2.14 Sơ đồ Use Case văn thư

2.2.1.4. Sơ đồ Use Case lãnh đạo đơn vị

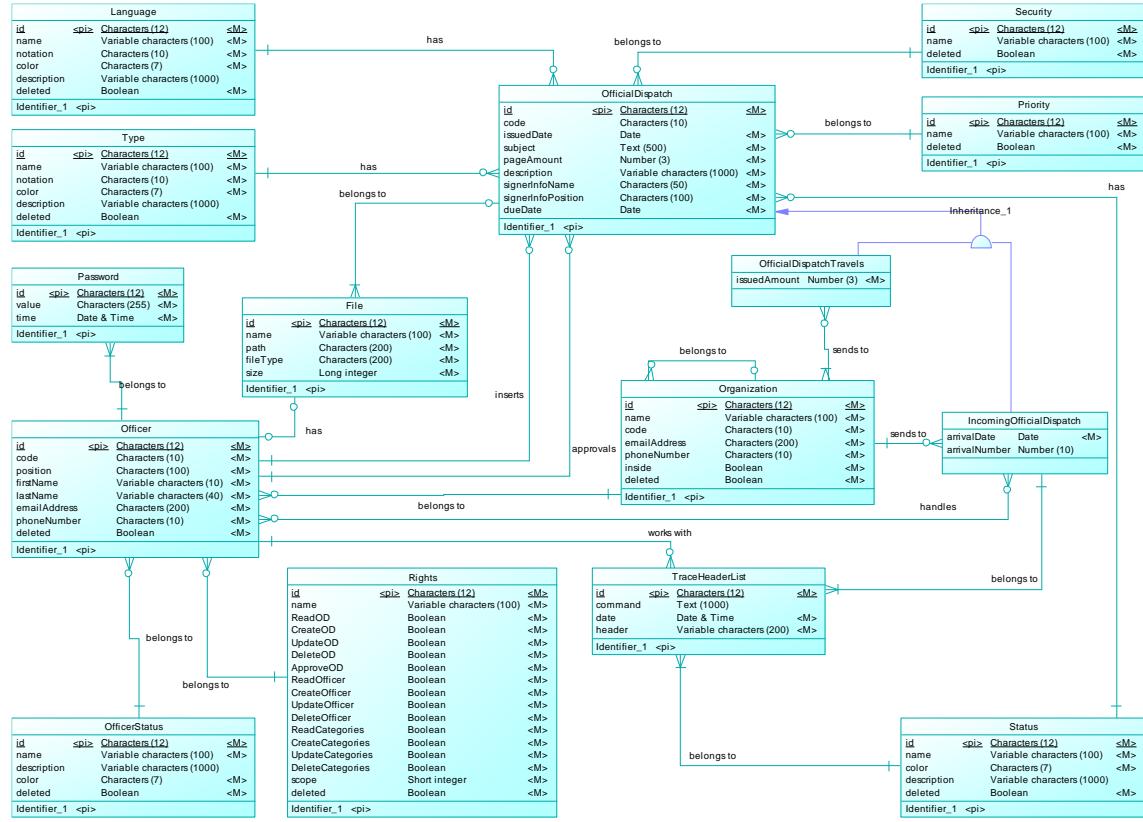
Lãnh đạo đơn vị sau khi đăng nhập sẽ có toàn bộ quyền của văn thư và các chức năng quản lý cán bộ, duyệt công văn đi, công văn đến, ban hành công văn đến và từ chối công văn đến.



Hình 2.15 Sơ đồ Use Case lãnh đạo đơn vị

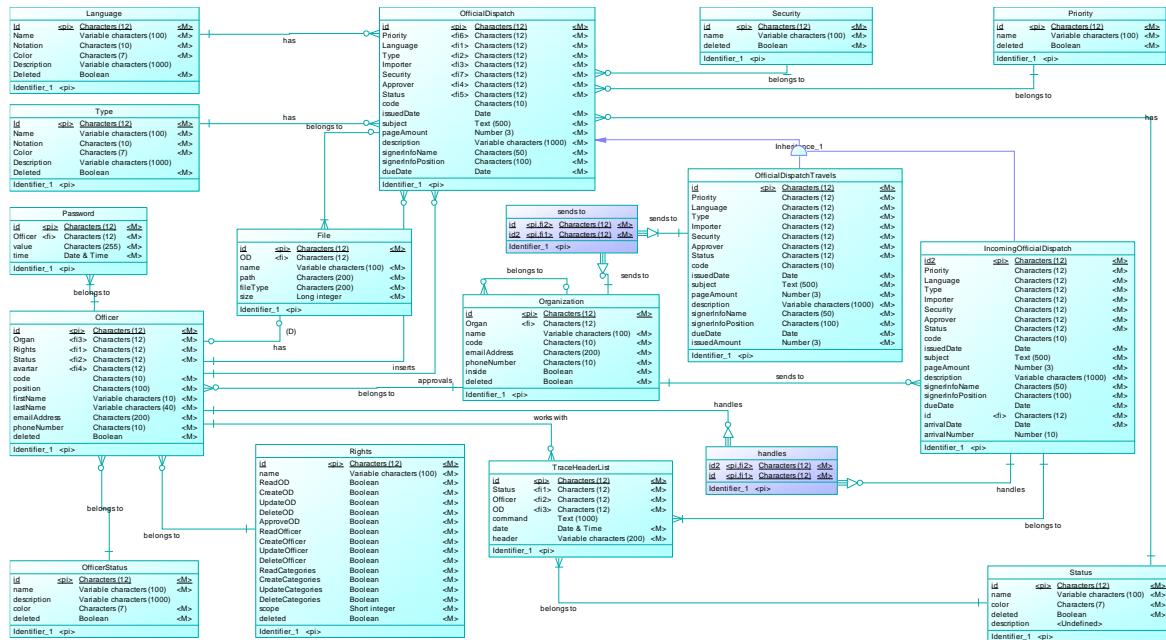
Ứng dụng optical character recognition vào hệ thống quản lý công văn

2.2.2. Sơ đồ ER



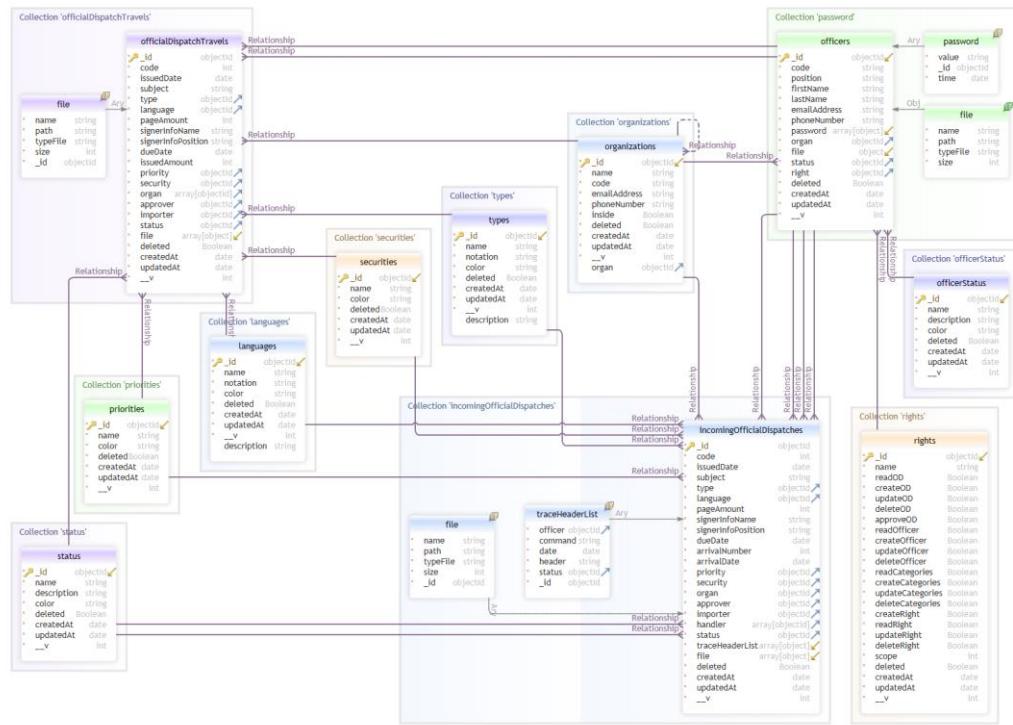
Hình 2.16 Sơ đồ ER

2.2.3. Sơ đồ LDM



Hình 2.17 Sơ đồ LDM

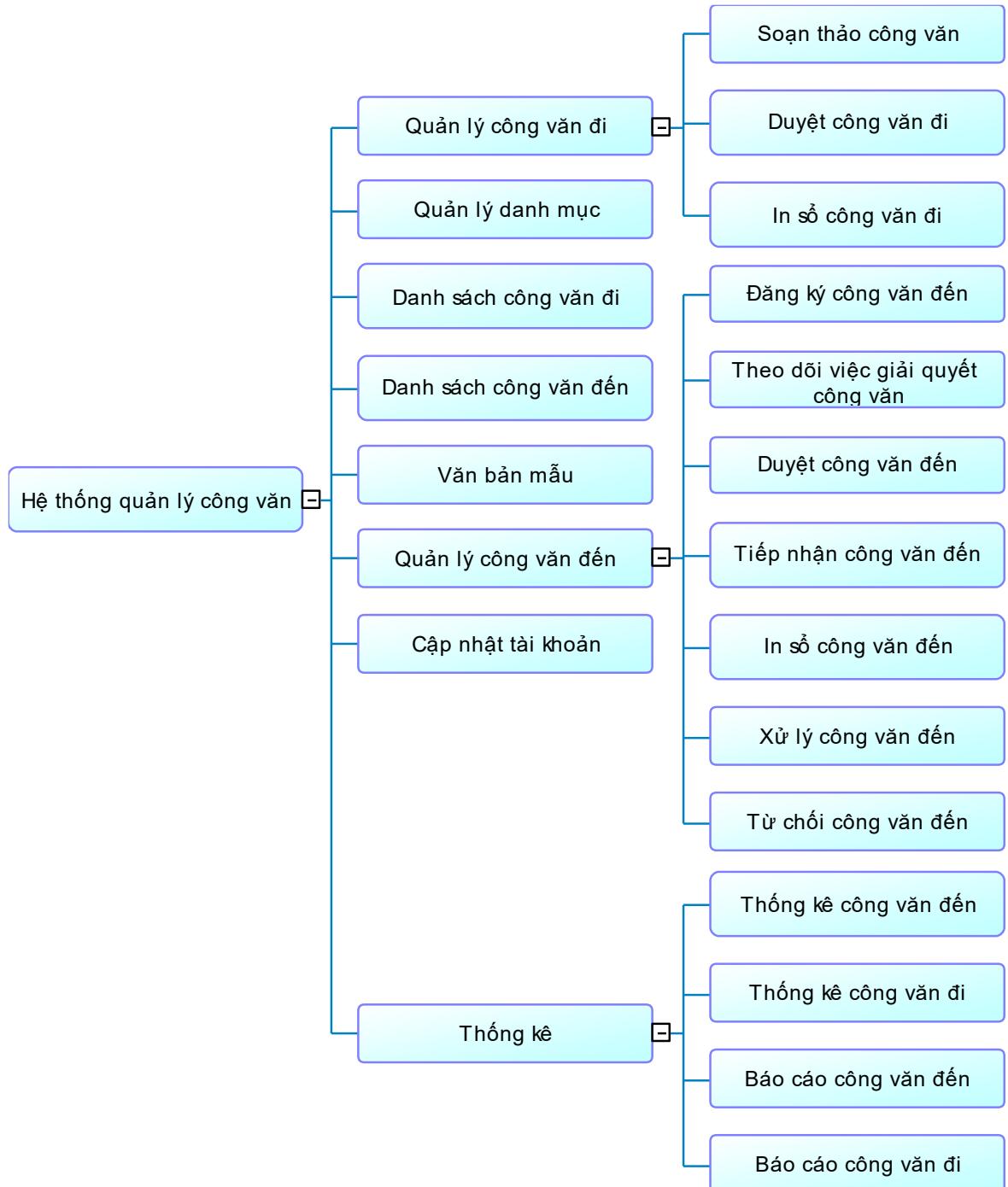
2.2.4. Sơ đồ LDM của MongoDB



Hình 2.18 Sơ đồ LDM của MongoDB

Ứng dụng optical character recognition vào hệ thống quản lý công văn

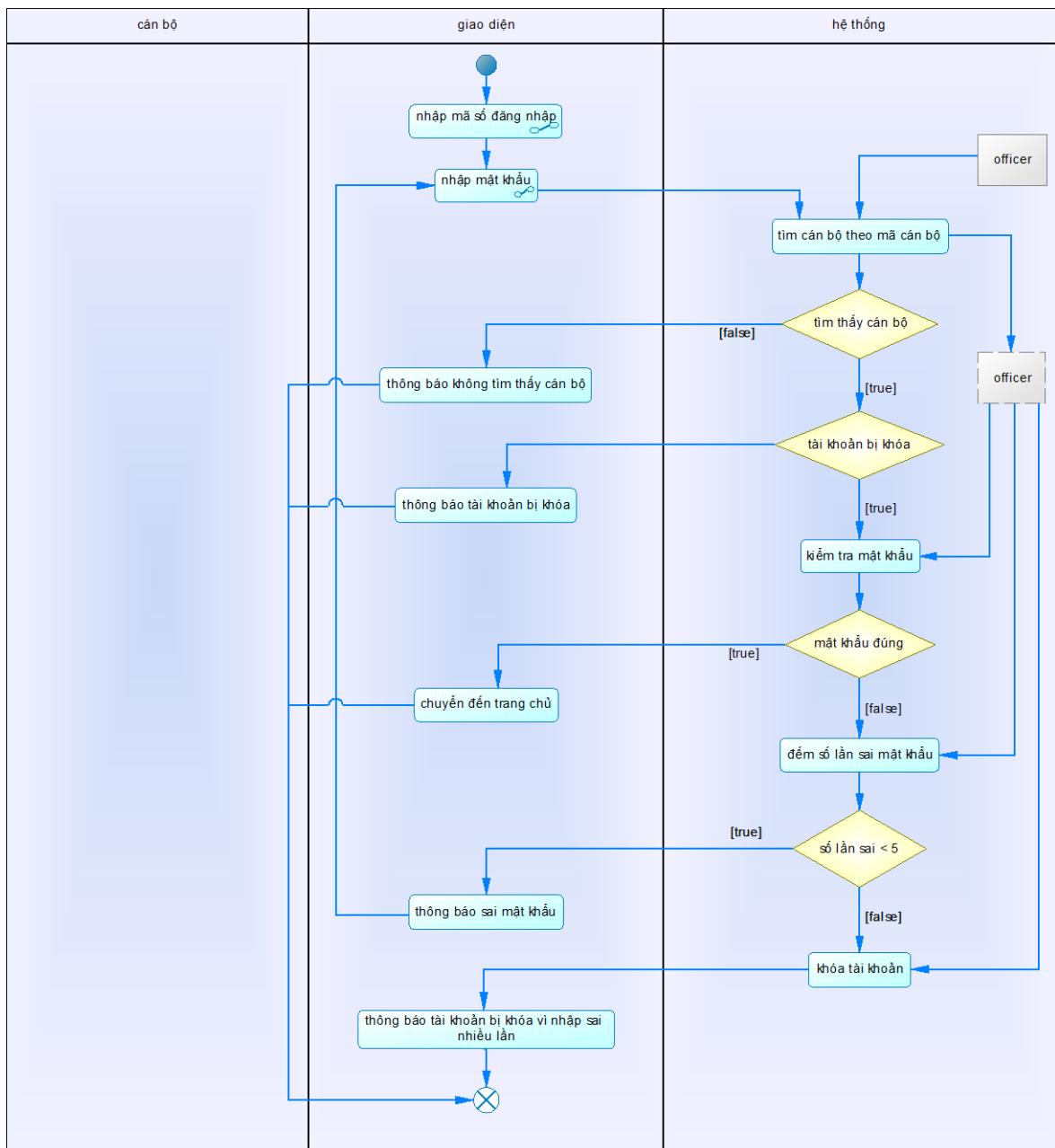
2.2.5. Sơ đồ phân rã chức năng



Hình 2.19 Sơ đồ phân rã chức năng

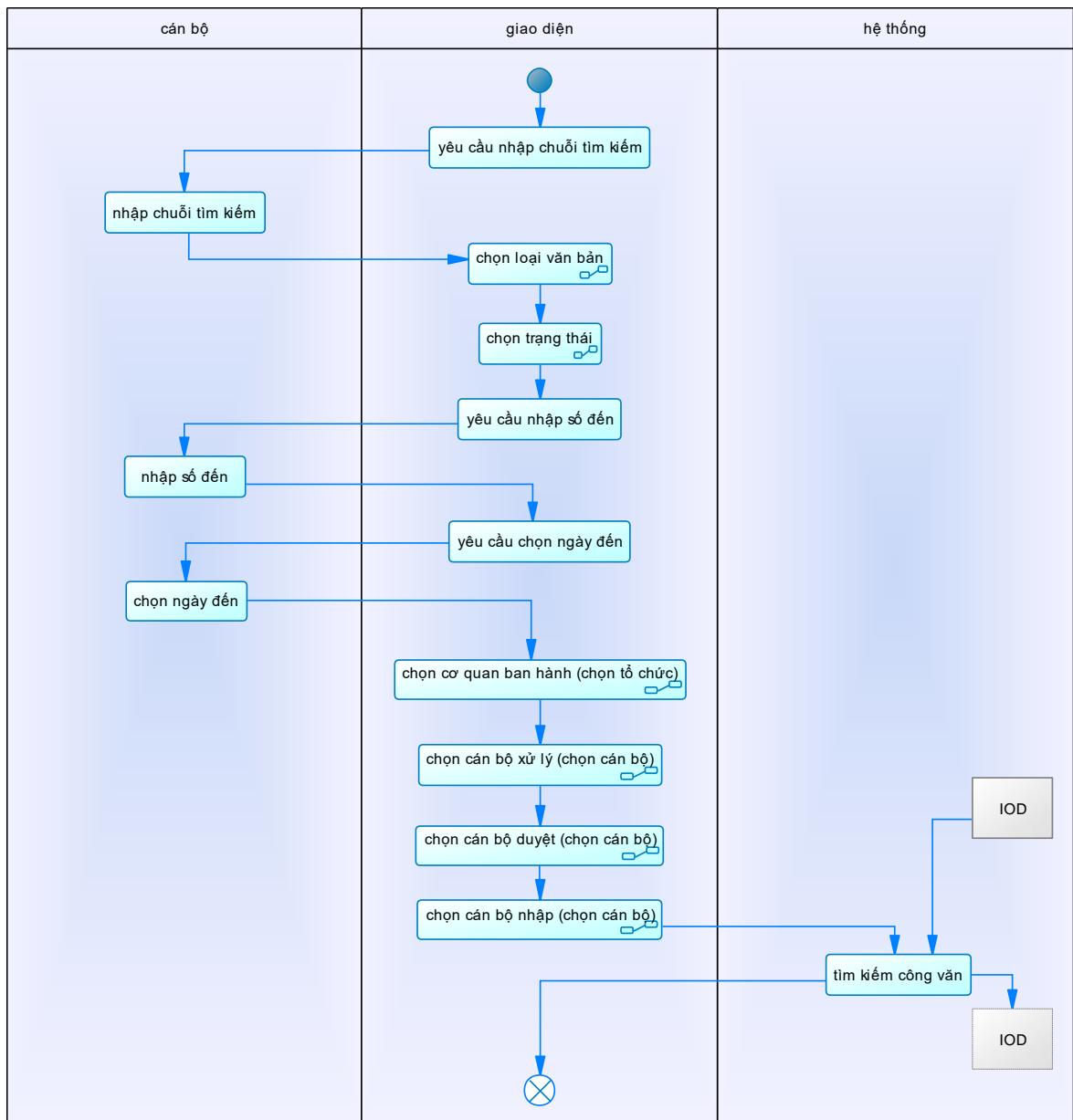
2.2.6. Sơ đồ hoạt động

2.2.6.1. Sơ đồ hoạt động chức năng đăng nhập



Hình 2.20 Sơ đồ hoạt động chức năng đăng nhập

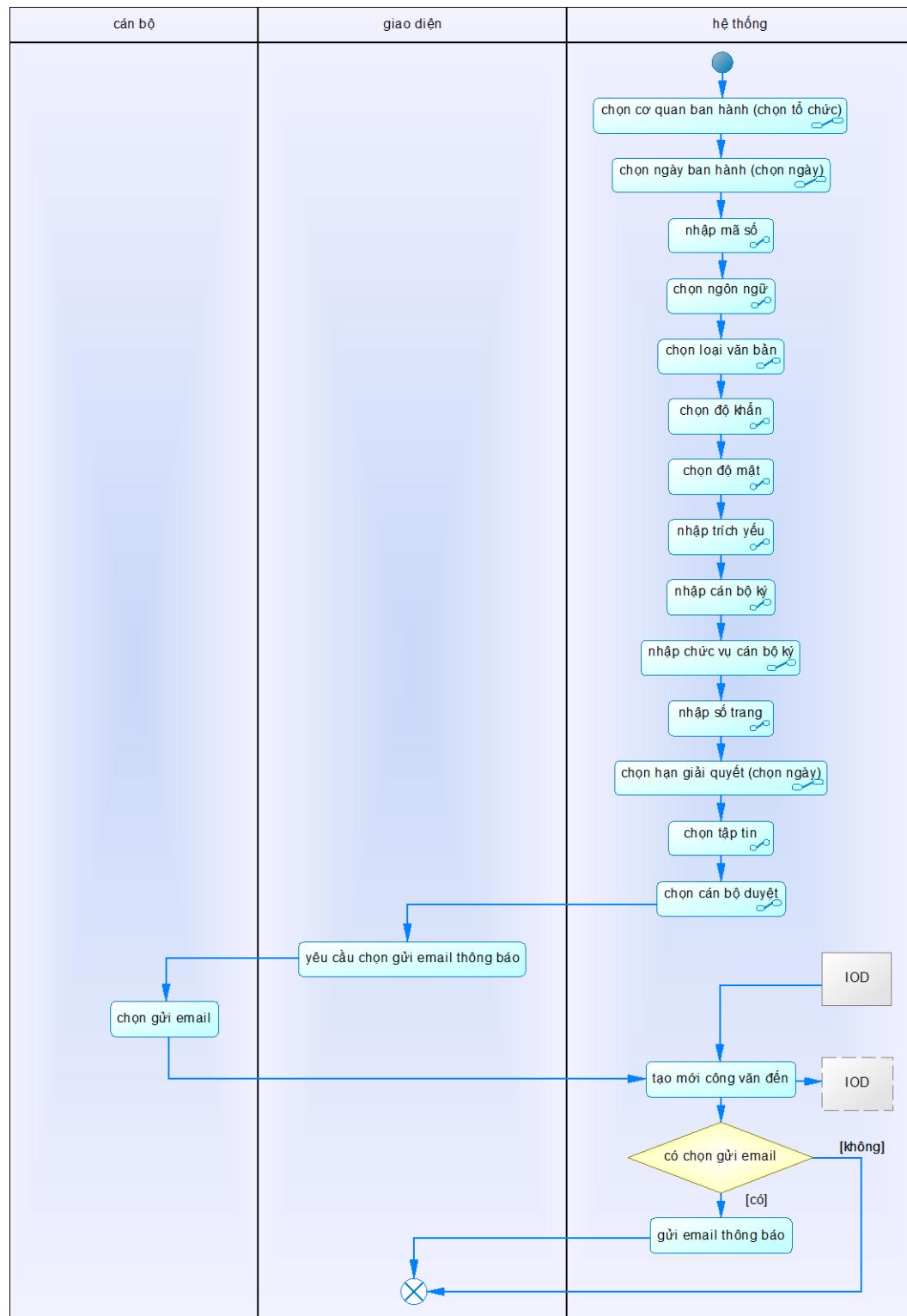
2.2.6.2. Sơ đồ hoạt động tìm công văn đến



Hình 2.21 Sơ đồ hoạt động tìm công văn đến

Ứng dụng optical character recognition vào hệ thống quản lý công văn

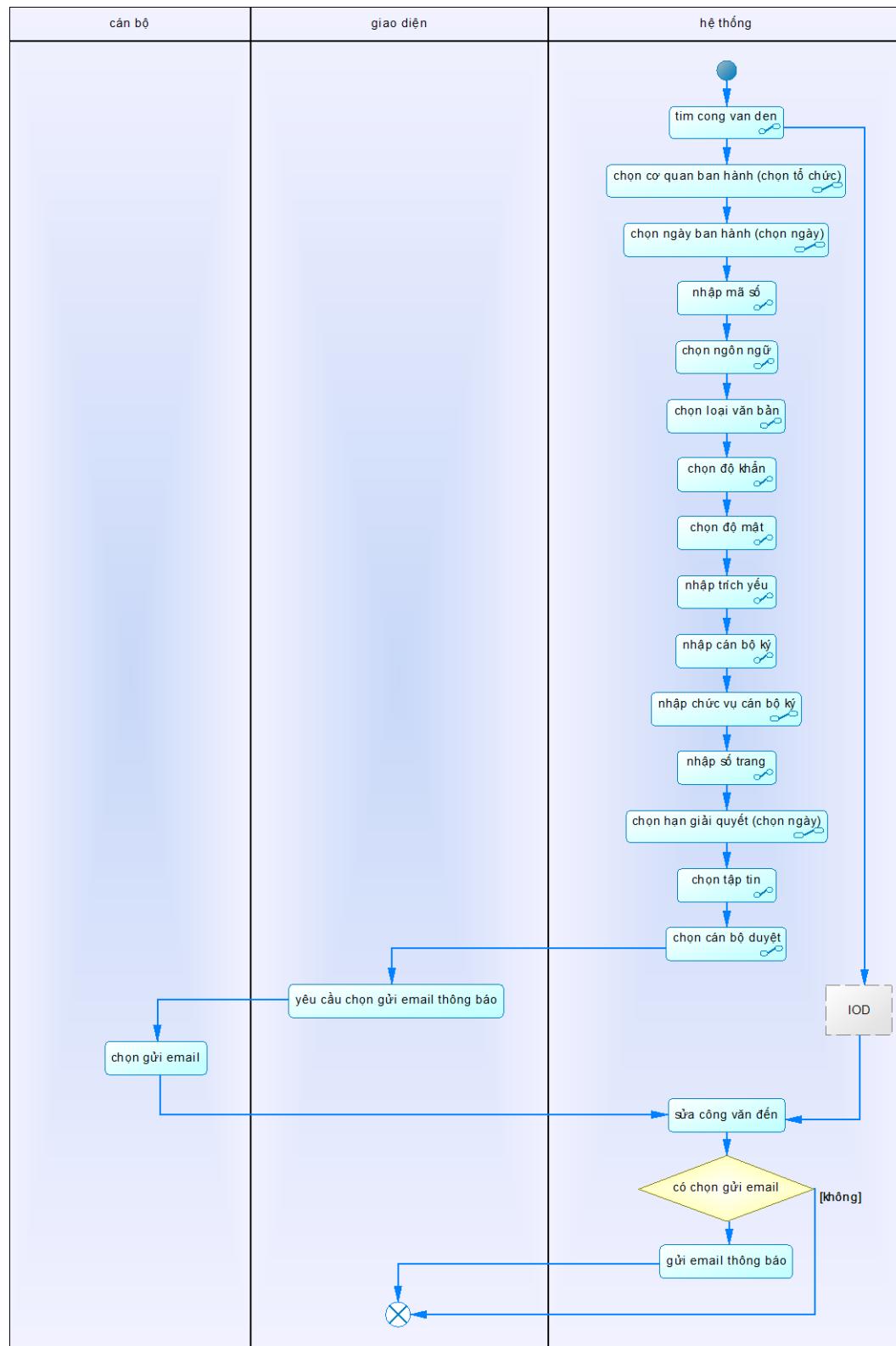
2.2.6.3. Sơ đồ hoạt động chức năng thêm công văn đến



Hình 2.22 Sơ đồ hoạt động thêm công văn đến

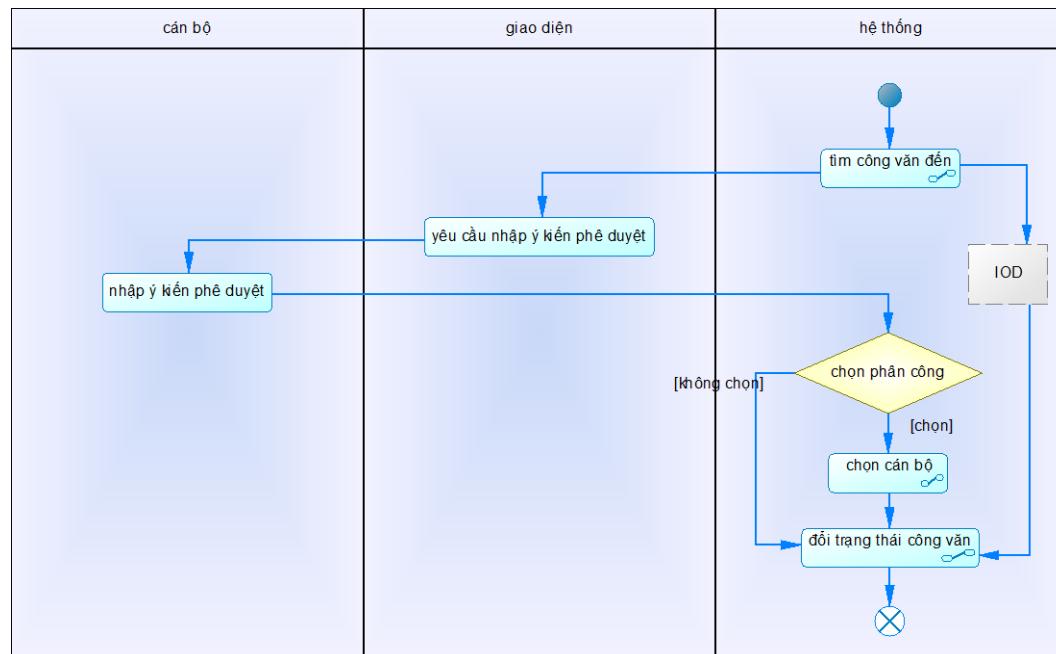
Ứng dụng optical character recognition vào hệ thống quản lý công văn

2.2.6.4. Sơ đồ hoạt động chức năng sửa công văn đến



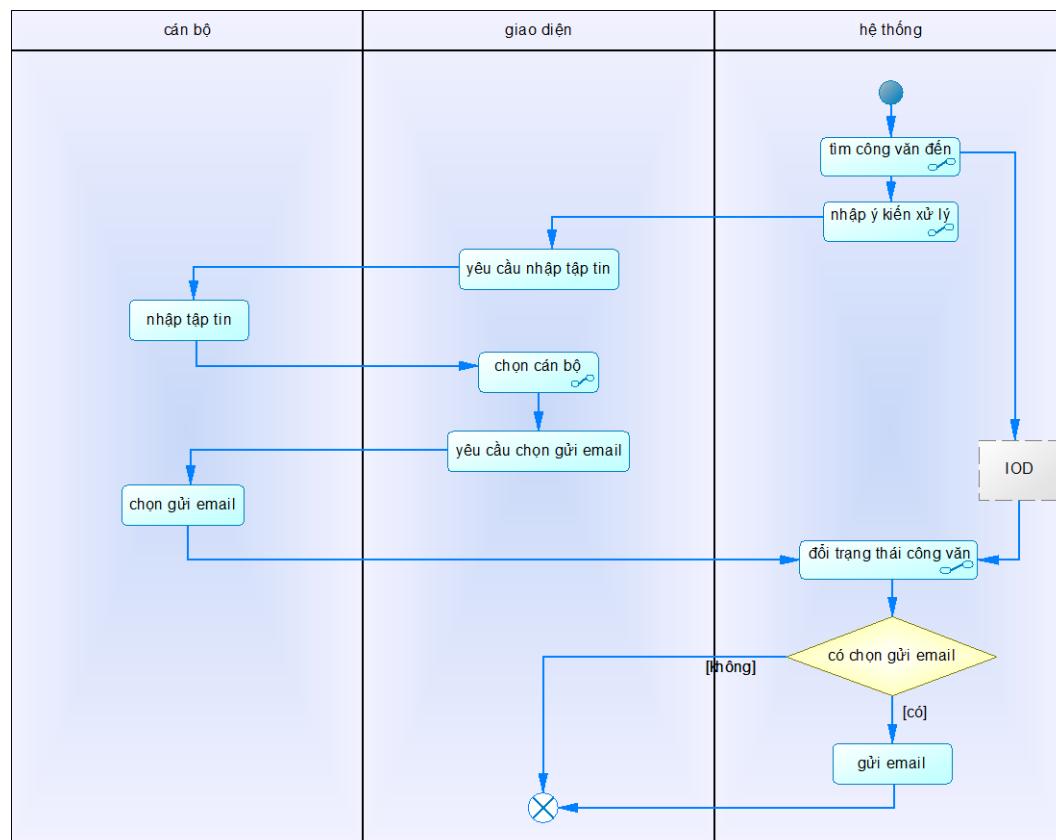
Hình 2.23 Sơ đồ hoạt động chức năng sửa công văn đến

2.2.6.5. Sơ đồ hoạt động chức năng phê duyệt công văn đến



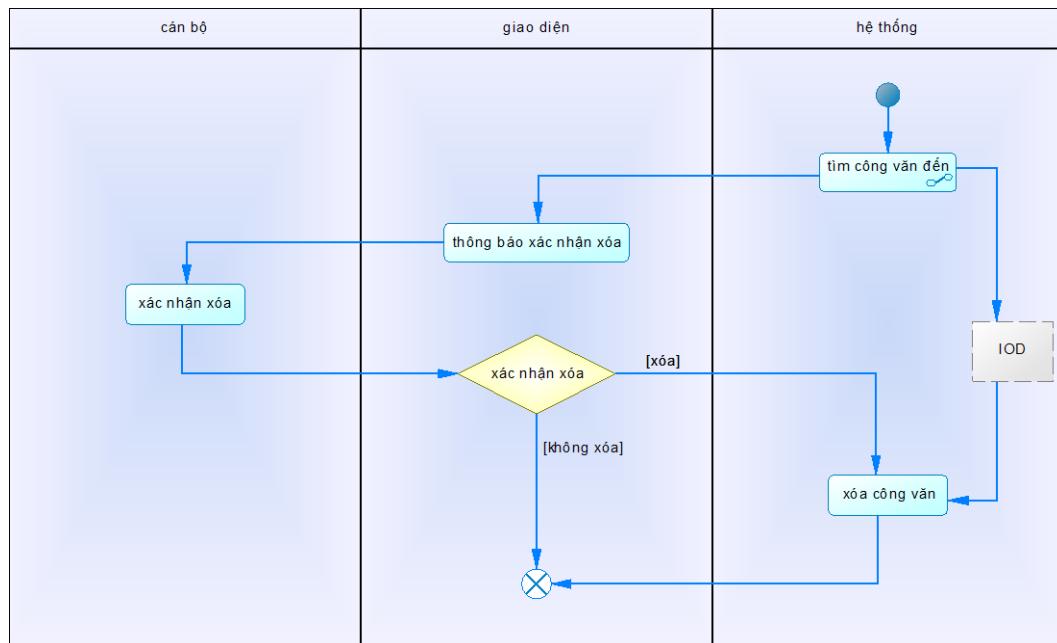
Hình 2.24 Sơ đồ hoạt động chức năng phê duyệt công văn đến

2.2.6.6. Sơ đồ hoạt động chức năng xử lý công văn đến



Hình 2.25 Sơ đồ hoạt động chức năng xử lý công văn đến

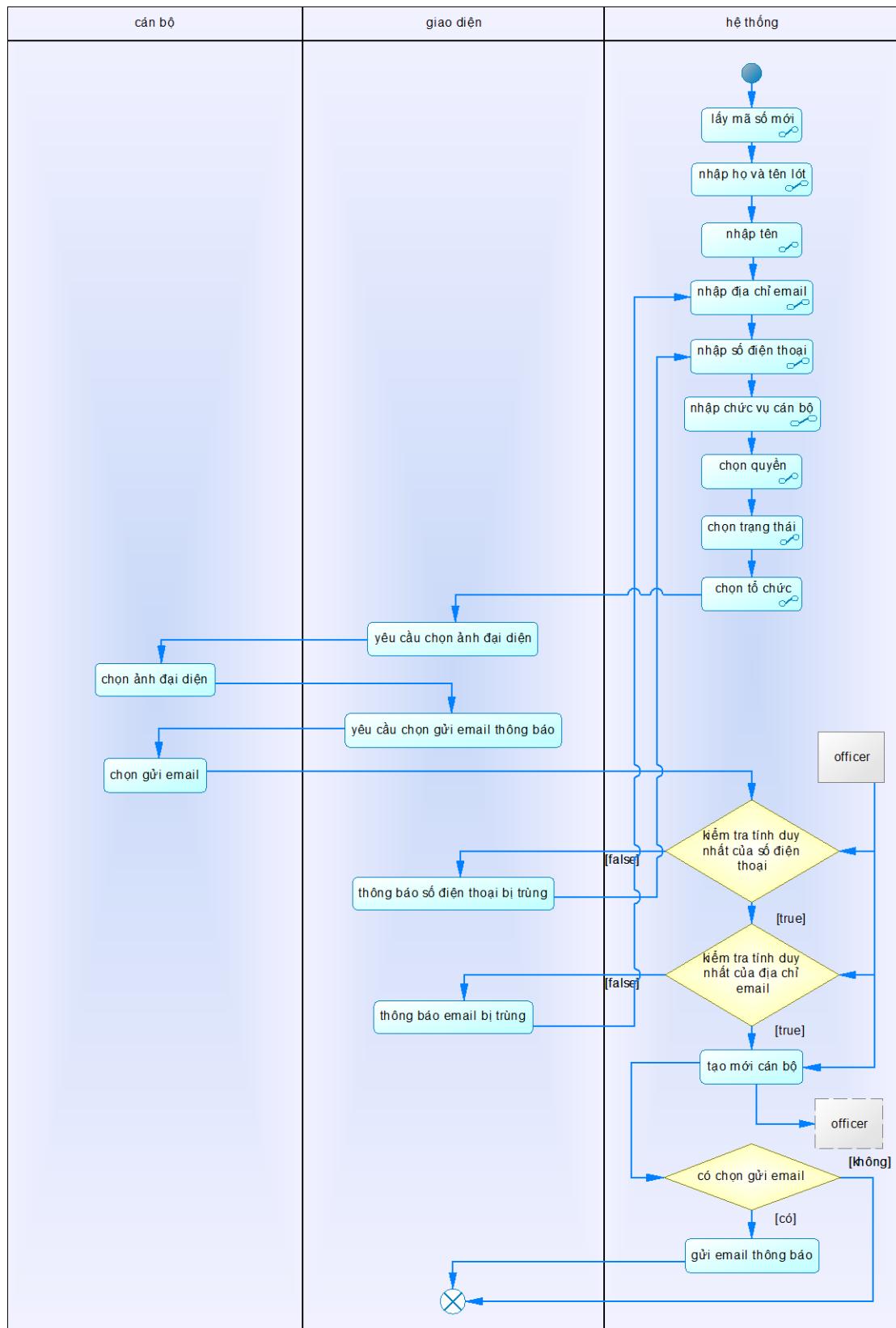
2.2.6.7. Sơ đồ hoạt động chức năng xóa công văn đến



Hình 2.26 Sơ đồ hoạt động chức năng xóa công văn đến

2.2.6.8. Sơ đồ hoạt động thông kê công văn đến

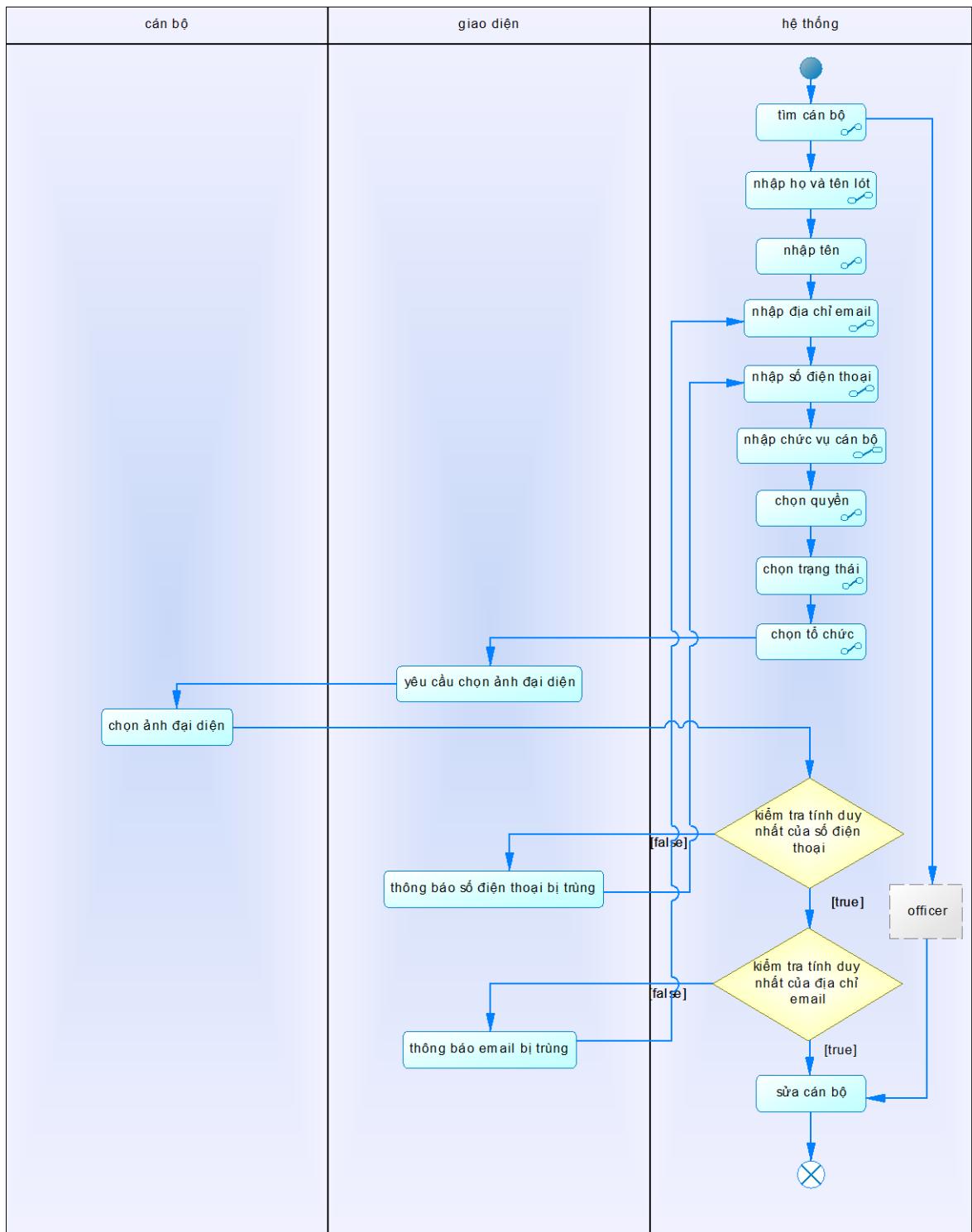
2.2.6.9. Sơ đồ hoạt động thêm cán bộ



Hình 2.27 Sơ đồ hoạt động thêm cán bộ

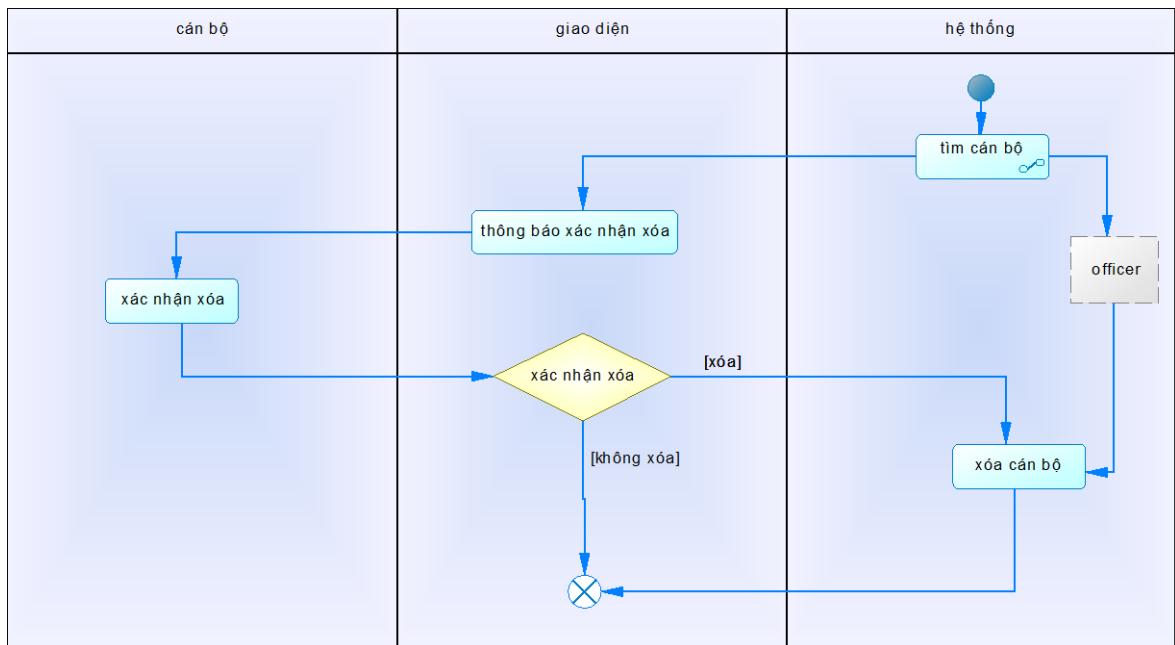
Ứng dụng optical character recognition vào hệ thống quản lý công văn

2.2.6.10. Sơ đồ hoạt động sửa cán bộ



Hình 2.28 Sơ đồ hoạt động sửa cán bộ

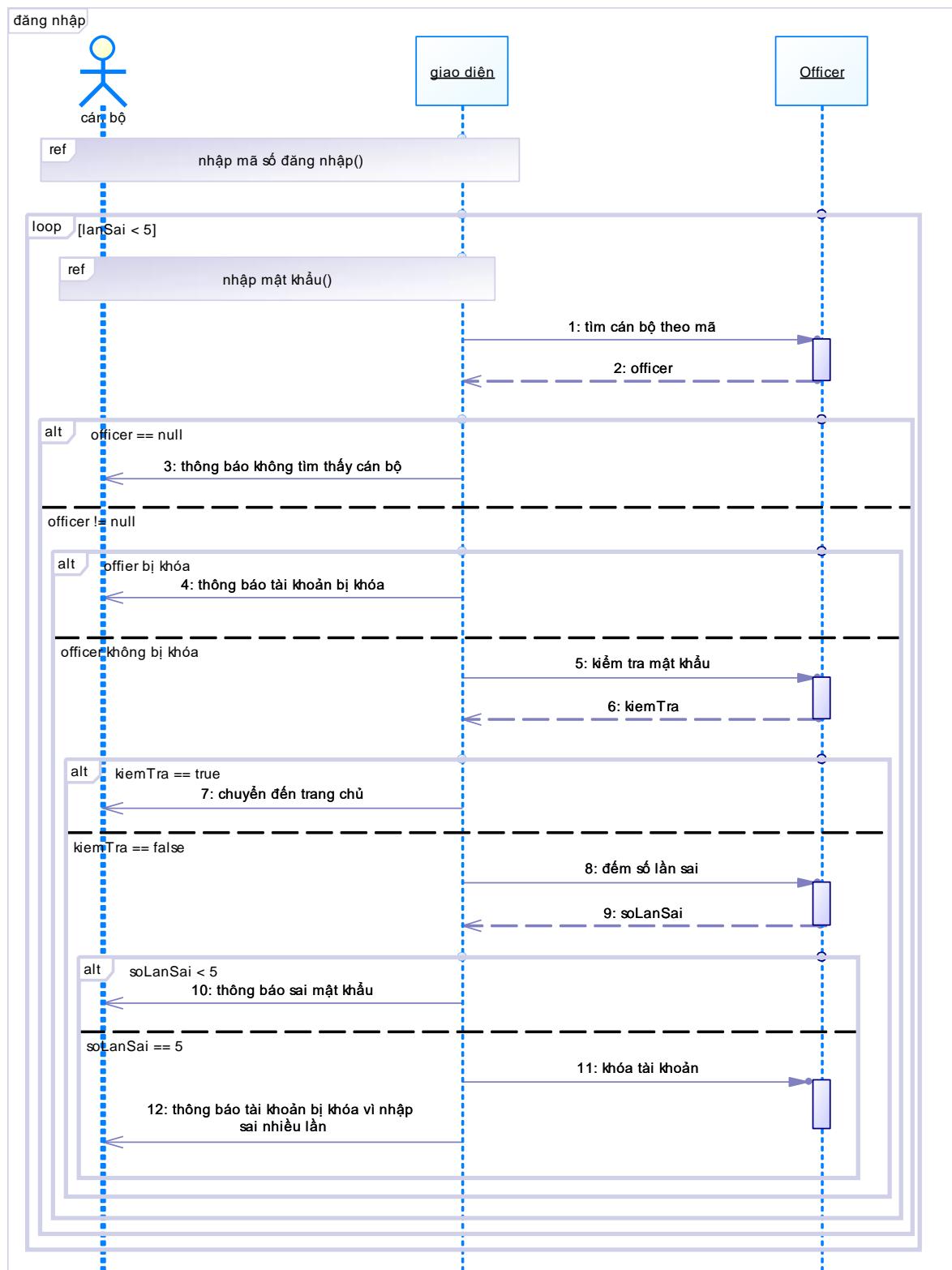
2.2.6.11. Sơ đồ hoạt động xóa cán bộ



Hình 2.29 Sơ đồ hoạt động xóa cán bộ

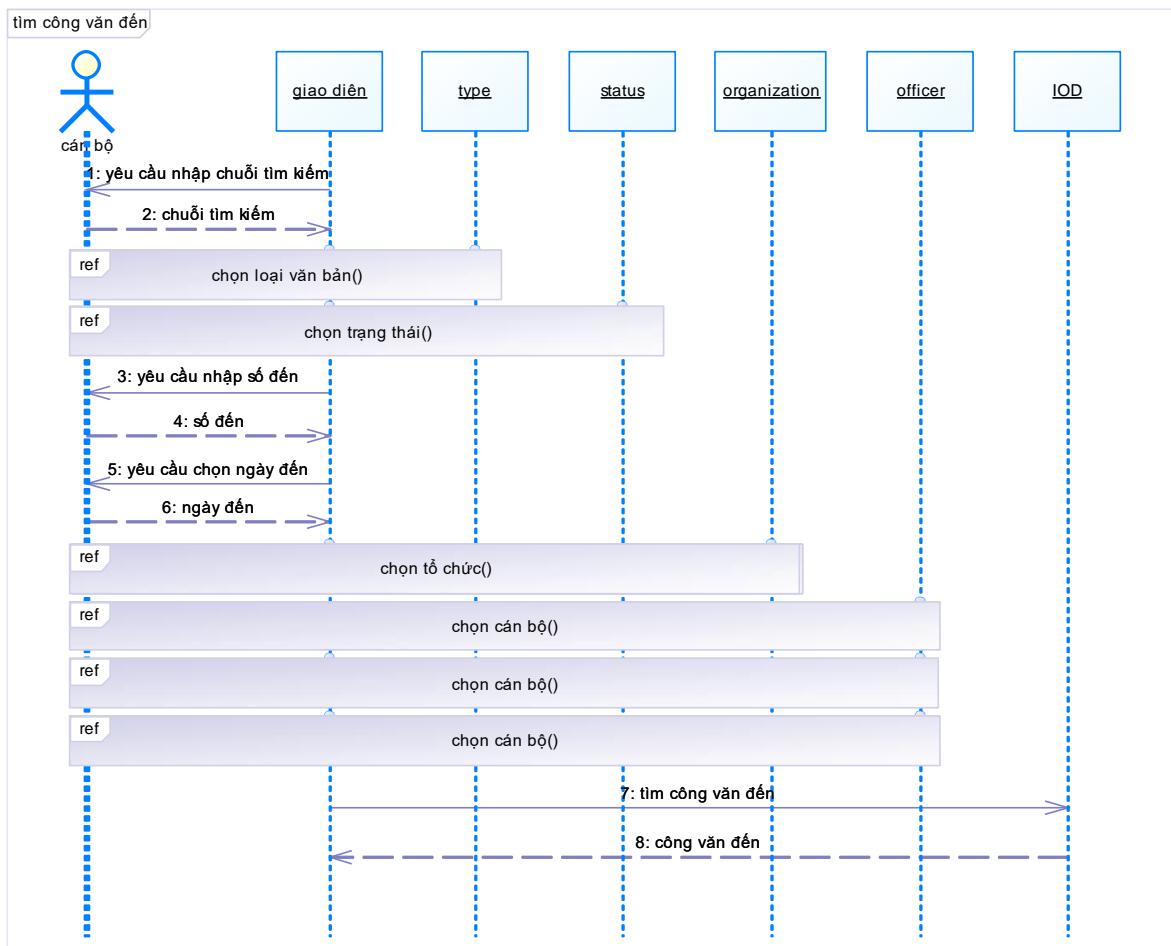
2.2.7. Sơ đồ tuần tự

2.2.7.1. Sơ đồ tuần tự chức năng đăng nhập



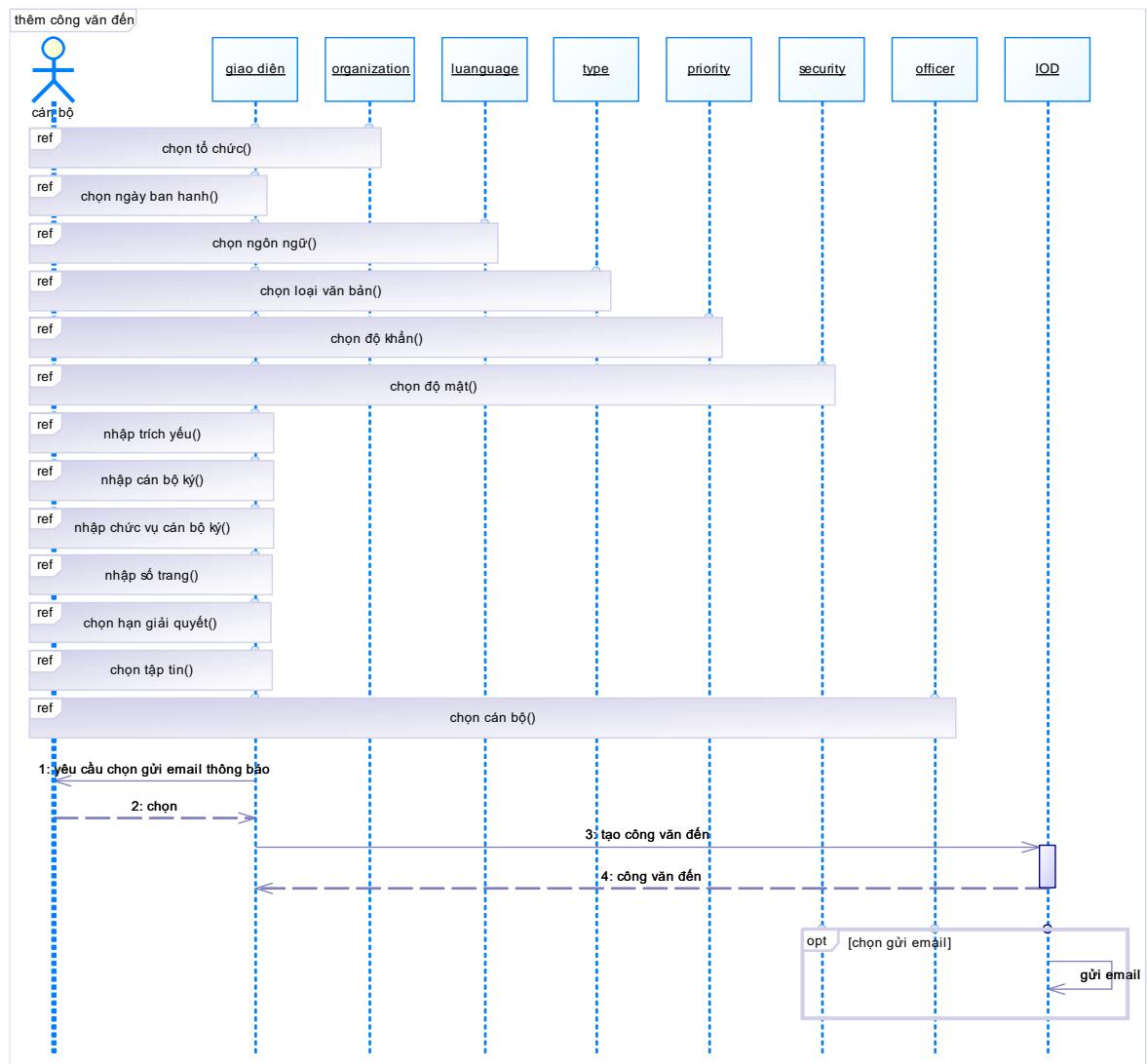
Hình 2.30 Sơ đồ tuần tự chức năng đăng nhập

2.2.7.2. Sơ đồ tuần tự chức năng tìm công văn đến



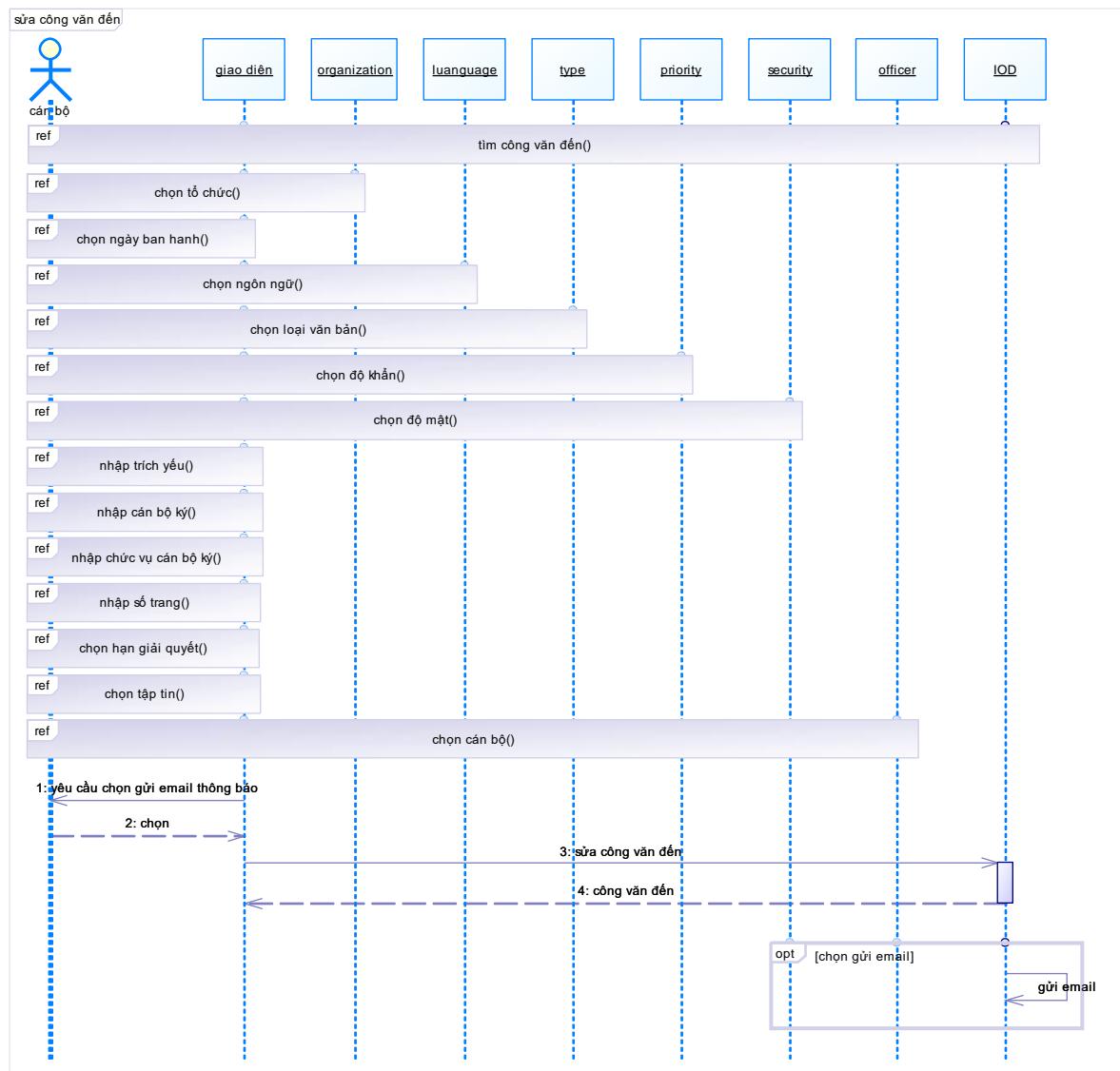
Hình 2.31 Sơ đồ tuần tự chức năng tìm công văn đến

2.2.7.3. Sơ đồ tuần tự chức năng thêm công văn đến



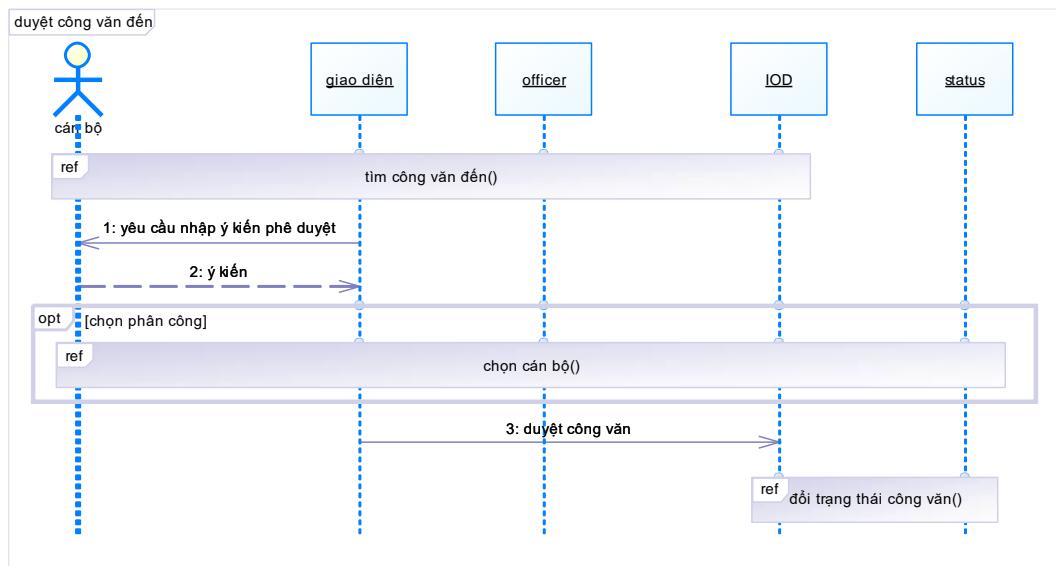
Hình 2.32 Sơ đồ tuần tự chức năng thêm công văn đến

2.2.7.4. Sơ đồ tuần tự chức năng sửa công văn đến



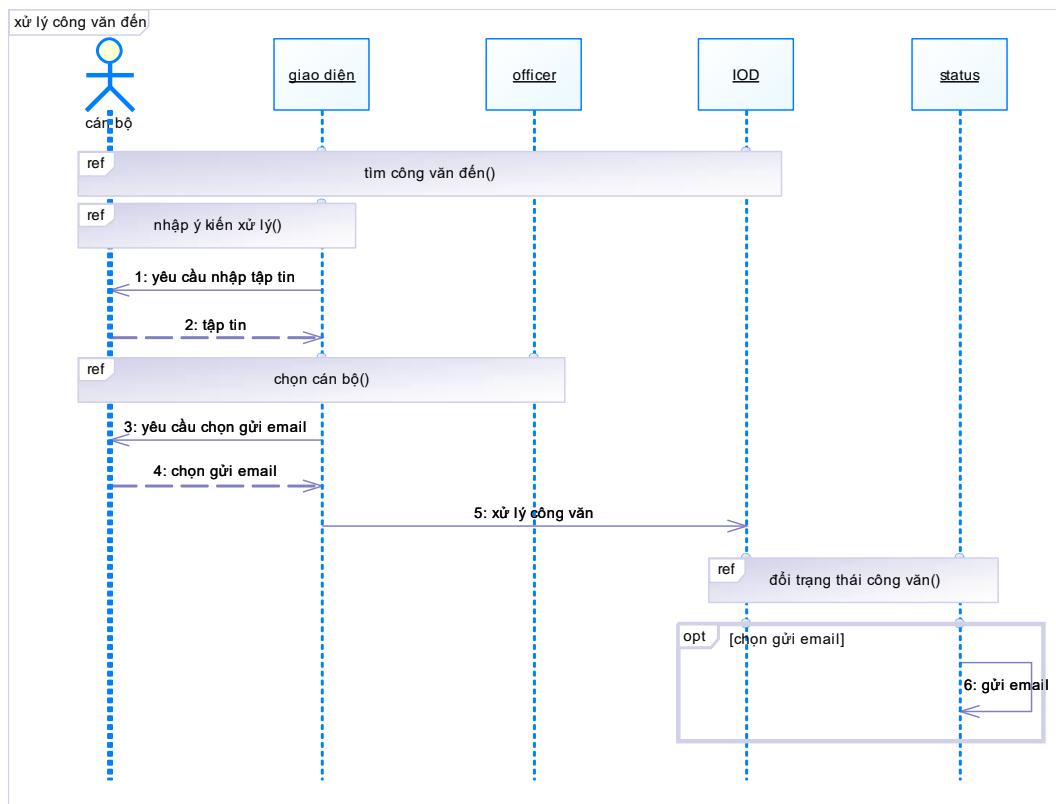
Hình 2.33 Sơ đồ tuần tự chức năng sửa công văn đến

2.2.7.5. Sơ đồ tuần tự chức năng phê duyệt công văn đến



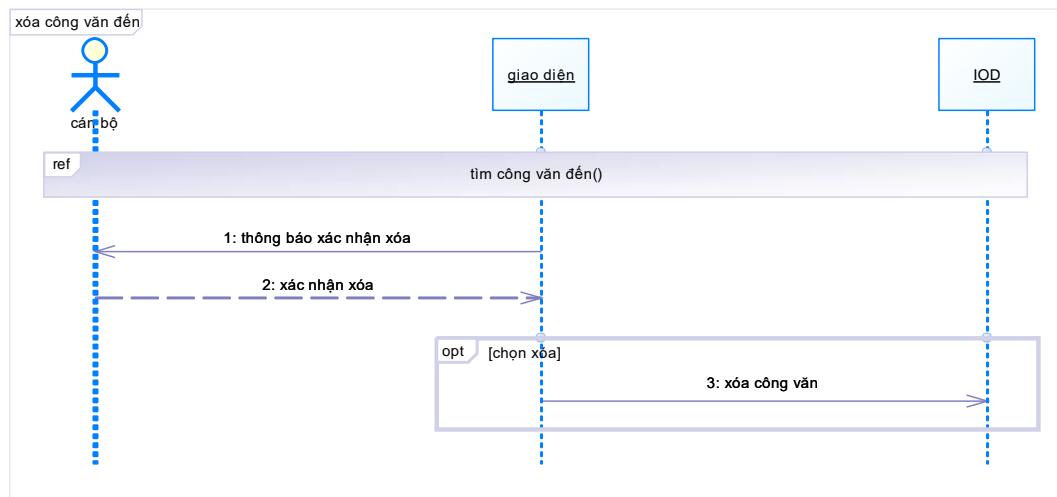
Hình 2.34 Sơ đồ tuần tự chức năng phê duyệt công văn đến

2.2.7.6. Sơ đồ tuần tự chức năng xử lý công văn đến



Hình 2.35 Sơ đồ tuần tự chức năng xử lý công văn đến

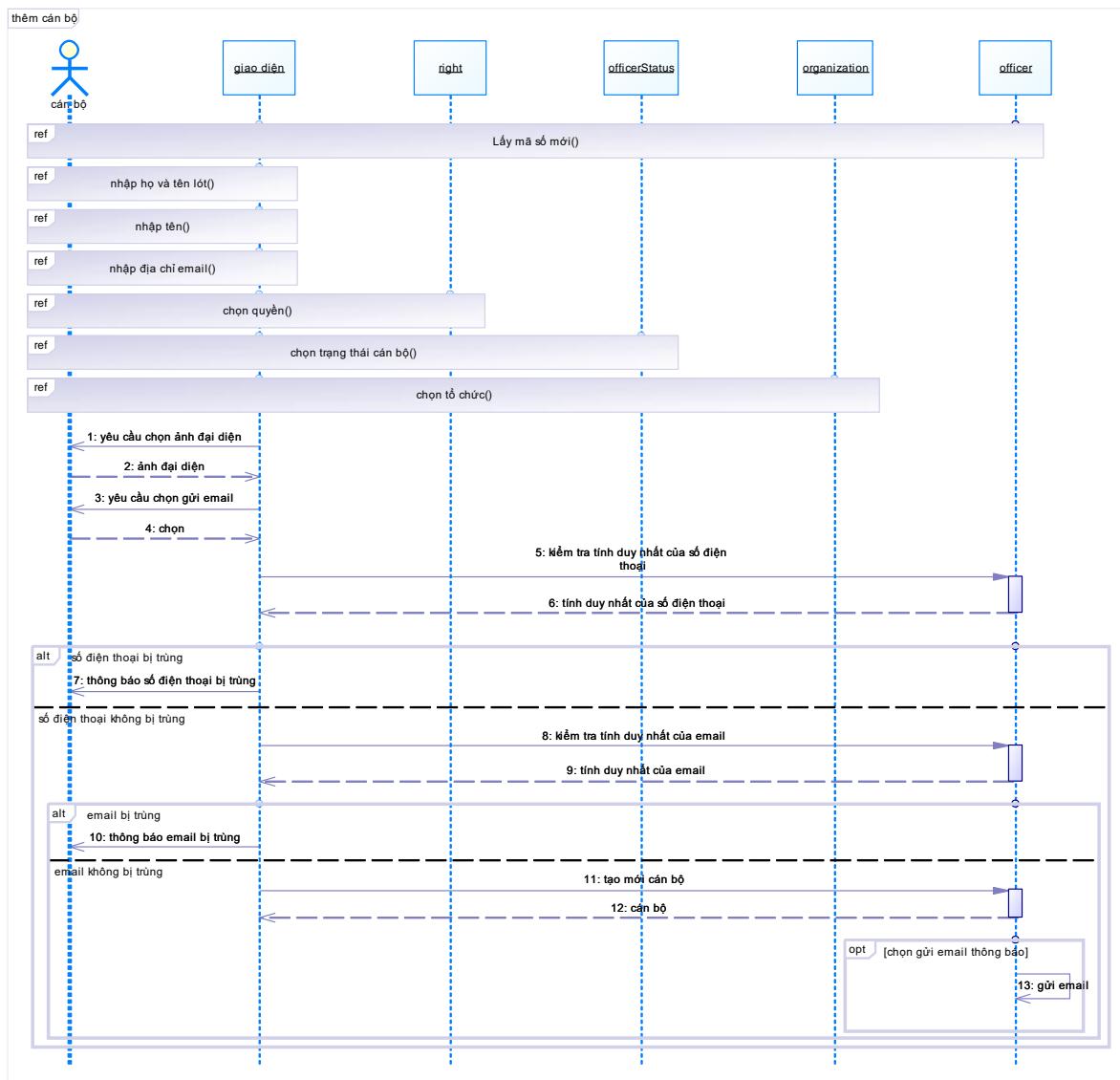
2.2.7.7. Sơ đồ tuần tự chức năng xóa công văn đến



Hình 2.36 Sơ đồ tuần tự chức năng xóa công văn đến

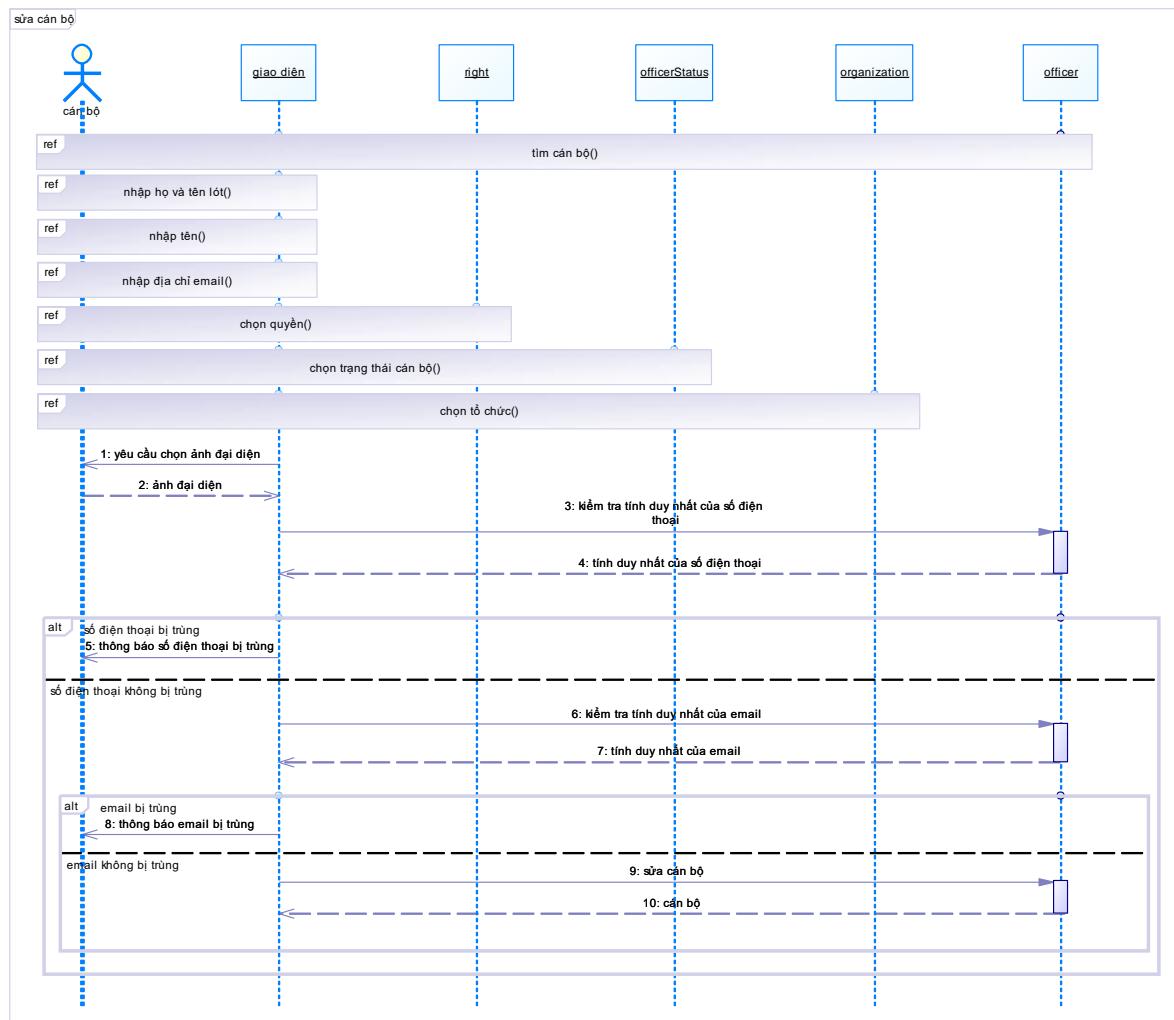
2.2.7.8. Sơ đồ tuần tự thông kê công văn đến

2.2.7.9. Sơ đồ tuần tự thêm cán bộ



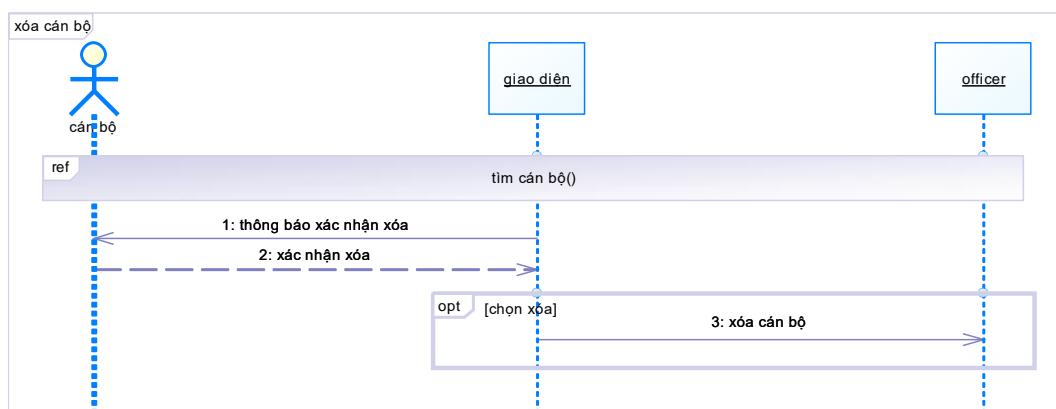
Hình 2.37 Sơ đồ tuần tự thêm cán bộ

2.2.7.10. Sơ đồ tuần tự sửa cán bộ



Hình 2.38 Sơ đồ tuần tự sửa cán bộ

2.2.7.11. Sơ đồ tuần tự xóa cán bộ

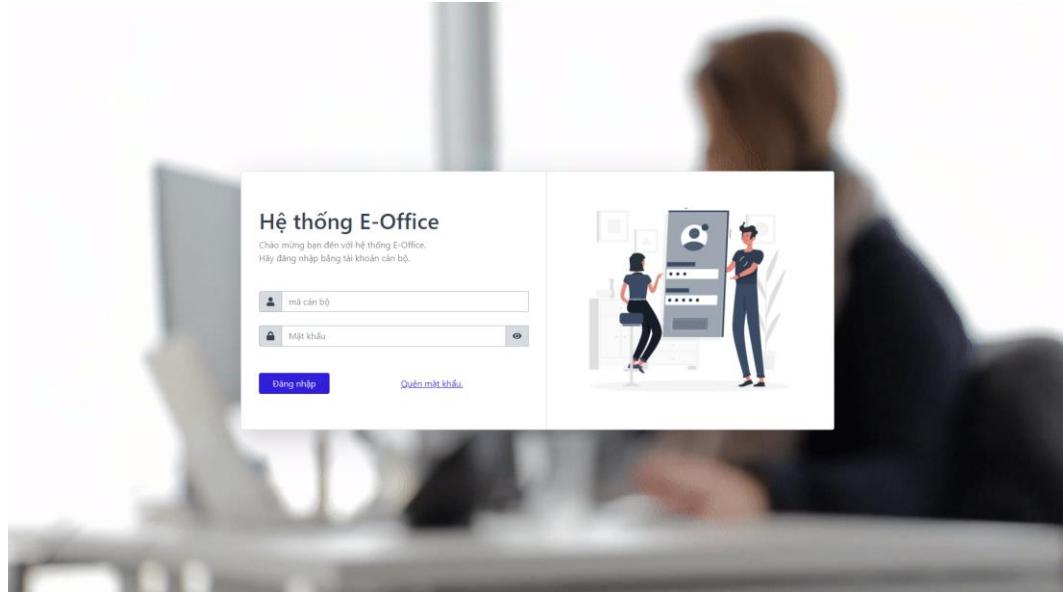


Hình 2.39 Sơ đồ tuần tự xóa cán bộ

2.3. GIAO DIỆN ỨNG DỤNG

2.3.1. Giao diện đăng nhập

Cán bộ đăng nhập vào hệ thống để thực hiện các chức năng.



Hình 2.40 Giao diện đăng nhập

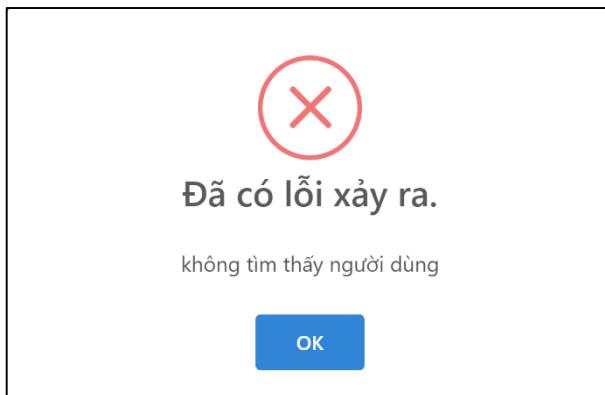
Khi cán bộ không nhập hoặc nhập không đầy đủ thông tin đăng nhập sẽ xuất hiện dòng thông báo như sau:

A screenshot of the E-Office login page with validation errors. The error messages are: "Bạn phải nhập mã số." next to the employee number field and "Bạn phải nhập mật khẩu." next to the password field. The rest of the page is identical to the one in Figure 2.40.

Hình 2.41 Thông báo chưa nhập đủ thông tin đăng nhập

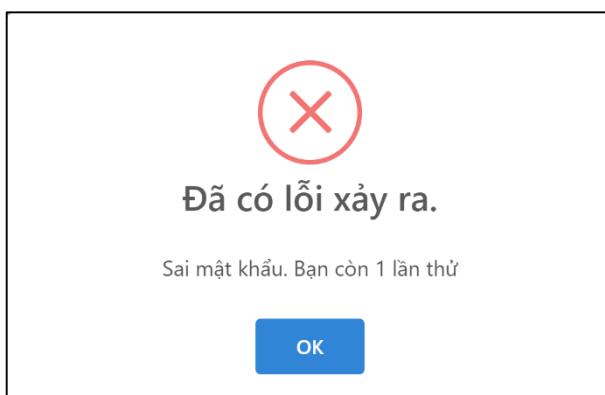
Ứng dụng optical character recognition vào hệ thống quản lý công văn

Khi nhập sai mã nhân viên sẽ xuất hiện thông báo **không tìm thấy người dùng**.



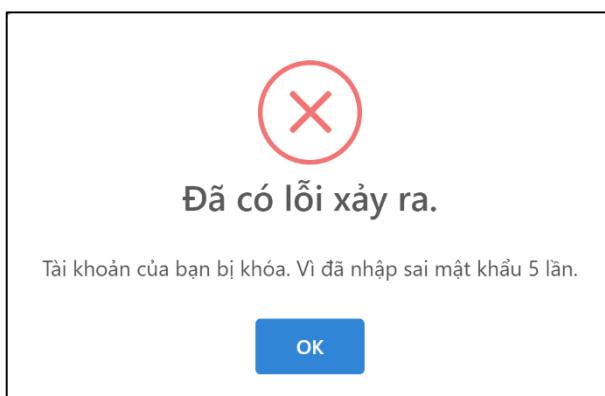
Hình 2.42 Thông báo không tìm thấy người dùng

Khi nhập sai mật khẩu sẽ hiện thông báo **sai mật khẩu**.



Hình 2.43 Thông báo sai mật khẩu

Hệ thống sẽ cảnh báo số lần đăng nhập còn lại nếu cán bộ đăng nhập sai 5 lần liên tiếp tài khoản sẽ bị khóa.



Hình 2.44 Thông báo tài khoản bị khóa vì đăng nhập sai nhiều lần

Ứng dụng optical character recognition vào hệ thống quản lý công văn

Trong trường hợp tài khoản bị khóa người dùng sẽ nhận được thông báo **tài khoản bị khóa**.



Hình 2.45 Thông báo tài khoản bị khóa

2.3.2. Giao diện xem danh sách công văn đến

A screenshot of a web-based document management system interface. The title bar says "Văn bản đến". The main area is a table titled "Danh sách văn bản đến" showing 10 entries. The columns are: #, Ngày đến, Ngày ph..., Số đ..., Mã số, Trich yếu, Cơ quan ban hành, Trạng thái, and Thảo tác. Each row has a "Chi tiết" link in the last column. The table includes filters and search functions at the top.

Hình 2.46 Giao diện danh sách công văn đến

Cán bộ có thể xem danh sách công văn đến khi truy cập vào giao diện danh sách công văn đến. Tùy theo quyền của mỗi cán bộ mà danh sách công văn hiển thị sẽ khác nhau. Tại giao diện này cán bộ có thể tìm kiếm công văn theo loại văn bản, trạng thái, số đến. Ngoài ra cán bộ cũng có thể tìm kiếm theo nhiều tiêu chí khác khi nhấn vào nút **Tìm kiếm nâng cao** Tìm kiếm nâng cao.

Ứng dụng optical character recognition vào hệ thống quản lý công văn

Hình 2.47 Giao diện tìm kiếm nâng cao

Từ giao diện xem danh sách cán bộ có thể thực hiện các thao tác khác là thêm mới Thêm mới, in danh sách In danh sách, xem chi tiết Chi tiết, chỉnh sửa Chỉnh sửa, xóa Xóa, phê duyệt Phê duyệt, xử lý Xử lý và ban hành Ban hành.

2.3.3. Xem chi tiết công văn đến

Thông tin văn bản

Ngày phát hành văn bản đến	21/11/2022	Mã số văn bản đến	300 (300/2022/BC-CUSC)
Ngày đến	29/11/2022	Số đến	4850
Loại	Báo cáo	Ngôn ngữ	Tiếng Pháp
Trich yếu văn bản đến	Hampden-Sydney College in Virginia, looked up one		
Số trang văn bản đến	1	Cán bộ ký văn bản đến	Võ Xuân Phong
Hạn giải quyết văn bản đến	08/12/2022	Chức vụ cán bộ ký văn bản đến	Cán bộ
Độ khẩn	Hàng đầu	Độ mật	Không
Cơ quan ban hành	Trung tâm Công nghệ Phần mềm (CUSC)		

Nội dung toàn văn

test.pdf 9.77 KB	NhatKý.docx 9.77 KB	typeFile.csv 9.77 KB	từ.xlsx 9.77 KB
Tải tập tin			
Xem trước	Xem trước	Xem trước	Xem trước
New Project.png 9.77 KB	New Project.svg 9.77 KB		
Tải tập tin	Tải tập tin		
Xem trước	Xem trước		

Trạng thái xử lý

Mô tả văn bản đến	
Trạng thái văn bản đến	Đã báo cáo

Thông tin khác

Cán bộ duyệt văn bản đến	000011 Trần Thảo Hằng (Cán bộ)	Cán bộ nhập văn bản đến	000036 Dương Thúy Mai (Cán bộ)
Cán bộ xử lý	000038 Nguyễn Ái Liên (Cán bộ)		
	000007 Huỳnh Hiếu Trang (Cán bộ)		
	000012 Hồ Khải Bình (Cán bộ)		
ID	6388192540a1352cbda1852d	Phiên bản	0
Thời điểm khởi tạo	10:02:06, Thứ Năm, 01/12/2022	Thời điểm cập nhật	10:02:06, Thứ Năm, 01/12/2022

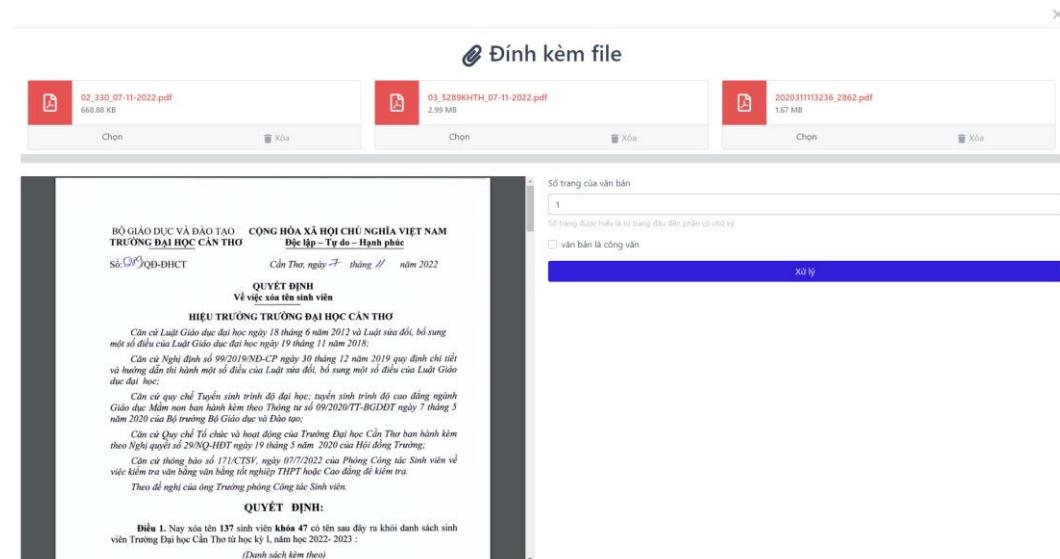
Hình 2.48 Giao diện chi tiết công văn đến

Ứng dụng optical character recognition vào hệ thống quản lý công văn

Tùy giao diện danh sách công văn đến chọn để chuyển qua giao diện thông tin chi tiết. Ở đây cán bộ sẽ được xem thông tin chi tiết công văn và thực hiện các thao tác: phê duyệt, triển khai, xử lý hay hủy duyệt.

2.3.4. Giao diện thêm công văn đến

Để thêm một công văn đến văn thư vào chức năng thêm mới trên giao diện danh sách công văn đến. Trước khi vào giao diện nhập liệu hệ thống sẽ hiện giao diện nhập file đính kèm. Tại đây chúng ta có thể chọn file công văn và chọn xử lý hoặc bỏ qua bước này bằng cách chọn bên góc trên bên phải màn hình.



Hình 2.49 Giao diện nhập file đính kèm

Khi chọn chức năng xử lý ta cần cho hệ thống biết văn bản có bao nhiêu trang (số trang được hiểu là từ trang đầu đến phần có chữ ký) và văn bản có phải là công văn hay không. Sau đó ấn vào nút **Xử lý** .

Sau khi thực hiện xử lý hệ thống sẽ hiện giao diện kết quả. Tại đây văn thư sẽ chọn những trường dữ liệu nhận dạng đúng.

Ứng dụng optical character recognition vào hệ thống quản lý công văn

Kết quả trích xuất						
Thành phần	Hình ảnh	Phân nhận dạng được	Dự đoán	Ghi chú	Chọn	
Mã số	Số:330/TB-CTSV	Số:330/TB-CTSV	370		<input checked="" type="checkbox"/>	
Ngày phát hành	năm 2022	năm 2022	01/01/2022		<input checked="" type="checkbox"/>	
Cần Thơ, ngày 7 tháng 11	Cần Thơ, ngày * tháng *					
THÔNG BÁO Về việc đánh giá điểm rèn luyện Học kỳ 1, 2022-2023	. THÔNG BÁO Về việc đánh giá điểm rèn luyện Học kỳ 1, 2022-2023	Về việc đánh giá điểm rèn luyện Học kỳ 1, 2022-2023	Dữ liệu nhận dạng trên thân văn bản	<input checked="" type="checkbox"/>		
Trích yếu	Số:330/TB-CTSV	Số:330/TB-CTSV	Dữ liệu nhận dạng từ vị trí mã số	<input type="checkbox"/>		
Loại	THÔNG BÁO Về việc đánh giá điểm rèn luyện Học kỳ 1, 2022-2023	. THÔNG BÁO Về việc đánh giá điểm rèn luyện Học kỳ 1, 2022-2023	Thông báo	Dữ liệu nhận dạng trên thân văn bản	<input type="checkbox"/>	
Tổ chức cấp trên	Số:330/TB-CTSV	Số:330/TB-CTSV	Thông báo	Dữ liệu nhận dạng từ mã số	<input checked="" type="checkbox"/>	
TRƯỜNG ĐẠI HỌC CẦN THƠ PHÒNG CÔNG TÁC SINH VIÊN	ngƯỜNG ĐẠI HỌC CẦN THƠ PHÒNG CÔNG TÁC SINH VIÊN	Phòng Công tác Sinh viên	Dữ liệu nhận dạng trên thân văn bản	<input type="checkbox"/>		
Số:330/TB-CTSV	Số:330/TB-CTSV	Phòng Công tác Sinh viên	Dữ liệu nhận dạng từ mã số	<input checked="" type="checkbox"/>		
TRƯỜNG PHÒNG						

Hình 2.50 Giao diện kết quả xử lý công văn

Sau bước xử lý văn thư sẽ nhập lại các thông tin còn thiếu vào các trường thông tin. Các trường có ký hiệu là các trường bắt buộc nhập.

The screenshot shows the software interface for document processing. On the left, there's a sidebar with icons for file operations like Open, Save, Print, etc. The main area has two parts: a preview window on the left and an input form on the right.

Preview Window Content:

- Header: TRƯỜNG ĐẠI HỌC CẦN THƠ
PHÒNG CÔNG TÁC SINH VIÊN
- Header: CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM
Độc lập - Tự do - Hạnh phúc
- Date: Cần Thơ, ngày 7 tháng 11 năm 2022
- Section: THÔNG BÁO
Về việc đánh giá điểm rèn luyện Học kỳ 1, 2022-2023
- Text: Kính gửi: Thủ trưởng đơn vị
Thực hiện kế hoạch đánh giá điểm rèn luyện học kỳ 1, 2022-2023; Phòng Công tác Sinh viên xin thông báo đến các đơn vị quản lý đào tạo triển khai thực hiện:
1. Sinh viên phải thực hiện kiểm tra trực tuyến "Sinh hoạt công dân sinh viên".
Thực hiện ý kiến chỉ đạo của Ban Giám hiệu trong công tác tuyên truyền, sinh hoạt công dân sinh viên (SV). Trong đó thông báo về tổ chức kiểm tra trực tiếp trên website Phòng Công tác Sinh viên.
- Đăng dẫn tài liệu tham khảo: <https://doa.ctu.edu.vn/tin-an/3/379-mot-so-tai-lieu-van-ban-co-lien-quan-1.html>
- Đường dẫn kiểm tra: <https://elearning.ctu.edu.vn/ctu/ctu-kieu-huoc-hoc.php?id=1158>
Để nộp hồ sơ, sinh viên cần đăng nhập vào Kế toán thanh toán số sinh viên trực tiếp đến kết quả đánh giá rèn luyện học kỳ 1, năm 2022-2023, chi tiết xem tại <https://doa.ctu.edu.vn/tin-an/3/61-tai-lieu-tham-khoa-sinh-hoat-dau-nam-dau-hoc-hoc-2022-2023-gioi-dinh-kiem-tra.html>
Trường hợp không đạt, SV sẽ bị trừ 03 điểm vào nội dung được nêu tại Điều 5, Mục 1, điều 4, do Phòng CTV cấp nhận trong Hướng dẫn đánh giá kết quả tín luận.
L. Các mức thời gian thực hiện:
TT Nộp đúng công việc Thời gian thực hiện
1 Các đơn vị nhận dữ liệu tham gia hoạt động 04/11 → 15/12/2022
2 Thời gian tổng hợp dữ liệu của Khoa 16/12 → 18/12/2022
3 Khoa sau nhận xét 10/1/2023

Input Form Fields (Right Side):

- Ngày phát hành văn bản đến: Ngày: 07/12/2021
- Mã số văn bản đến: Mã số: 330
- Ngôn ngữ văn bản đến: Tiếng Việt - vn
- Loại văn bản đến: Thông báo - TB
- Độ khẩn văn bản đến: Không
- Độ mật văn bản đến: Không
- Trích yếu văn bản đến: Võ việc đánh giá điểm rèn luyện Học kỳ 1, 2022-2023
- Cán bộ ký văn bản đến: Nguyễn Thành Tường
- Chức vụ cán bộ ký văn bản đến: Trưởng phòng
- Buttons: Lưu thông tin và tiếp tục, Lưu thông tin, Đặt lại, Tùy chỉnh

Hình 2.51 Giao diện thêm công văn

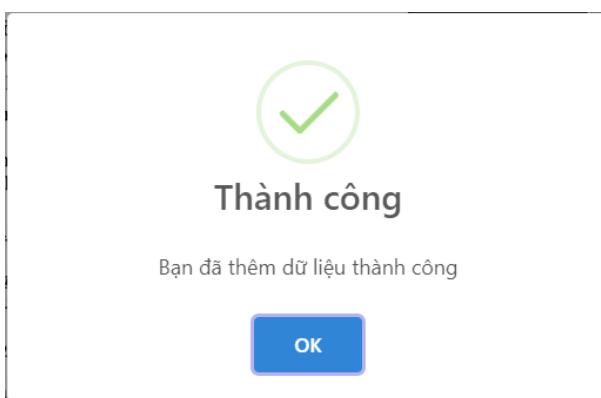
Khi chưa nhập hoặc nhập sai một trường nào đó mà chọn lưu thông tin thì sẽ hiện thông báo lỗi.

Ứng dụng optical character recognition vào hệ thống quản lý công văn

Ngôn ngữ văn bản đến i	Loại văn bản đến i
<input type="button" value="Chọn ngôn ngữ văn bản"/> <input type="button" value="▼"/>	<input type="button" value="Chọn loại văn bản"/> <input type="button" value="▼"/>
Bạn phải nhập ngôn ngữ.	
Độ khẩn văn bản đến i	Độ mật văn bản đến i
<input type="button" value="Chọn độ khẩn"/> <input type="button" value="▼"/>	<input type="button" value="Chọn độ mật"/> <input type="button" value="▼"/>
Bạn phải nhập độ khẩn.	
Bạn phải nhập độ mật.	

Hình 2.52 Thông báo lỗi khi không nhập các trường bắt buộc của công văn đi

Nếu muốn tiếp tục thêm mới công văn sau khi lưu chọn **Lưu thông tin và tiếp tục**, nếu chỉ chọn **Lưu thông tin** thì hệ thống sẽ trả về trang chủ sau khi lưu thành công. Trong quá trình nhập liệu nếu muốn xóa hết các trường đã nhập thì chọn nút **Đặt lại**.

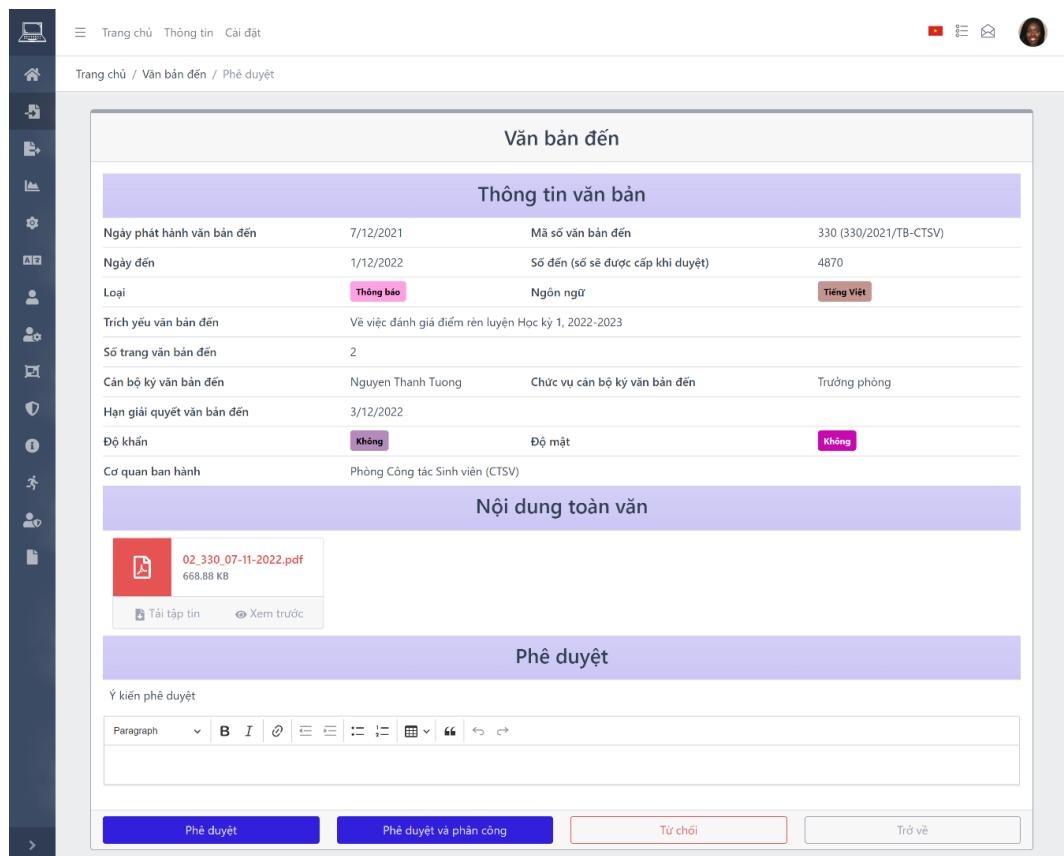


Hình 2.53 Thông báo thêm công văn đến thành công

Ứng dụng optical character recognition vào hệ thống quản lý công văn

2.3.5. Giao diện sửa công văn đến

2.3.6. Giao diện phê duyệt công văn đến



Hình 2.54 Giao diện phê duyệt công văn đến

Từ giao diện danh sách công văn đến ta chọn hoặc từ giao diện chi tiết ta chọn **Phê duyệt** để vào giao diện duyệt công văn đến. Cán bộ sẽ xem một số thông tin và nội dung toàn văn của văn bản cần duyệt. Sau khi xem xét cán bộ sẽ chọn **Phê duyệt** hoặc **Phê duyệt và phân công** nếu muốn phê duyệt hoặc chọn **Tù chối** nếu muốn từ chối văn bản.

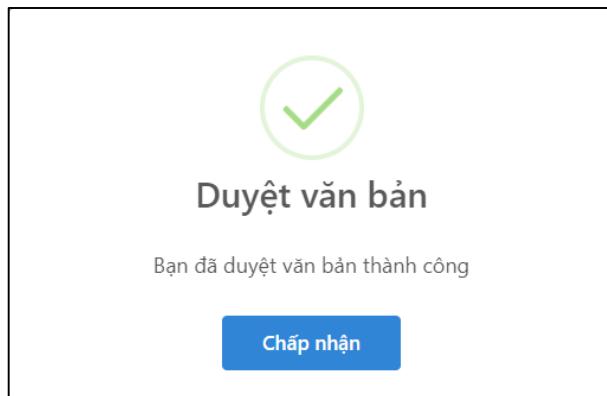
Khi chọn **Phê duyệt và phân công** sẽ xuất hiện hộp thoại để cán bộ chọn cán bộ để xử lý công văn.



Hình 2.55 Hộp thoại chọn cán bộ xử lý công văn

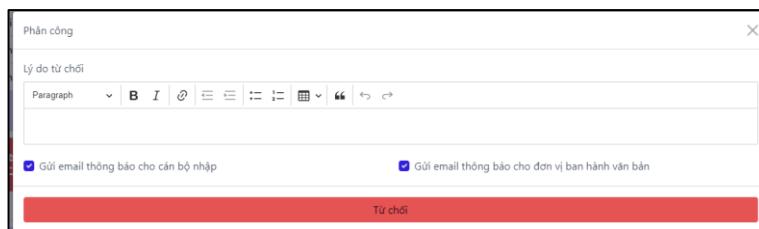
Ứng dụng optical character recognition vào hệ thống quản lý công văn

Sau khi phê duyệt thành công sẽ có thông báo như sau



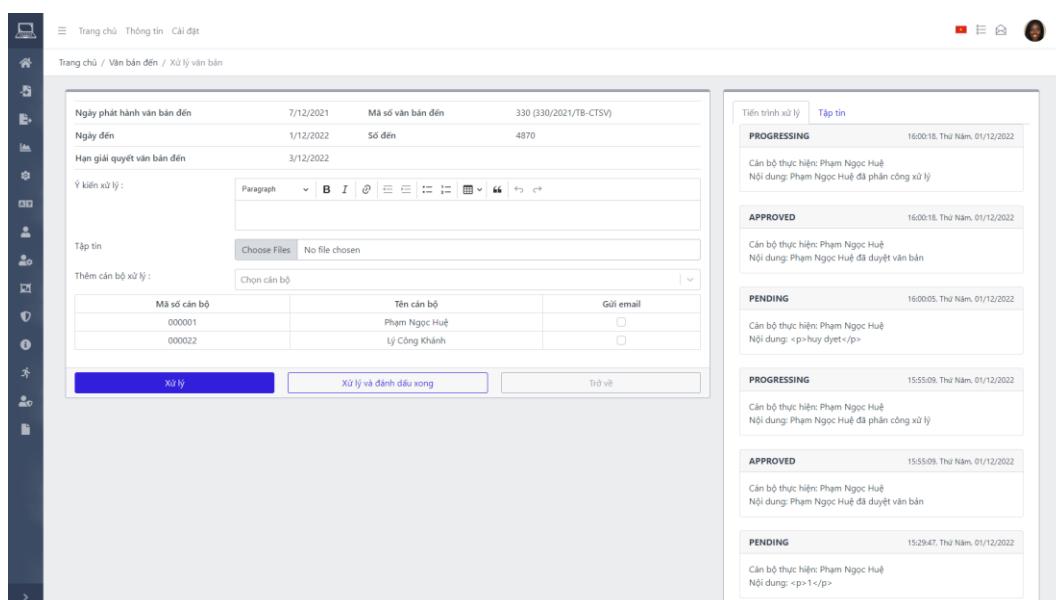
Hình 2.56 Thông báo duyệt thành công

Khi chọn **Tùy chọn** cũng sẽ xuất hiện một hộp thoại để từ chối công văn. Cán bộ phải nhập lý do từ chối và chọn gửi email thông báo cho cán bộ nhập và cho đơn vị ban hành nếu cần.



Hình 2.57 Hộp thoại từ chối công văn

2.3.7. Giao diện xử lý công văn đến



Hình 2.58 Giao diện xử lý công văn

Ứng dụng optical character recognition vào hệ thống quản lý công văn

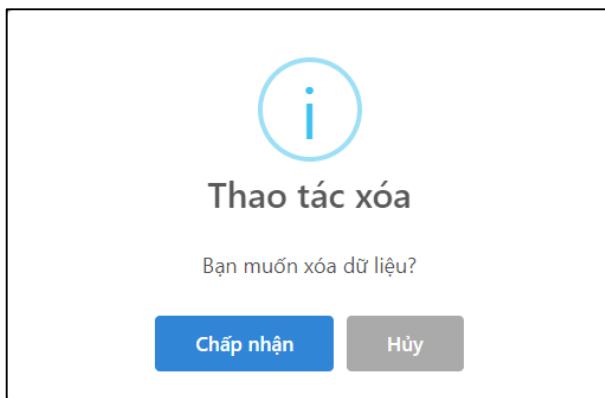
Từ giao diện danh sách công văn đến chọn hoặc từ giao diện chi tiết chọn **Xử lý** để vào giao diện xử lý công văn đến. Tại đây cán bộ sẽ được cung cấp một số thông tin cơ bản như ngày phát hành, mã số, ngày đến, số đến và hạn giải quyết. Bên phải là các tập tin và tiến trình xử lý công văn.

Khi xử lý cán bộ cần nhập ý kiến xử lý. Có thể thêm tập tin hoặc thêm cán bộ xử lý. Sau khi hoàn thành việc xử lý công văn cán bộ xử lý chọn **Xử lý và đánh dấu xong** hoặc nếu còn cần xử lý thêm thì chỉ chọn nút **Xử lý**.

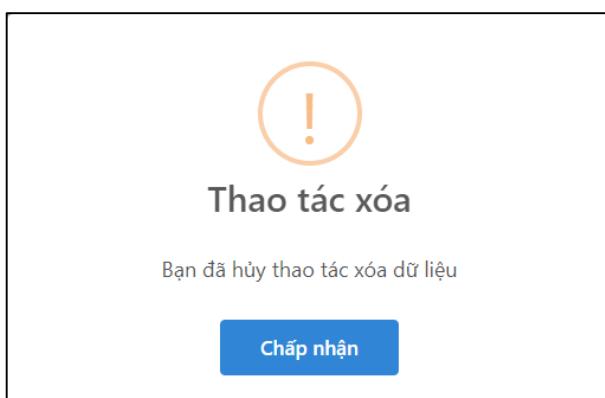
2.3.8. Giao diện xóa công văn đến

Từ giao diện xem danh sách công văn đến chọn hoặc tích vào ô checkbox và chọn nút để xóa công văn. Chỉ cán bộ nhập công văn và chỉ có những công văn ở trạng thái chờ duyệt, từ chối mới có thể xóa được.

Khi xóa sẽ có một thông báo xác nhận việc xóa như sau

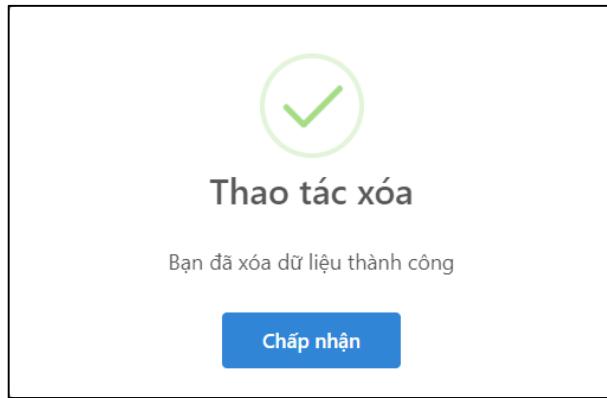


Hình 2.59 Thông báo xác nhận xóa công văn đến



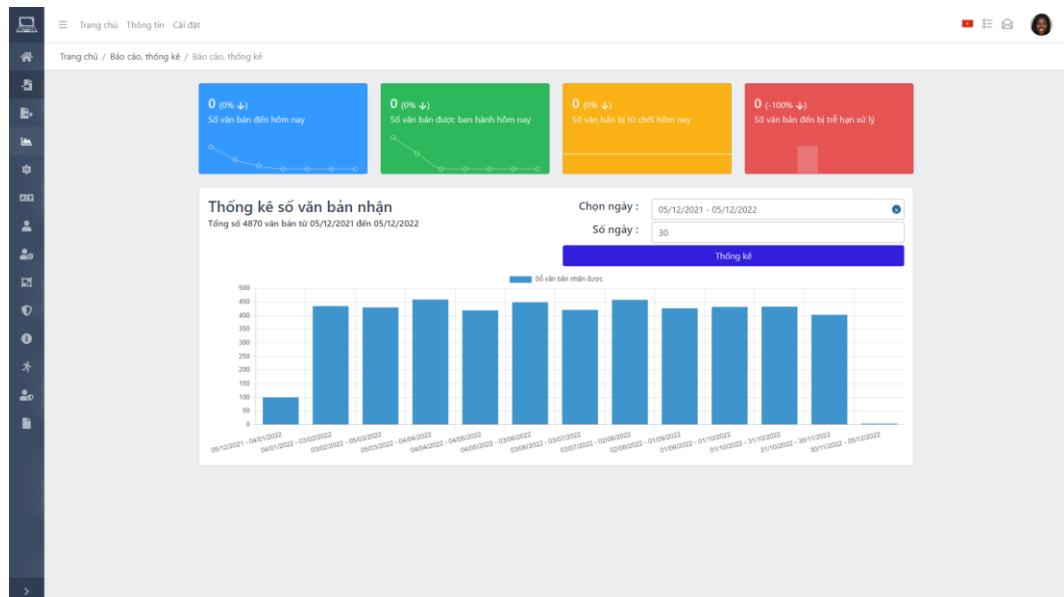
Hình 2.60 Thông báo đã hủy thao tác xóa công văn đến

Ứng dụng optical character recognition vào hệ thống quản lý công văn



Hình 2.61 Thông báo xóa công văn đến thành công

2.3.9. Giao diện báo cáo thống kê công văn đến



Hình 2.62 Giao diện thống kê công văn đến

Giao diện thống kê công văn đến cho biết số công văn đến, số công văn đến được ban hành, số công văn đến bị từ chối và số công văn đến bị trễ hạn trong 7 ngày gần đây. Ngoài ra còn có biểu đồ cột thể hiện số lượng công văn đến theo ngày mà cán bộ muốn xem.

Ứng dụng optical character recognition vào hệ thống quản lý công văn

STT	Đơn vị	Số văn bản đã nhận									
		Tổng	Chờ duyệt/cấp số	Đã duyệt/cấp số	Đã xử lý	Từ chối	Trễ hạn xử lý	Đã triển khai	Đã báo cáo	Đã phát hành	Chờ xử lý
1	Phòng Đào tạo	114	1	17	16	12	32	11	14	11	0
2	Trung tâm Bồi dưỡng Nghiệp vụ Sư phạm	89	0	13	12	8	26	11	9	10	0
3	Trường Bách khoa	105	0	10	11	8	34	15	17	10	0
4	Công Đoàn Trường	116	0	13	8	19	40	15	13	8	0
5	Hội Cựu Chiến binh	120	0	11	13	14	38	16	12	16	0
6	Trung tâm Chuyển giao Công nghệ và Dịch vụ	106	0	18	5	9	39	11	10	14	0
7	Trường Công nghệ thông tin và Truyền thông	110	0	14	8	12	41	14	10	11	0
8	Hội Cựu Sinh viên	117	0	14	17	9	36	11	12	18	0
9	Phòng Công tác Chính trị	99	0	12	14	10	30	10	9	14	0
10	Phòng Công tác Sinh viên	100	0	6	10	20	29	6	13	16	0
11	Trung tâm Công nghệ Phần mềm	93	0	12	13	10	32	9	8	9	0
12	Khoa dự bị Dân tộc	113	0	10	11	13	46	12	12	9	0
13	Trung tâm Điện - Điện tử	100	0	5	17	8	29	15	14	12	0
14	Trung tâm Đánh giá năng lực Ngoại ngữ	108	0	14	12	7	36	17	9	13	0
15	Đoàn Thanh niên CSHCM & Hội Sinh viên	119	0	8	15	14	41	12	13	16	0
16	Trung tâm Đào tạo, NC và Tư vấn kinh tế	119	0	9	16	16	39	15	9	15	0
17	Tổng số	~n	~n	~n	~n	~n	~n	~n	~n	~n	~n

Hình 2.63 Giao diện báo cáo công văn đến

Giao diện báo cáo công văn đến cho biết số lượng công văn đến theo từng đơn vị gửi theo từng loại trong khoảng thời gian mà cán bộ muốn xem. Bên cạnh đó giao diện báo cáo còn cho phép in báo cáo.

CHƯƠNG 3 KIỂM THỦ VÀ ĐÁNH GIÁ

3.1. ĐÁNH GIÁ MÔ HÌNH

3.1.1. Mô tả tập dữ liệu

Dữ liệu được sử dụng là 204 hình ảnh của 99 văn bản hành chính được mô tả chi tiết trong Bảng 3.1

Bảng 3.1 Bảng thống kê loại văn bản hành chính trong tập dữ liệu

STT	Loại văn bản	Số lượng
1	Công văn	53
2	Chương trình	1
3	Chỉ thị	1
4	Kế hoạch	3
5	Thông báo	26
6	Hướng dẫn	1
7	Quyết định	14

Sau khi tiến hành xử lý bóc tách các thành phần và gán nhãn ta thu được 2786 mẫu dữ liệu.

Tổng số thuộc tính: 8 thuộc tính

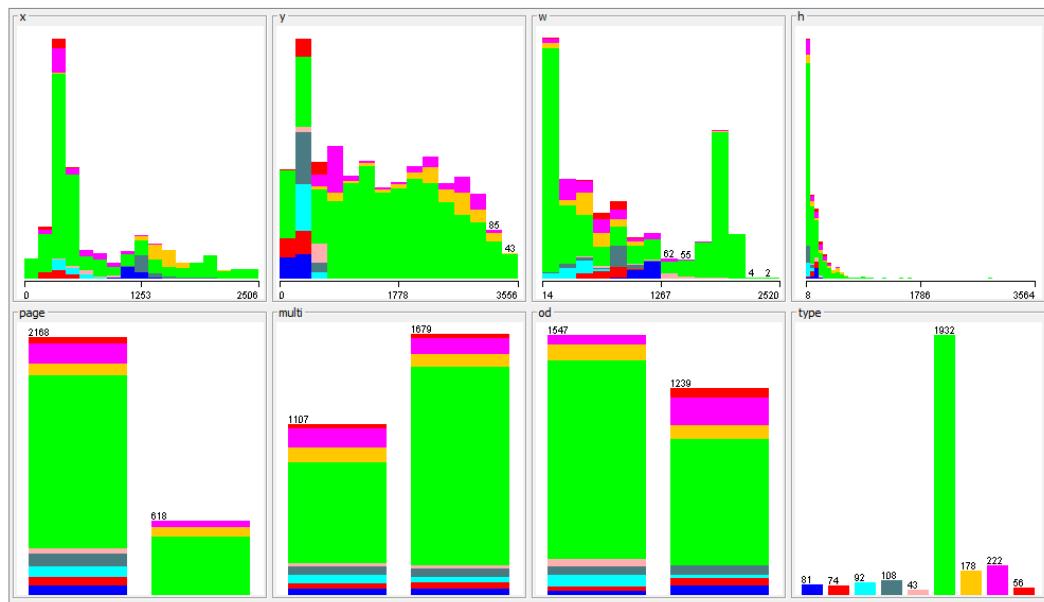
- Thuộc tính lớp: type
- Các lớp:
 - 1: Quốc hiệu và Tiêu ngữ
 - 2: Tên cơ quan tổ chức ban hành
 - 3: Số và ký hiệu
 - 4: Địa danh và thời gian ban hành
 - 5: Tên loại và trích yếu
 - 6: Phần nội dung hoặc lỗi
 - 7: Phần chữ ký
 - 8: Nơi nhận
 - 9: Phần trích dẫn và ký hiệu của công văn
- Tổng số mẫu tin:

Ứng dụng optical character recognition vào hệ thống quản lý công văn

- Tổng số mẫu tin trong lớp 1: 81
- Tổng số mẫu tin trong lớp 2: 74
- Tổng số mẫu tin trong lớp 3: 92
- Tổng số mẫu tin trong lớp 4: 108
- Tổng số mẫu tin trong lớp 5: 43
- Tổng số mẫu tin trong lớp 6: 1932
- Tổng số mẫu tin trong lớp 7: 178
- Tổng số mẫu tin trong lớp 8: 222
- Tổng số mẫu tin trong lớp 9: 56

Bảng 3.2 Bảng mô tả thuộc tính của tập dữ liệu

STT	Tên thuộc tính	Kiểu dữ liệu	Ý nghĩa
1	x	Numeric	Tọa độ x của khối văn bản
2	y	Numeric	Tọa độ y của khối văn bản
3	w	Numeric	Chiều rộng của khối văn bản
4	h	Numeric	Chiều cao của khối văn bản
5	page	Nominal $\{0, 1, 2\}$	Khối văn bản thuộc trang nào. <ul style="list-style-type: none"> ○ 0 nếu khối văn bản thuộc trang đầu. ○ 1 nếu khối văn bản nằm ở trang cuối
6	multiple	Nominal $\{0, 1\}$	Số lượng trang của văn bản <ul style="list-style-type: none"> ○ 0 nếu văn bản có 1 trang ○ 1 nếu văn bản có nhiều trang
7	od	Nominal $\{0, 1\}$	văn bản có phải là công văn hay không <ul style="list-style-type: none"> ○ 0 nếu không là công văn ○ 1 nếu là công văn
8	type	Nominal $\{1, 2, 3, 4, 5, 6, 7, 8, 9\}$	Loại của khối văn bản



Hình 3.1 Biểu diễn tập dữ liệu

3.1.2. Đánh giá mô hình

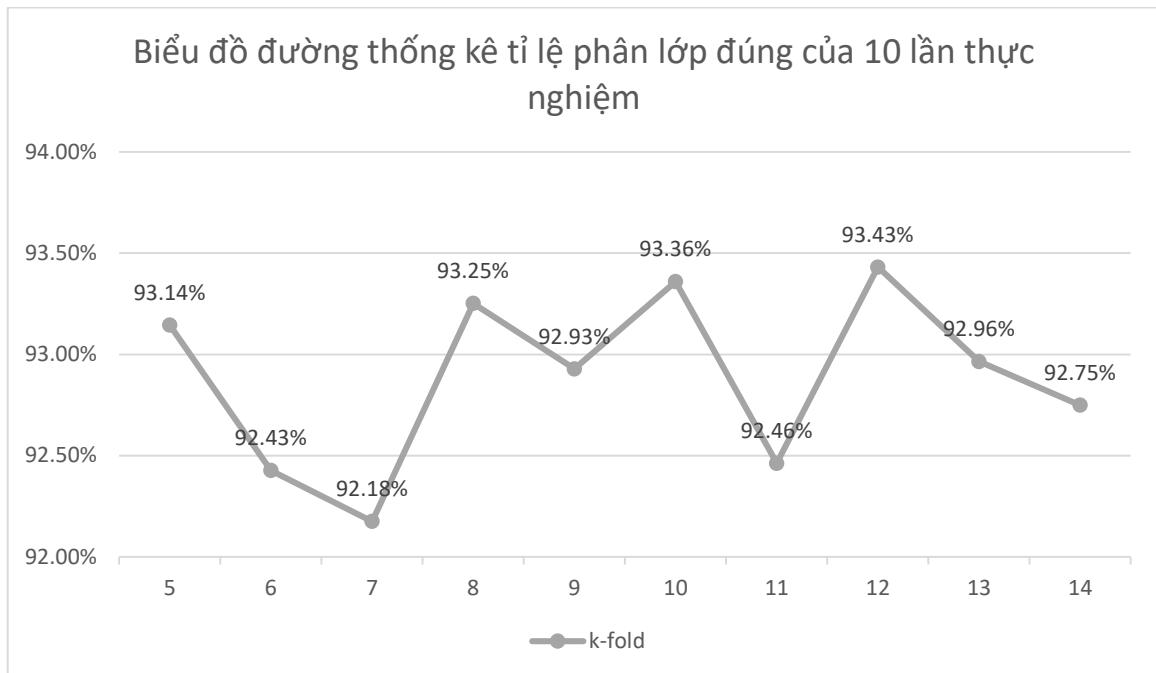
Từ tập dữ liệu ở trên ta tiến hành khai khoáng dữ liệu bằng cách k-fold. Ở đây ta sẽ chạy thực nghiệm 10 lần với giá trị k lần lượt từ 7 đến 16. Kết quả thống kê của 10 lần chạy thực nghiệm được thể hiện trong Bảng 3.3 và Hình 3.2.

Bảng 3.3 Thống kê các lần chạy thực nghiệm bằng cách k-fold

Lần chạy	k	Tổng số bản ghi	Số mẫu phân lớp đúng	Số mẫu phân lớp sai	Tỉ lệ phân lớp đúng (%)	Tỉ lệ phân lớp sai (%)
1	7	2786	2595	191	93.1443	6.8557
2	8	2786	2575	211	92.4264	7.5736
3	9	2786	2568	218	92.1752	7.8248
4	10	2786	2598	188	93.252	6.748
5	11	2786	2589	197	92.9289	7.0711
6	12	2786	2601	185	93.3597	6.6403
7	13	2786	2576	210	92.4623	7.5377
8	14	2786	2603	183	93.4314	6.5686
9	15	2786	2590	196	92.9648	7.0352

Ứng dụng optical character recognition vào hệ thống quản lý công văn

Lần chạy	k	Tổng số bản ghi	Số mẫu phân lớp đúng	Số mẫu phân lớp sai	Tỉ lệ phân lớp đúng (%)	Tỉ lệ phân lớp sai (%)
10	16	2786	2584	202	92.7495	7.2505

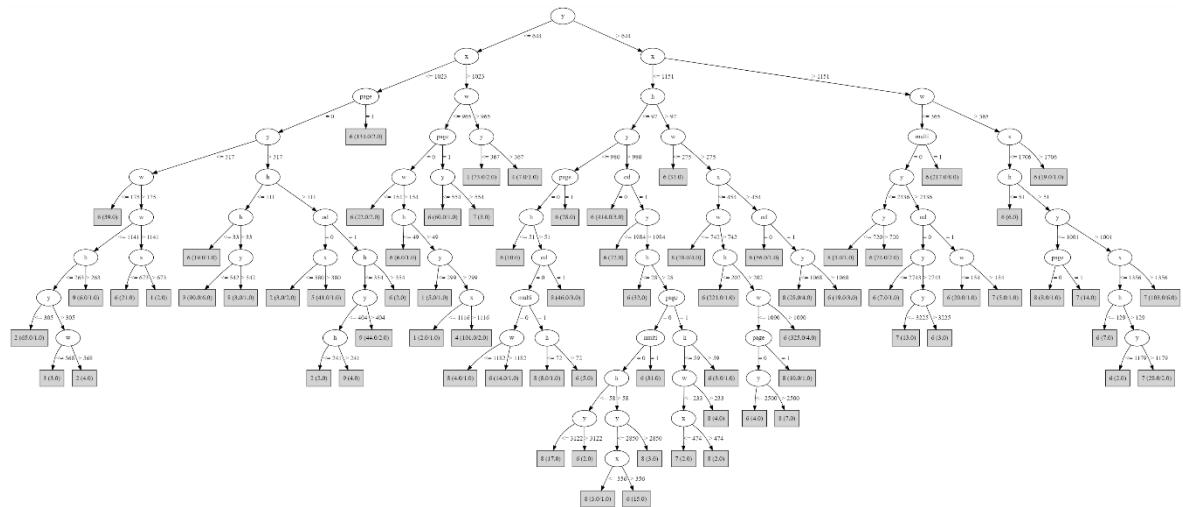


Hình 3.2 Biểu đồ đường thống kê tỉ lệ phân lớp đúng của 10 lần thực nghiệm

Dựa trên biểu đồ thống kê ta thấy được $k=14$ cho tỉ lệ phân lớp đúng cao nhất là 93.4314%. Vậy ta sẽ chọn $k=14$ để đánh giá thuật toán.

Từ kết quả phân khai khoáng dữ liệu ta được cây quyết định có kích thước 143 và số lượng lá là 72.

Ứng dụng optical character recognition vào hệ thống quản lý công văn



Hình 3.3 Kết quả cây quyết định với k=14

Kết quả ma trận Confusion (Bảng 2.4) là một ma trận vuông 9×9 . Nhìn vào ta thấy lớp 1 có 10 mẫu dự đoán sai, tương tự lớp 2 có 7 mẫu, lớp 3 có 10 mẫu, lớp 4 có 6 mẫu, lớp 5 có 6 mẫu, lớp 6 có 59 mẫu, lớp 7 có 31 mẫu, lớp 8 có 47 mẫu và lớp 9 có 7 mẫu dự đoán sai.

Bảng 3.4 Ma trận Confusion

Dự đoán Thực tế	1	2	3	4	5	6	7	8	9
1	71	2	0	7	1	0	0	0	0
2	0	67	3	0	0	0	0	0	4
3	1	5	82	0	1	3	0	0	0
4	0	0	1	102	0	4	0	1	0
5	1	0	0	1	37	2	0	2	0
6	2	1	1	4	1	1873	23	25	2
7	0	0	0	0	0	28	147	3	0
8	0	0	2	0	2	35	6	175	2
9	0	1	4	1	0	1	0	0	49

Dưới đây là bảng thống kê chi tiết kết quả của giải thuật cây quyết định.

Bảng 3.5 Bảng thông số kết quả chạy giải thuật

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0.877	0.001	0.947	0.877	0.910	0.908	0.978	0.848	1
0.905	0.003	0.882	0.905	0.893	0.890	0.970	0.846	2
0.891	0.004	0.882	0.891	0.886	0.883	0.968	0.810	3
0.944	0.005	0.887	0.944	0.915	0.912	0.983	0.877	4
0.860	0.002	0.881	0.860	0.871	0.869	0.959	0.825	5
0.969	0.085	0.962	0.969	0.966	0.888	0.953	0.961	6
0.826	0.011	0.835	0.826	0.831	0.819	0.930	0.738	7
0.788	0.012	0.850	0.788	0.818	0.803	0.926	0.731	8
0.875	0.003	0.860	0.875	0.867	0.865	0.968	0.787	9
0.934	0.061	0.934	0.934	0.934	0.878	0.953	0.908	Weighted Avg.

Đánh giá các giá trị quan trọng:

- TP Rate: Cao (93,4%)
- FP Rate: Thấp (6.1%)
- Precision: Cao (93,4%)
- Recall: Cao (93,4%)
- F1 (F-Measure): Cao (93,4%)

3.2. KIỂM THỬ CHỨC NĂNG PHÂN TÍCH VĂN BẢN

3.2.1. Môi trường kiểm thử

Phần cứng: Laptop có cấu hình như sau

- Bộ xử lý: Intel® Core(TM) i5-8265U CPU @ 1.60GHz 1.80 GHz
- RAM: 12 GB
- Hệ điều hành: Windows 10 Enterprise LTSC 64-bit
- Có kết nối mạng internet

Dữ liệu kiểm thử: Chuẩn bị 6 văn bản hành chính được đánh mã số lần lượt là vb1, vb2, vb3, vb4, vb5 và vb6. Trong đó văn bản vb1, vb2, vb3 không nằm trong dữ liệu huấn luyện vb4, vb5, vb6 nằm trong dữ liệu huấn luyện mô hình.

3.2.2. Kết quả kiểm thử

Bảng 3.6 Kết quả kiểm thử chức năng phân tích văn bản đến

STT	Đặc tả	Kết quả mong đợi	Kết quả đạt được	Đánh giá
	vb1			
1	Tỷ lệ dự đoán thành phần trong văn bản	$\geq 93\%$	97,3% (36/37)	Thành công
2	Tỷ lệ trích xuất dữ liệu trong văn bản	$\geq 50\%$	57,14% (4/7)	Thành công
3	Thời gian thực hiện trích xuất	≤ 120 giây	105.822 giây	Thành công
	vb2			
4	Tỷ lệ dự đoán thành phần trong văn bản	$\geq 93\%$	100% (35/35)	Thành công
5	Tỷ lệ trích xuất dữ liệu trong văn bản	$\geq 50\%$	71.43% (5/7)	Thành công
6	Thời gian thực hiện trích xuất	≤ 120 giây	93.854 giây	Thành công
	vb3			
7	Tỷ lệ dự đoán thành phần trong văn bản	$\geq 93\%$	100% (35/35)	Thành công
8	Tỷ lệ trích xuất dữ liệu trong văn bản	$\geq 50\%$	57,14% (4/7)	Thành công
9	Thời gian thực hiện trích xuất	≤ 120 giây	99.697 giây	Thành công
	vb4			
10	Tỷ lệ dự đoán thành phần trong văn bản	$\geq 93\%$	90% (9/10)	Thất bại
11	Tỷ lệ trích xuất dữ liệu trong văn bản	$\geq 50\%$	42,86% (3/7)	Thất bại

Ứng dụng optical character recognition vào hệ thống quản lý công văn

STT	Đặc tả	Kết quả mong đợi	Kết quả đạt được	Đánh giá
12	Thời gian thực hiện trích xuất	<= 120 giây	55,199 giây	Thành công
	vb5			
13	Tỷ lệ dự đoán thành phần trong văn bản	>= 93%	100% (27/27)	Thành công
14	Tỷ lệ trích xuất dữ liệu trong văn bản	>= 50%	42,86% (3/7)	Thất bại
15	Thời gian thực hiện trích xuất	<= 120 giây	106, 816 giây	Thành công
	vb6			
16	Tỷ lệ dự đoán thành phần trong văn bản	>= 93%	100% (10/10)	Thành công
17	Tỷ lệ trích xuất dữ liệu trong văn bản	>= 50%	42,86% (3/7)	Thất bại
18	Thời gian thực hiện trích xuất	<= 120 giây	52.446 giây	Thành công

PHẦN KẾT LUẬN

I. KẾT QUẢ ĐẠT ĐƯỢC

I.1. Kiến thức đạt được

- Hiểu được nghiệp vụ quản lý công văn.
- Áp dụng được kỹ thuật Data Mining trong việc phân tích công văn.
- Hiểu được cách công nghệ OCR và ứng dụng được vào việc quản lý công văn.
- Sử dụng tốt các công cụ phân tích thiết kế hệ thống để có cái nhìn tổng quan về một hệ thống thông tin.
- Xây dựng được hệ thống quản lý công văn có tích hợp Data Mining và OCR.
- Học được các công nghệ mới là React, Node, MongoDB.

I.2. Kinh nghiệm thực tiễn

- Biết được các bước quan trọng trong việc xây dựng một hệ thống.
- Có khả năng thu thập dữ liệu để xây dựng một mô hình máy học đơn giản.
- Biết được cách tìm kiếm tài liệu về các công nghệ mới, có khả năng tìm kiếm và đánh giá các nguồn tài liệu tham khảo.
- Có thêm nhiều kinh nghiệm trong việc thiết kế, phân tích và lập trình hệ thống thông qua việc giải quyết các vấn đề gặp phải trong quá trình làm luận văn.

I.3. Hệ thống

- Giao diện ứng dụng đẹp, dễ thao tác.
- Có đủ các chức năng cần thiết cho quá trình quản lý công văn.
- Các chức năng cần thiết cho quá trình quản lý trở nên nhanh chóng, dễ dàng và độ chính xác cao, đảm bảo toàn vẹn dữ liệu.

II. HẠN CHẾ

Do hạn chế về thời gian, kiến thức và khả năng phân bổ thời gian. Vì vậy hệ thống vẫn chưa thực sự hoàn thiện:

- Tập dữ liệu còn rất ít cho việc huấn luyện, kiểm nghiệm mô hình.
- Thời gian chờ của hệ thống còn dài.
- Khả năng dự đoán còn phụ thuộc nhiều vào chất lượng văn bản đầu vào.
- Khả năng trích xuất phụ thuộc hoàn toàn vào tesseract nên khả năng nhận dạng tiếng Việt và chữ viết tay còn hạn chế.
- Hệ thống vẫn chưa hoàn thành được các chức năng phụ.

III. HƯỚNG PHÁT TRIỂN

Hệ thống sẽ được phát triển để thời gian xử lý, dự đoán và trích xuất công văn được cải thiện hơn so với phiên bản hiện tại. Cùng với việc xây dựng một mô hình tốt hơn để tăng tỷ lệ dự đoán đúng.

Cải thiện các bước xử lý ảnh để tăng khả năng nhận dạng ký tự của Tesseract. Áp dụng thêm các kỹ thuật xử lý chuỗi để trích xuất được thông tin chính xác hơn.

Mở rộng thêm các chức năng như nhắn tin trực tiếp, gửi và nhận email trực tiếp trên hệ thống, chức năng xem lịch làm việc, lịch giải quyết công văn, tự động nhắc nhở khi công văn gần đến hạn giải quyết.

TÀI LIỆU THAM KHẢO

1. Thông T.T. (2019), *Ứng dụng học máy trong nhận dạng công văn các cơ quan Đảng tỉnh Quảng Bình*, (Luận văn thạc sĩ), Trường Đại học Bách khoa - Đại học Đà Nẵng, Đà Nẵng.
2. Memon J., Sami M., Khan R.A. và cộng sự. (2020). Handwritten optical character recognition (OCR): A comprehensive systematic literature review (SLR). *IEEE Access*, **8**, 142642–142668.
3. Mori S., Nishida H., và Yamada H. (1999), *Optical character recognition*, John Wiley & Sons, Inc.
4. Eikvil L. (1993). Optical character recognition. *citeeseer ist psu edu/142042 html*, **26**.
5. Smith R. (2007). An overview of the Tesseract OCR engine. *Ninth international conference on document analysis and recognition (ICDAR 2007)*, IEEE, 629–633.
6. Hiệp T.Đ. (2013), *Nghiên cứu nhận dạng chữ in trong ảnh scan, ứng dụng vào trích lọc thông tin trích yếu của văn bản hành chính*, (Luận văn thạc sĩ), Trường Đại học Cần Thơ.
7. Nghị Đ.T. (2011). Chương 4 Máy học cây quyết định. *Khai mỏ dữ liệu: Minh họa bằng ngôn ngữ R*. Đại học Cần Thơ, Cần Thơ, 22.
8. Trung L.T. (2010), *Xây dựng cây quyết định trên tập dữ liệu nhỏ*, (Luận văn thạc sĩ), Trường Đại học Cần Thơ, Cần Thơ.
9. Đệ T.C. và Khang P.N. (2012). Phân loại văn bản với máy học Vector hỗ trợ và cây quyết định. *Tạp chí Khoa học Trường Đại học Cần Thơ*, **(21a)**, 52–63.
10. Kiên P.T. (2019), *Khai thác và phân tích dữ liệu nhằm quản lý rủi ro trong giao dịch tín dụng*, (Luận văn thạc sĩ), Trường Đại học Công nghệ - Đại học Quốc gia Hà Nội, Hà Nội.
11. Poll: Deployed Data Mining Techniques. <https://www.kdnuggets.com/polls/2004/deployed_data_mining_techniques.htm>, accessed: 16/11/2022.
12. Nghị Đ.T. (2011). Chương 2 Giải thuật k láng giềng. *Khai mỏ dữ liệu: Minh họa bằng ngôn ngữ R*. Đại học Cần Thơ, Cần Thơ, 12.
13. Chính phủ (2020), *Về công tác văn thư*, (Số 30/2020/NĐ-CP).
14. Trang chủ - Hệ thống quản lý công văn. <<https://qlcv.ctu.edu.vn/>>, accessed: 03/12/2022.
15. Trang chủ. <<https://dap.ctu.edu.vn/>>, accessed: 03/12/2022.
16. Hiệp H.X. và Linh N.T.T. (2018), *Giáo trình thiết kế web*, Nhà xuất bản Đại học Cần Thơ, Cần Thơ.

Ứng dụng optical character recognition vào hệ thống quản lý công văn

17. Holeczek Ł. Introduction. <<http://coreui.io/v1/docs/getting-started/introduction/>>, accessed: 27/11/2022.
18. Holmes S. (2013), *Mongoose for Application Development*, Packt Publishing Ltd.
19. Chodorow K. và Dirolf M. (2010), *MongoDB: The Definitive Guide*, O'Reilly Media.
20. Banks A. và Porcello E. (2017), *Learning React: functional web development with React and Redux*, O'Reilly Media, Inc.
21. Powers S. (2012), *Learning Node*, O'Reilly Media, Inc.
22. Brown E. (2019), *Web development with node and express: leveraging the JavaScript stack*, O'Reilly Media.
23. Weka 3 - Data Mining with Open Source Machine Learning Software in Java. <<https://www.cs.waikato.ac.nz/ml/weka/>>, accessed: 27/11/2022.
24. Kaehler A. và Bradski G. (2016), *Learning OpenCV 3: computer vision in C++ with the OpenCV library*, O'Reilly Media, Inc.
25. OpenCV: Color conversions. <https://docs.opencv.org/3.4/de/d25/imgproc_color_conversions.html>, accessed: 29/11/2022.

**PHỤ LỤC I
BIỂU MẪU THƯỜNG GẶP**

Mẫu 1 Nghị quyết (cá biệt)

TÊN CQ, TC CHỦ QUẢN¹ CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM

TÊN CƠ QUAN, TỔ CHỨC²

Độc lập – Tự do – Hạnh phúc

Số: /NQ-...³...

...⁴ ..., ngày tháng năm

NGHỊ QUYẾT

.....
.....

THẨM QUYỀN BAN HÀNH

Căn cứ.....;

Căn cứ.....;

QUYẾT ĐỊNH

.....
.....
.....
...../.

Nơi nhận:

- Như điều,;
-;
- Lưu: VT, ...⁷ ...⁸.

QUYỀN HẠN, CHỨC VỤ CỦA NGƯỜI KÝ

(Chữ ký của người có thẩm quyền,
dấu/chữ ký số của cơ quan, tổ chức)

Họ và tên

Ghi chú:

¹ Tên cơ quan, tổ chức chủ quản trực tiếp (nếu có).

² Tên cơ quan, tổ chức ban hành nghị quyết.

³ Chữ viết tắt tên cơ quan, tổ chức ban hành nghị quyết.

⁴ Địa danh.

⁵ Trích yếu nội dung nghị quyết.

⁶ Nội dung nghị quyết.

⁷ Chữ viết tắt tên đơn vị soạn thảo văn bản và số lượng bản lưu (nếu cần).

⁸ Ký hiệu người soạn thảo văn bản và số lượng bản phát hành (nếu cần).

Mẫu 2 Quyết định (cá biệt) quy định trực tiếp

TÊN CQ, TC CHỦ QUẢN¹

CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM

TÊN CƠ QUAN, TỔ CHỨC²

Độc lập – Tự do – Hạnh phúc

Số: /QĐ-....³...

...⁴ ..., ngày tháng năm

QUYẾT ĐỊNH

Về việc⁵

THẨM QUYỀN BAN HÀNH⁶

Căn cứ⁷

Căn cứ;

Theo đề nghị của

QUYẾT ĐỊNH:

Điều 1.....⁸

Điều.....

Nơi nhận:

- Như điều,;
-;
- Lưu: VT, ...⁹ ...¹⁰.

QUYỀN HẠN, CHỨC VỤ CỦA NGƯỜI KÝ

(Chữ ký của người có thẩm quyền,
dấu/chữ ký số của cơ quan, tổ chức)

Họ và tên

Ghi chú:

¹ Tên cơ quan, tổ chức chủ quản trực tiếp (nếu có).

² Tên cơ quan, tổ chức hoặc chức danh nhà nước ban hành quyết định.

³ Chữ viết tắt tên cơ quan, tổ chức hoặc chức danh nhà nước ban hành quyết định.

⁴ Địa danh

⁵ Trích yếu nội dung quyết định.

⁶ Thẩm quyền ban hành quyết định thuộc về người đứng đầu cơ quan, tổ chức thì ghi chức vụ của người đứng đầu; nếu thẩm quyền ban hành quyết định thuộc về tập thể lãnh đạo hoặc cơ quan, tổ chức thì ghi tên tập thể hoặc tên cơ quan, tổ chức đó.

⁷ Các căn cứ để ban hành quyết định.

⁸ Nội dung quyết định.

⁹ Chữ viết tắt tên đơn vị soạn thảo văn bản và số lượng bản lưu (nếu cần).

¹⁰ Ký hiệu người soạn thảo văn bản và số lượng bản phát hành (nếu cần).

Mẫu 3 Văn bản có tên loại

TÊN CQ, TC CHỦ QUẢN¹ CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM

TÊN CƠ QUAN, TỔ CHỨC²

Độc lập – Tự do – Hạnh phúc

Số: /...³....-....⁴...

...⁵..., ngày tháng năm

TÊN LOẠI VĂN BẢN⁶

.....⁷

.....⁸

Noi nhận:

- Như điều;
-;
- Lưu: VT,...⁹...¹⁰.

QUYỀN HẠN, CHỨC VỤ CỦA NGƯỜI KÝ

(Chữ ký của người có thẩm quyền,
dấu/chữ ký số của cơ quan, tổ chức)

Họ và tên

¹ Tên cơ quan, tổ chức chủ quản trực tiếp (nếu có).

² Tên cơ quan, tổ chức hoặc chức danh nhà nước ban hành văn bản.

³ Chữ viết tắt tên loại văn bản.

⁴ Chữ viết tắt tên cơ quan, tổ chức hoặc chức danh nhà nước ban hành văn bản.

⁵ Địa danh

⁶ Tên loại văn bản.

Mẫu này áp dụng chung đối với các hình thức văn bản hành chính có ghi tên loại gồm: chi thị, quy chế, quy định, thôn| cáo, thông báo, hướng dẫn, chương trình, kế hoạch, phương án, đề án, dự án, báo cáo, tờ trình, giấy ủy quyền, phiếu gửi, phiếu chuyên, phiếu báo.

⁷ Trích yếu nội dung văn bản.

⁸ Nội dung văn bản.

⁹ Chữ viết tắt tên đơn vị soạn thảo văn bản và số lượng bản lưu (nếu cần).

¹⁰ Ký hiệu người soạn thảo văn bản và số lượng bản phát hành (nếu cần).

Mẫu 4 Công văn

TÊN CQ, TC CHỦ QUẢN¹ CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM

TÊN CƠ QUAN, TỔ CHỨC²

Độc lập – Tự do – Hạnh phúc

Số: /...³.....⁴...

...⁶..., ngày tháng năm

V/v⁵.....

Kính gửi:

-;

-

7

.....

.....
...../.
.....

Noi nhận:

- Như điều;
-;
- Lưu: VT, ...⁸...⁹.

QUYỀN HẠN, CHỨC VỤ CỦA NGƯỜI KÝ

(Chữ ký của người có thẩm quyền,
dấu/chữ ký số của cơ quan, tổ chức)

Họ và tên

10

¹ Tên cơ quan, tổ chức chủ quản trực tiếp (nếu có).

² Tên cơ quan, tổ chức hoặc chức danh nhà nước ban hành công văn.

³ Chữ viết tắt tên loại văn bản.

⁴ Chữ viết tắt tên cơ quan, tổ chức hoặc chức danh nhà nước ban hành công văn.

⁵ Trích yếu nội dung công văn.

⁶ Địa danh

⁷ Nội dung công văn.

⁸ Chữ viết tắt tên đơn vị soạn thảo văn bản và số lượng bản lưu (nếu cần).

⁹ Ký hiệu người soạn thảo văn bản và số lượng bản phát hành (nếu cần).

¹⁰ Địa chỉ cơ quan, tổ chức; thư điện tử; trang thông tin điện tử; số điện thoại; số Fax (nếu cần).

Mẫu 5 Công điện

TÊN CQ, TC CHỦ QUẢN¹

CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM

TÊN CƠ QUAN, TỔ CHỨC²

Độc lập – Tự do – Hạnh phúc

Số: .../CĐ-...³...

...⁴ ..., ngày tháng năm

CÔNG ĐIỆN

.....⁵

.....⁶ điện:

-⁷

-

.....⁸

.....
.....
.....
.....
.....

Noi nhận:

- Như điều;
-;
- Lưu: VT,...⁹...¹⁰.

QUYỀN HẠN, CHỨC VỤ CỦA NGƯỜI KÝ

(Chữ ký của người có thẩm quyền,
dấu/chữ ký số của cơ quan, tổ chức)

Họ và tên

Ghi chú:

¹ Tên cơ quan, tổ chức chủ quản trực tiếp (Nếu có).

² Tên cơ quan, tổ chức ban hành công điện.

³ Chữ viết tắt tên cơ quan, tổ chức ban hành công điện.

⁴ Địa danh.

⁵ Trích yếu nội dung điện.

⁶ Tên cơ quan, tổ chức hoặc chức danh của người đứng đầu.

⁷ Tên cơ quan, tổ chức nhận điện.

⁸ Nội dung điện.

⁹ Chữ viết tắt tên đơn vị soạn thảo văn bản và số lượng bản lưu (nếu cần).

¹⁰ Ký hiệu người soạn thảo văn bản và số lượng bản phát hành (nếu cần).

Mẫu 6 Giấy mời

TÊN CQ, TC CHỦ QUẢN¹

CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM

TÊN CƠ QUAN, TỔ CHỨC²

Độc lập – Tự do – Hạnh phúc

Số: /GM-...³ ...

...⁴ ..., ngày tháng năm

GIẤY MỜI

.....⁵

.....² trân trọng kính mời:⁶

Tới dự⁷

Chủ trì:

Thời gian:

Địa điểm:

.....⁸/.⁹

Noi nhận:

- Như điều;
-;
- Lưu: VT, ...⁹ ...¹⁰.

QUYỀN HẠN, CHỨC VỤ CỦA NGƯỜI KÝ

(Chữ ký của người có thẩm quyền,
dấu/chữ ký số của cơ quan, tổ chức)

Họ và tên

Ghi chú:

¹ Tên cơ quan, tổ chức chủ quản trực tiếp (Nếu có).

² Tên cơ quan, tổ chức ban hành giấy mời.

³ Chữ viết tắt tên cơ quan, tổ chức ban hành giấy mời.

⁴ Địa danh.

⁵ Trích yếu nội dung cuộc họp.

⁶ Tên cơ quan, tổ chức hoặc họ và tên, chức vụ, đơn vị công tác của người được mời.

⁷ Tên (nội dung) của cuộc họp, hội thảo, hội nghị v.v...

⁸ Các vấn đề cần lưu ý.

⁹ Chữ viết tắt tên đơn vị soạn thảo văn bản và số lượng bản lưu (nếu cần).

¹⁰ Ký hiệu người soạn thảo văn bản và số lượng bản phát hành (nếu cần).

Mẫu 7 Giấy giới thiệu

TÊN CQ, TC CHỦ QUẢN¹

CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM

TÊN CƠ QUAN, TỔ CHỨC²

Độc lập – Tự do – Hạnh phúc

Số: .../GGT-...³...

...⁴ ..., ngày tháng năm

GIẤY GIỚI THIỆU

.....² trân trọng giới thiệu:

Ông (bà)⁵

Chức vụ:

Được cử đến⁶

Về việc:

Đề nghị Quý cơ quan tạo điều kiện để ông bà có tên ở trên hoàn thành nhiệm vụ.

Giấy này có giá trị đến hết ngày/..

Noi nhận:

- Như trên

- Lưu: VT.

QUYỀN HẠN, CHỨC VỤ CỦA NGƯỜI KÝ

(Chữ ký của người có thẩm quyền,
dấu/chữ ký số của cơ quan, tổ chức)

Họ và tên

Ghi chú:

¹ Tên cơ quan, tổ chức chủ quản trực tiếp (Nếu có).

² Tên cơ quan, tổ chức ban hành văn bản (cấp giấy giới thiệu).

³ Chữ viết tắt tên cơ quan, tổ chức ban hành văn bản.

⁴ Địa danh.

⁵ Họ và tên, chức vụ và đơn vị công tác của người được giới thiệu.

⁶ Tên cơ quan, tổ chức được giới thiệu tới làm việc.

Mẫu 8 Biên bản

TÊN CQ, TC CHỦ QUẢN¹

CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM

TÊN CƠ QUAN, TỔ CHỨC²

Độc lập – Tự do – Hạnh phúc

Số: .../BB-...³...

GIẤY MỜI

.....⁴.....

Thời gian bắt đầu:

Địa điểm:

Thành phần tham dự:

Chủ trì (chủ tọa):

Nội dung (theo diễn biến cuộc họp/hội nghị/hội thảo):

.....
Cuộc họp (hội nghị, hội thảo) kết thúc vào ... giờ ..., ngày ... tháng ... năm
..../.

THUẾ KÝ

(Chữ ký)

Họ và tên

CHỦ TỌA

(Chữ ký của người có thẩm quyền,
đầu/chữ ký số của cơ quan, tổ chức (nếu có))⁵

Họ và tên

Noi nhận:

-;
- Lưu: VT, Hò sơ.

Ghi chú:

¹ Tên cơ quan, tổ chức chủ quản trực tiếp (Nếu có).

² Tên cơ quan, tổ chức ban hành văn bản.

³ Chữ viết tắt tên cơ quan, tổ chức ban hành văn bản.

⁴ Tên cuộc họp hoặc hội nghị, hội thảo.

⁵ Ghi chức vụ chính quyền (nếu cần).

Mẫu 9 Giấy nghỉ phép

TÊN CQ, TC CHỦ QUẢN¹

CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM

TÊN CƠ QUAN, TỔ CHỨC²

Độc lập – Tự do – Hạnh phúc

Số: .../GNP-...³...

...⁴ ..., ngày tháng năm

GIẤY NGHỈ PHÉP

Xét Đơn đề nghị nghỉ phép ngày của ông (bà)
.....² cấp cho:
Ông (bà):⁵
Chức vụ:
Được nghỉ phép trong thời gian kể từ ngày đến hết ngày
..... tại⁶
Số ngày nghỉ phép nêu trên được tính vào thời gian⁷ /.

Noi nhận:

-.....⁸

- Lưu: VT, ...⁹

QUYỀN HẠN, CHỨC VỤ CỦA NGƯỜI KÝ

(Chữ ký của người có thẩm quyền,
dấu/chữ ký số của cơ quan, tổ chức)

Họ và tên

Xác nhận của cơ quan (tổ chức) hoặc
chính quyền địa phương nơi nghỉ phép
(nếu cần)

(Chữ ký, dấu)

Họ và tên

Ghi chú:

¹ Tên cơ quan, tổ chức chủ quản trực tiếp (Nếu có).

² Tên cơ quan, tổ chức cấp giấy nghỉ phép.

³ Chữ viết tắt tên cơ quan, tổ chức cấp giấy nghỉ phép.

⁴ Địa danh.

⁵ Họ và tên, chức vụ và đơn vị công tác của người được cấp giấy phép.

⁶ Nơi nghỉ phép.

⁷ Thời gian nghỉ theo Luật Lao động (nghỉ hàng năm có lương hoặc nghỉ không hưởng lương hoặc nghỉ việc riêng mà vẫn hưởng nguyên lương...).

⁸ Người được cấp giấy nghỉ phép.

⁹ Ký hiệu người soạn thảo văn bản và số lượng bản phát hành (nếu cần).

PHỤ LỤC II
CÁC BẢNG DỮ LIỆU

Tên bảng		Languages												
Mô tả tên bảng		Ngôn ngữ văn bản												
Thuộc tính														
STT	Tên thuộc tính	Kiểu	Kích thước	Số thập phân	Trị mặc định	Miền giá trị	Min	Max	Khóa chính	Duy nhất	Not null	Khóa ngoại	Ràng buộc toàn vẹn	Diễn giải
1	Id	varchar	12						x		x			ID
2	Name	varchar	100								x			Tên
3	Notation	varchar	10								x			Ký hiệu
4	Color	varchar	7		#12B7BC		7	7			x			Màu
5	Description	varchar	1000											Mô tả
6	Deleted	boolean			False						x			Đã xóa

Ứng dụng optical character recognition vào hệ thống quản lý công văn

Tên bảng	Type													
Mô tả tên bảng	Loại văn bản													
Thuộc tính														
STT	Tên thuộc tính	Kiểu	Kích thước	Số thập phân	Trị mặc định	Miền giá trị	Min	Max	Khóa chính	Duy nhất	Not null	Khóa ngoại	Ràng buộc toàn vẹn	Diễn giải
1	Id	varchar	12						x		x			ID
2	Name	varchar	100								x			Tên
3	Notation	varchar	10								x			Ký hiệu
4	Color	varchar	7		#12B7BC		7	7			x			Màu
5	Description	varchar	1000											Mô tả
6	Deleted	boolean			False						x			Đã xóa

Ứng dụng optical character recognition vào hệ thống quản lý công văn

Tên bảng	Security													
Mô tả tên bảng	Độ mật													
Thuộc tính														
STT	Tên thuộc tính	Kiểu	Kích thước	Số thập phân	Trị mặc định	Miền giá trị	Min	Max	Khóa chính	Duy nhất	Not null	Khóa ngoại	Ràng buộc toàn vẹn	Diễn giải
1	Id	varchar	12						x		x			ID
2	Name	varchar	100								x			Tên
3	Deleted	boolean			False						x			Đã xóa

Ứng dụng optical character recognition vào hệ thống quản lý công văn

Tên bảng	Priority													
Mô tả tên bảng	Độ khẩn													
Thuộc tính														
STT	Tên thuộc tính	Kiểu	Kích thước	Số thập phân	Trị mặc định	Miền giá trị	Min	Max	Khóa chính	Duy nhất	Not null	Khóa ngoại	Ràng buộc toàn vẹn	Diễn giải
1	Id	varchar	12						x		x			ID
2	Name	varchar	100								x			Tên
3	Deleted	boolean			False						x			Đã xóa

Ứng dụng optical character recognition vào hệ thống quản lý công văn

Tên bảng		Status												
Mô tả tên bảng		Trạng thái văn bản												
Thuộc tính														
STT	Tên thuộc tính	Kiểu	Kích thước	Số thập phân	Trị mặc định	Miền giá trị	Min	Max	Khóa chính	Duy nhất	Not null	Khóa ngoại	Ràng buộc toàn vẹn	Diễn giải
1	Id	varchar	12						x		x			ID
2	Name	varchar	100								x			Tên
3	Description	boolean	1000											Mô tả
4	Deleted	boolean			False						x			Đã xóa

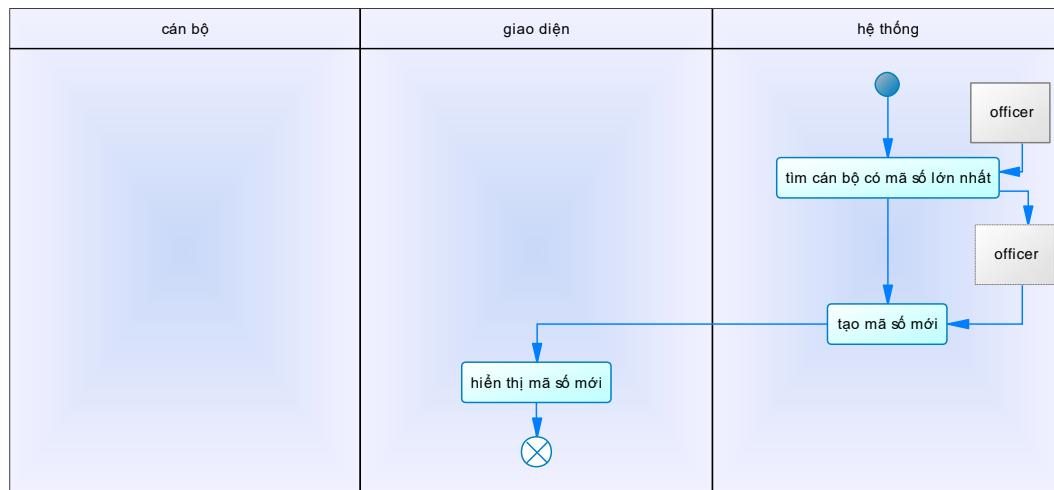
Ứng dụng optical character recognition vào hệ thống quản lý công văn

Tên bảng	OfficerStatus													
Mô tả tên bảng	Trạng thái cán bộ													
Thuộc tính														
STT	Tên thuộc tính	Kiểu	Kích thước	Số thập phân	Trị mặc định	Miền giá trị	Min	Max	Khóa chính	Duy nhất	Not null	Khóa ngoại	Ràng buộc toàn vẹn	Diễn giải
1	Id	varchar	12						x		x			ID
2	Name	varchar	100								x			Tên
3	Color	varchar	7		#12B7BC		7	7			x			Màu
4	Deleted	boolean			False						x			Đã xóa

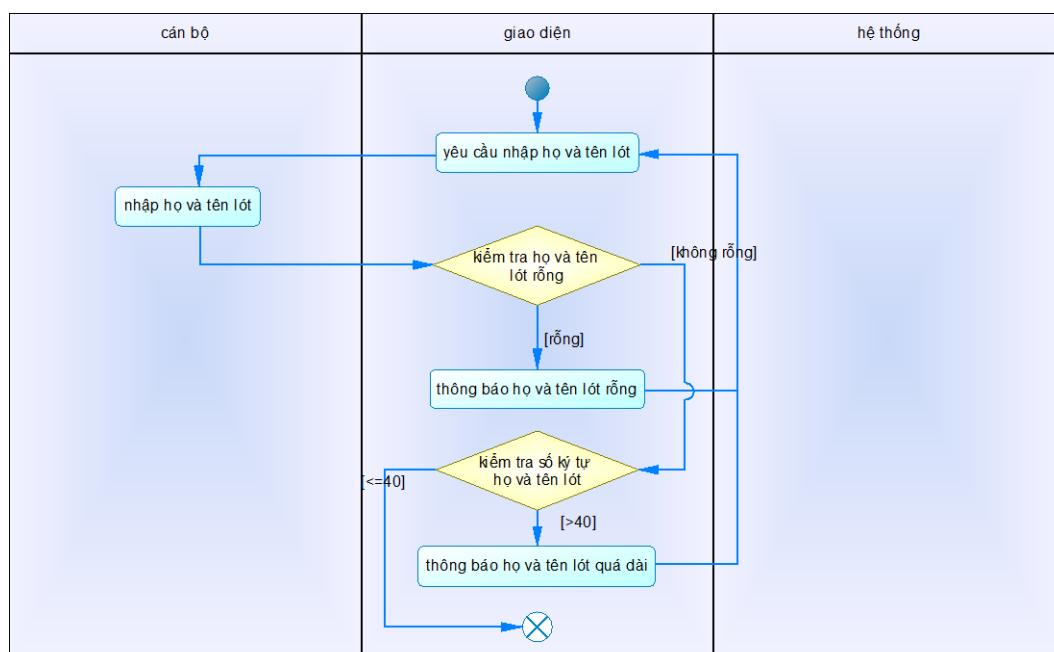
Ứng dụng optical character recognition vào hệ thống quản lý công văn

Tên bảng	Right													
Mô tả tên bảng	Quyền cán bộ													
Thuộc tính														
STT	Tên thuộc tính	Kiểu	Kích thước	Số thập phân	Trị mặc định	Miền giá trị	Min	Max	Khóa chính	Duy nhất	Not null	Khóa ngoại	Ràng buộc toàn vẹn	Diễn giải
1	Id	varchar	12						x		x			ID
2	Name	varchar	100								x			Tên
3	ReadOD	varchar	7		#12B7BC		7	7			x			Màu
4	Deleted	boolean			False						x			Đã xóa

PHỤ LỤC III SƠ ĐỒ HOẠT ĐỘNG CON

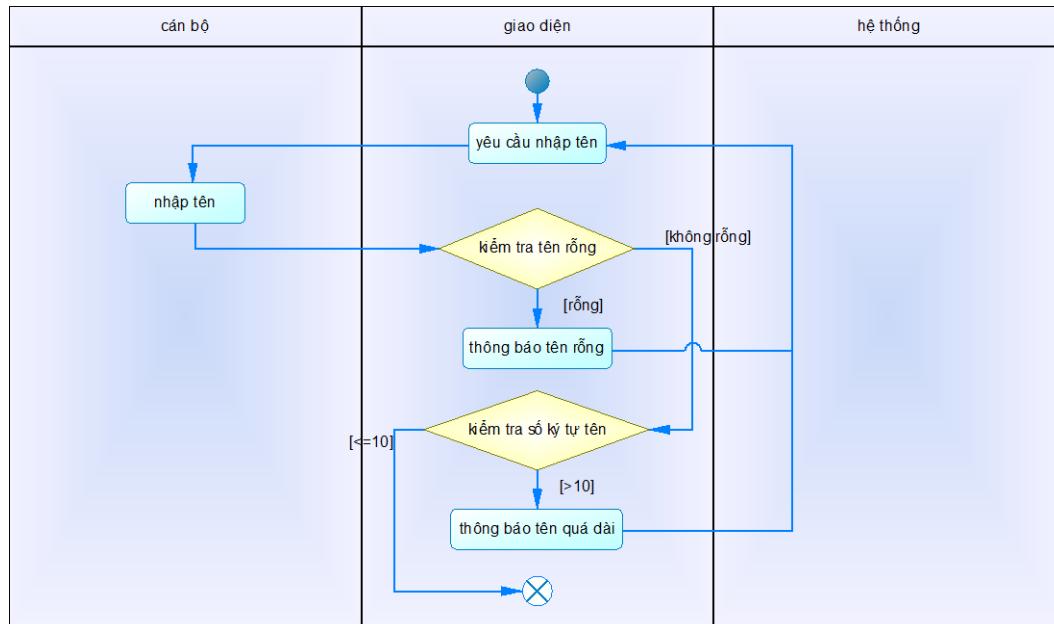


Sơ đồ III.1 Sơ đồ hoạt động lấy mã số cán bộ mới

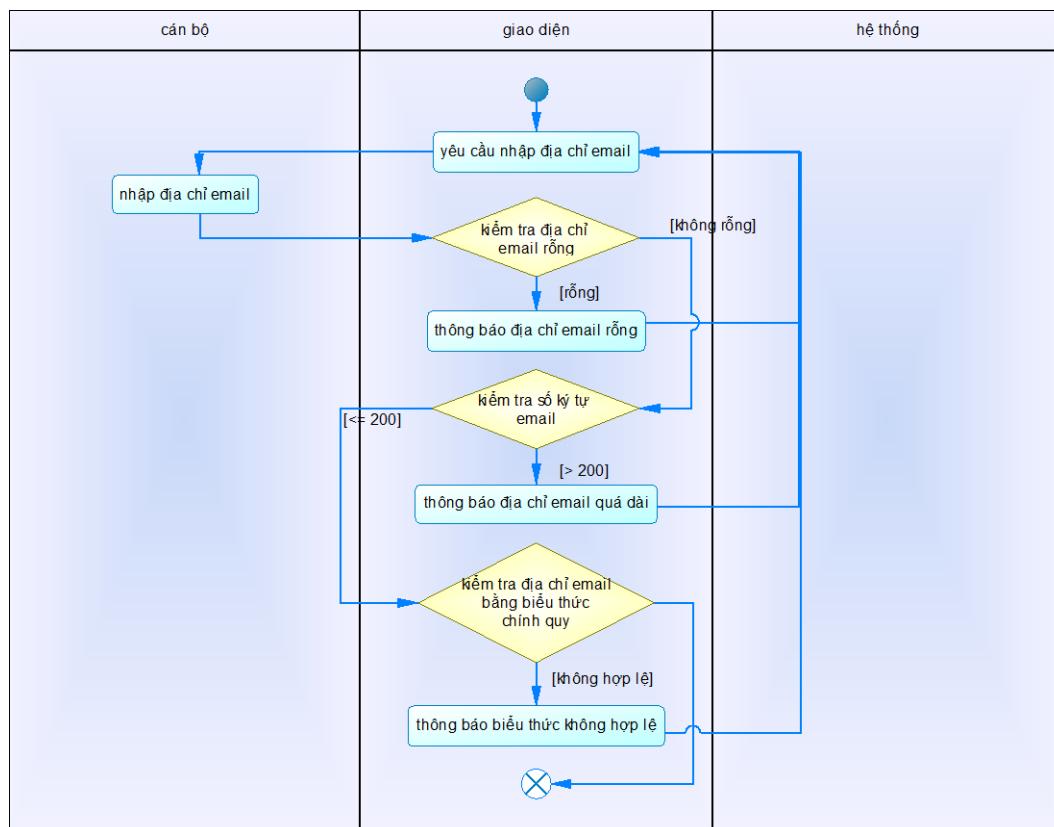


Sơ đồ III.2 Sơ đồ hoạt động nhập họ và tên lót cán bộ

Ứng dụng optical character recognition vào hệ thống quản lý công văn

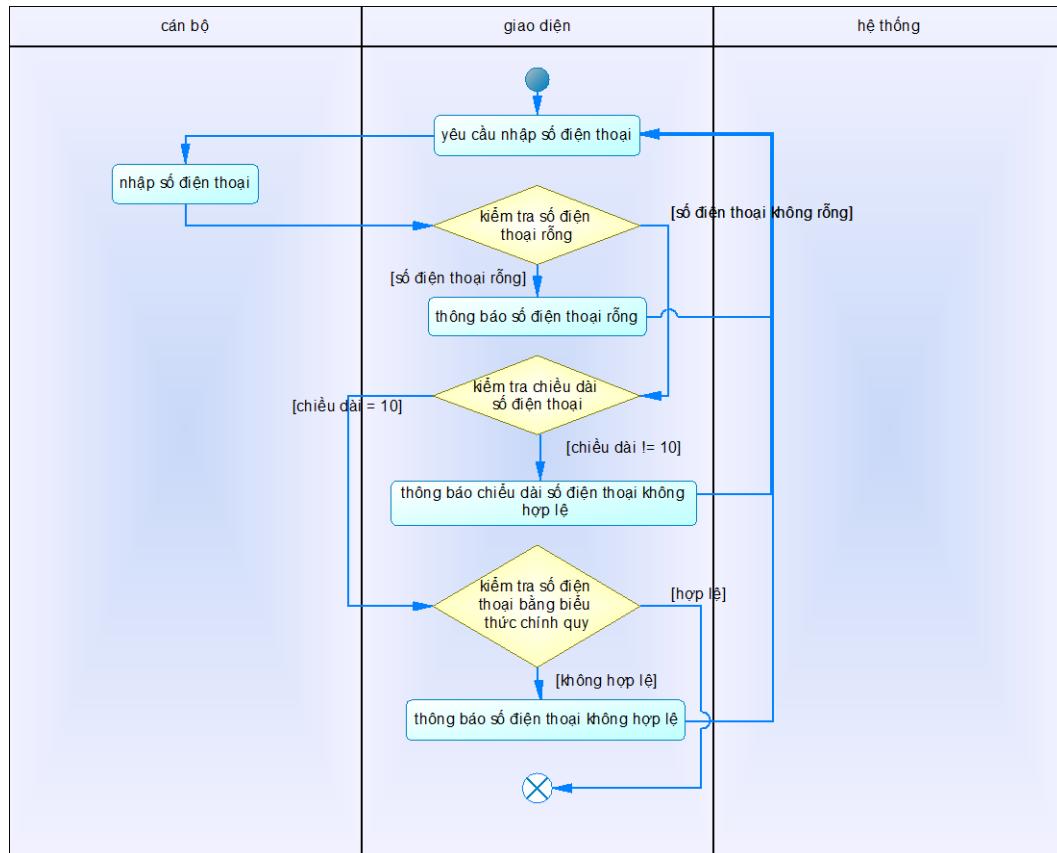


Sơ đồ III.3 Sơ đồ hoạt động nhập tên cán bộ

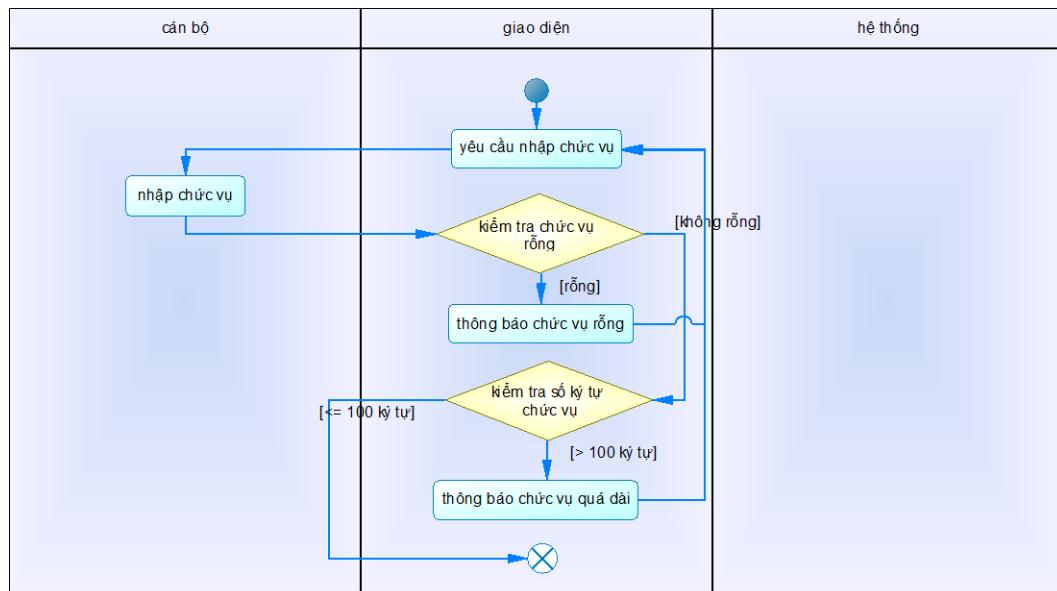


Sơ đồ III.4 Sơ đồ hoạt động nhập địa chỉ email

Ứng dụng optical character recognition vào hệ thống quản lý công văn

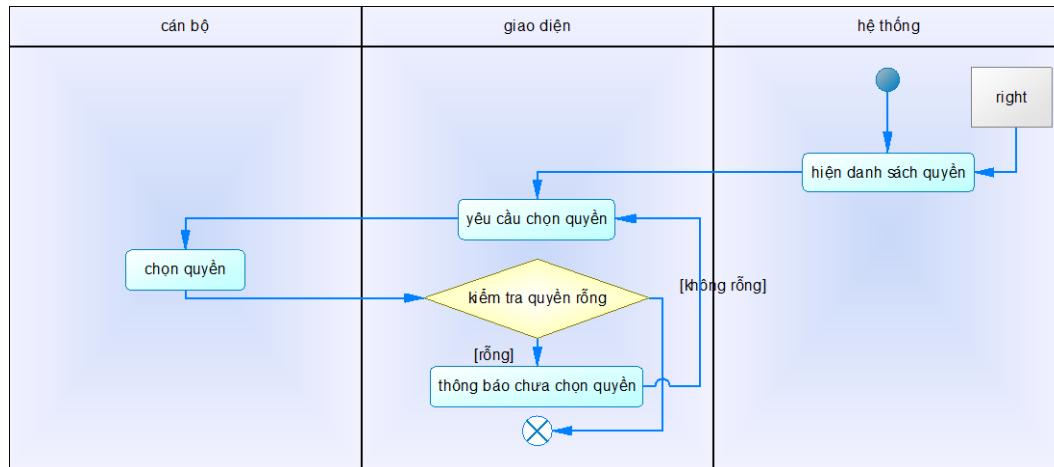


Sơ đồ III.5 Sơ đồ hoạt động nhập số điện thoại

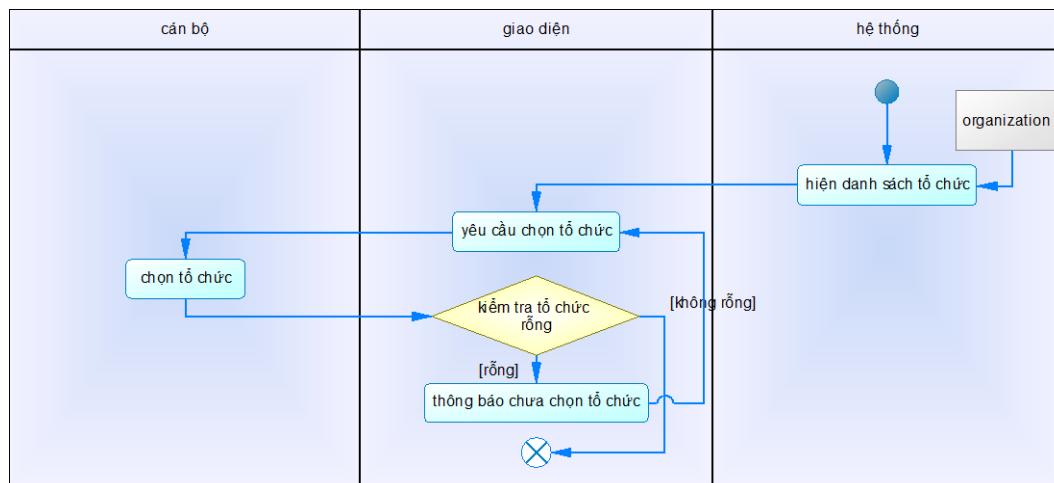


Sơ đồ III.6 Sơ đồ hoạt động nhập chức vụ

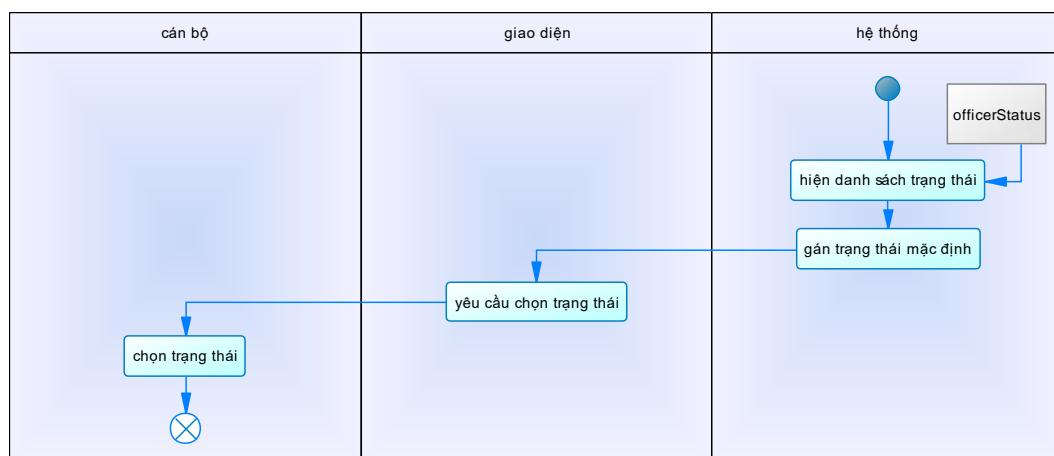
Ứng dụng optical character recognition vào hệ thống quản lý công văn



Sơ đồ III.7 Sơ đồ hoạt động chọn quyền

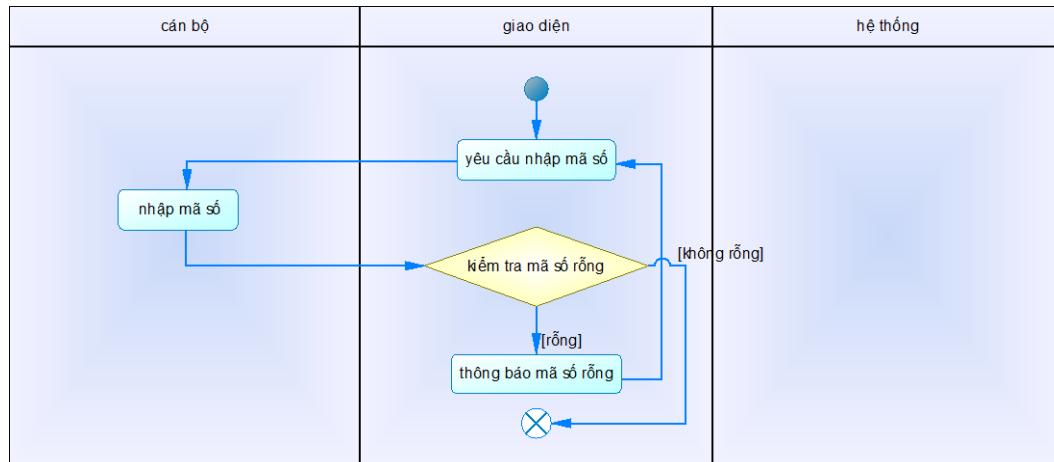


Sơ đồ III.8 Sơ đồ hoạt động chọn tổ chức

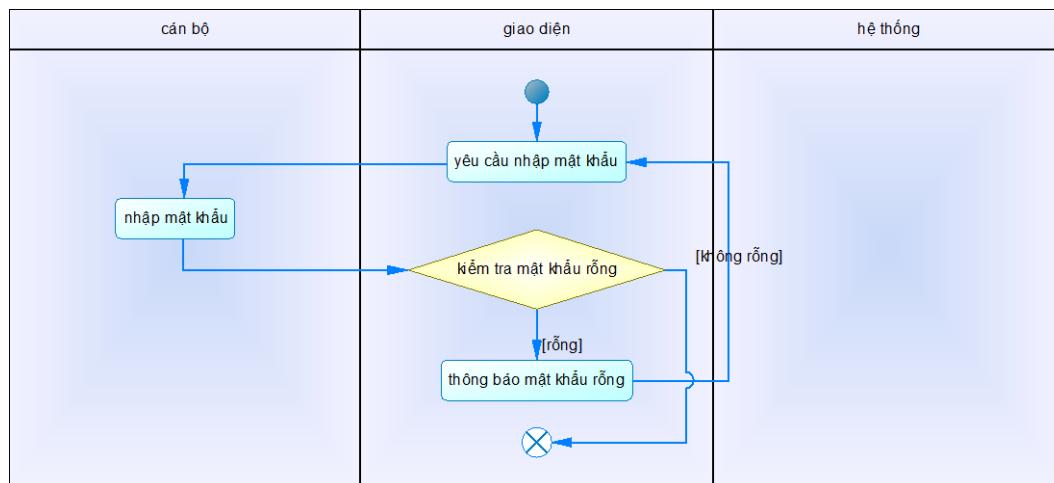


Sơ đồ III.9 Sơ đồ hoạt động chọn trạng thái cán bộ

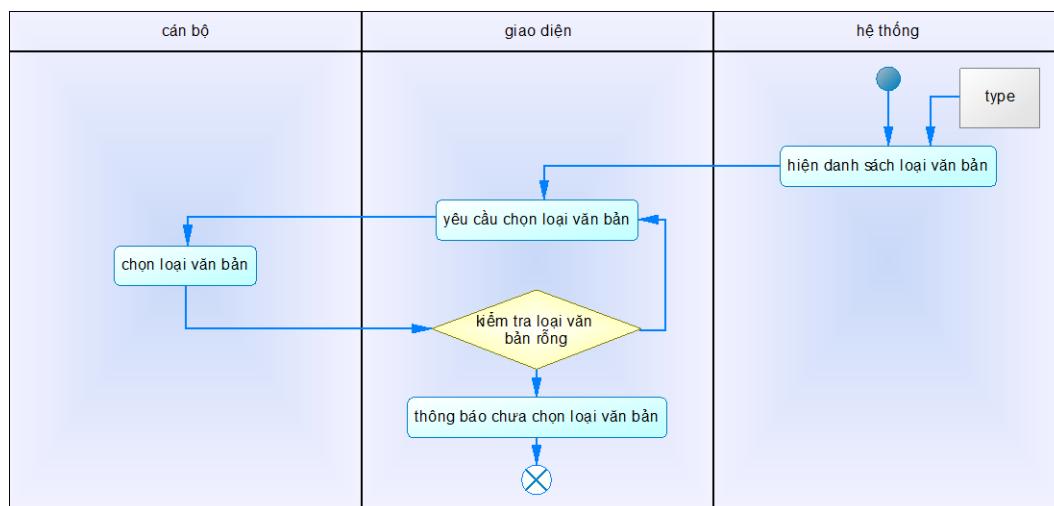
Ứng dụng optical character recognition vào hệ thống quản lý công văn



Sơ đồ III.10 Sơ đồ hoạt động nhập mã số đăng nhập

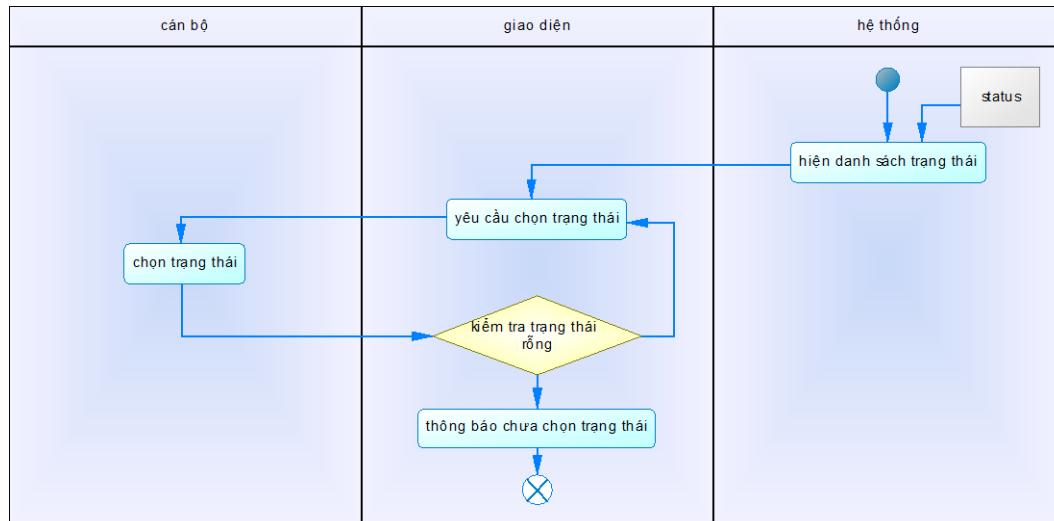


Sơ đồ III.11 Sơ đồ hoạt động nhập mật khẩu

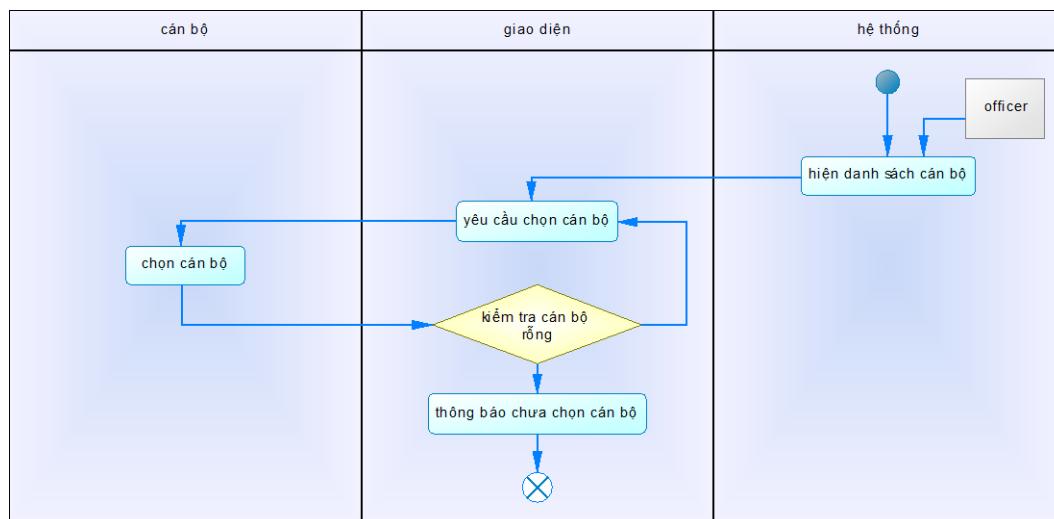


Sơ đồ III.12 Sơ đồ hoạt động chọn loại văn bản

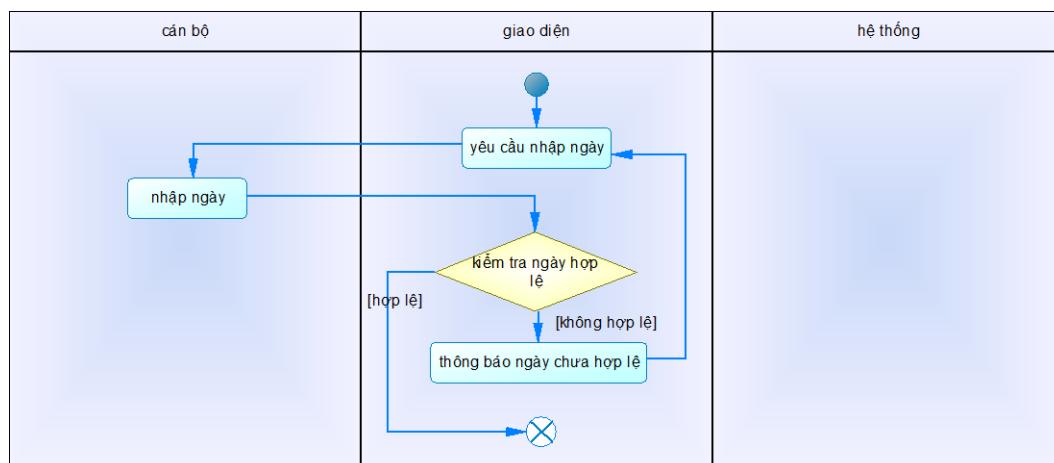
Ứng dụng optical character recognition vào hệ thống quản lý công văn



Sơ đồ III.13 Sơ đồ hoạt động chọn trạng thái

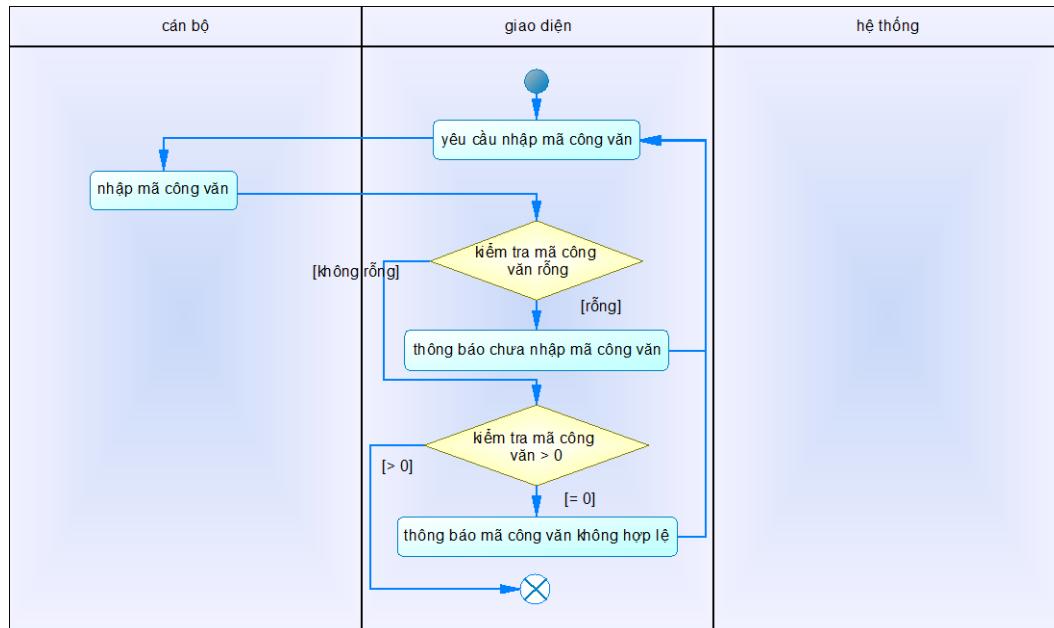


Sơ đồ III.14 Sơ đồ hoạt động chọn cán bộ

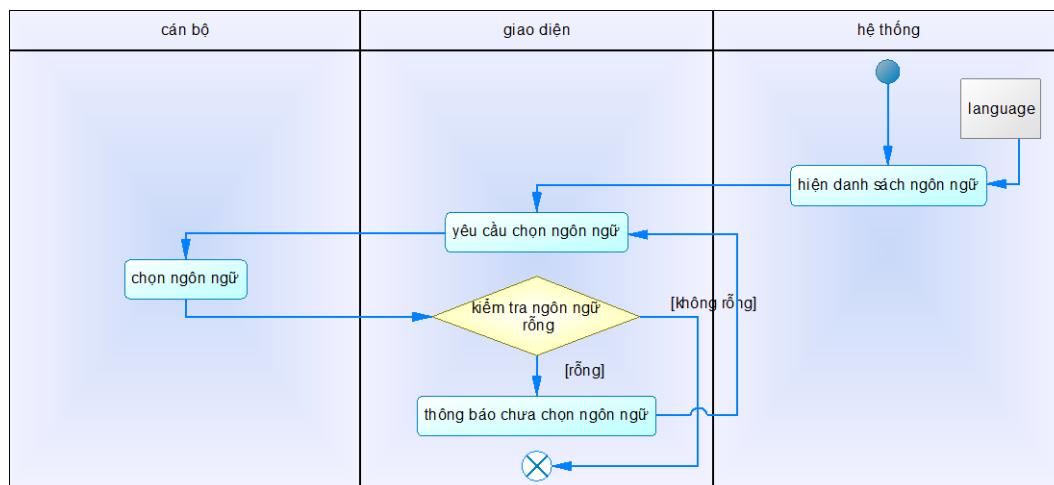


Sơ đồ III.15 Sơ đồ hoạt động chọn ngày

Ứng dụng optical character recognition vào hệ thống quản lý công văn

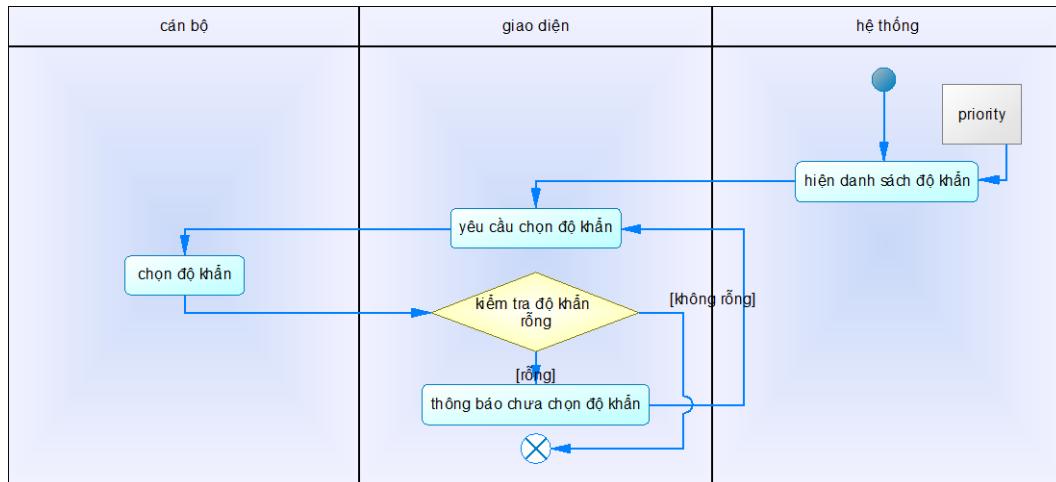


Sơ đồ III.16 Sơ đồ hoạt động nhập mã công văn



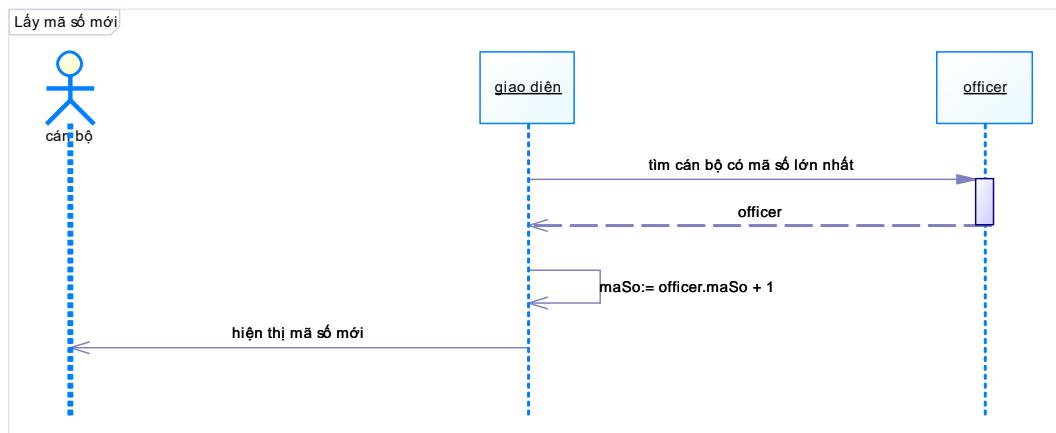
Sơ đồ III.17 Sơ đồ hoạt động chọn ngôn ngữ

Ứng dụng optical character recognition vào hệ thống quản lý công văn

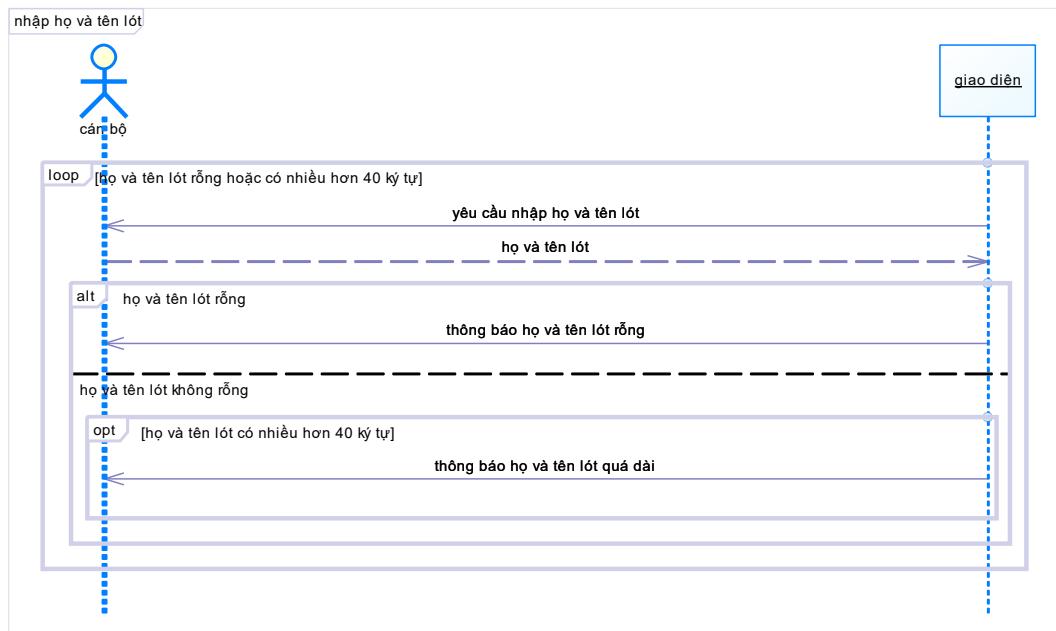


Sơ đồ III.18 Sơ đồ hoạt động chọn độ khẩn

PHỤ LỤC IV SƠ ĐỒ TUẦN TỤ CON

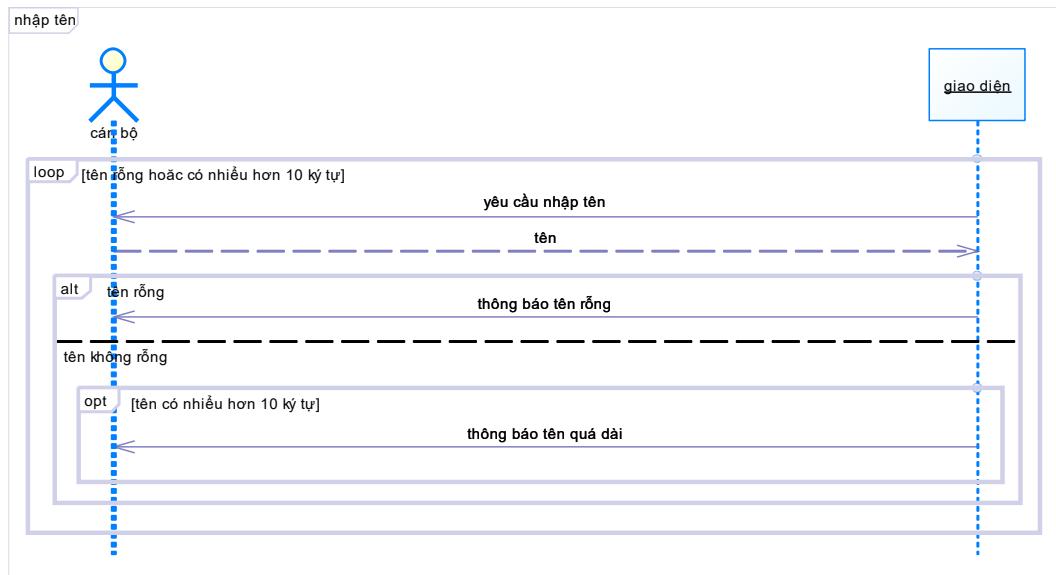


Sơ đồ III.1 Sơ đồ tuần tự lấy mã số mới



Sơ đồ III.2 Sơ đồ tuần tự nhập họ và tên lót

Ứng dụng optical character recognition vào hệ thống quản lý công văn

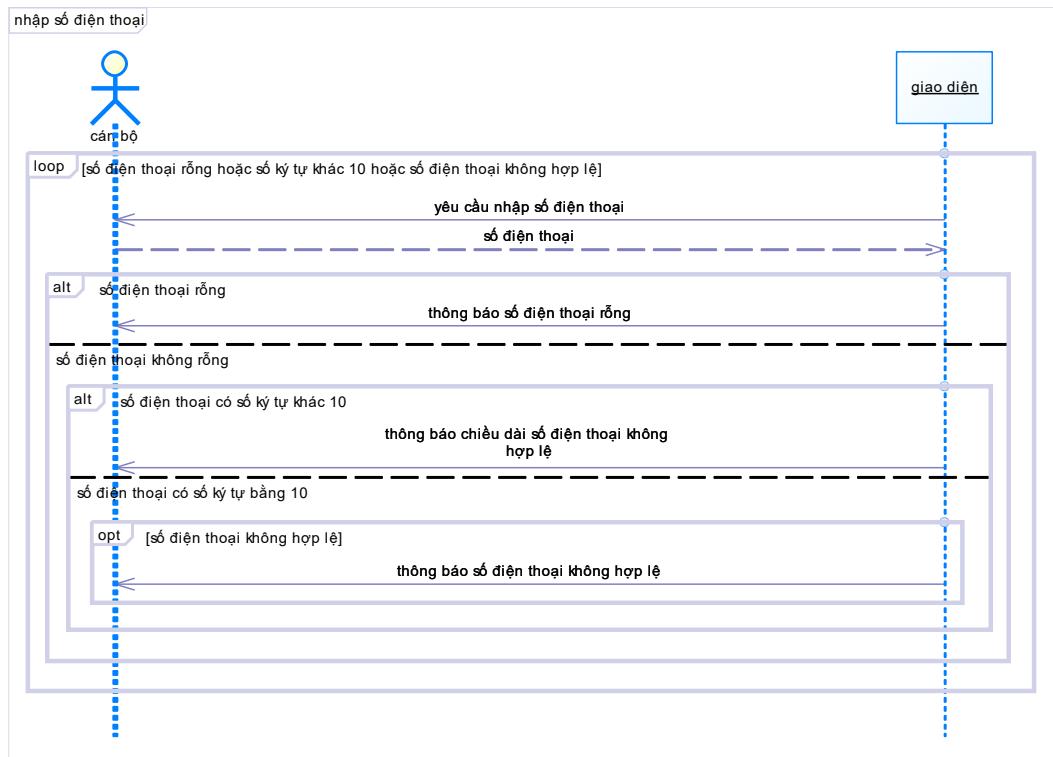


Sơ đồ III.3 Sơ đồ tuần tự nhập tên

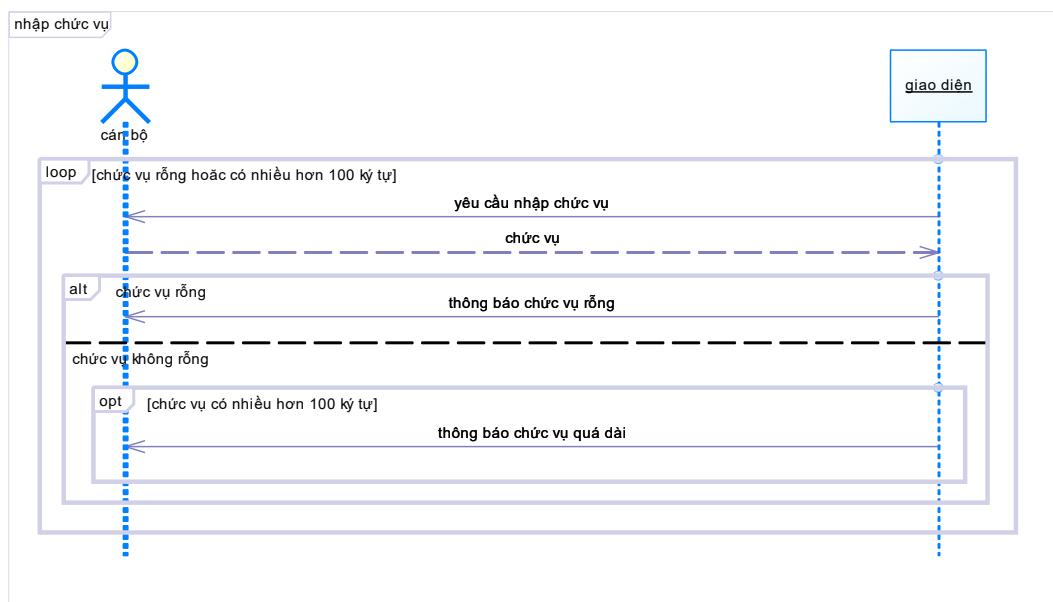


Sơ đồ III.4 Sơ đồ tuần tự nhập địa chỉ email

Ứng dụng optical character recognition vào hệ thống quản lý công văn

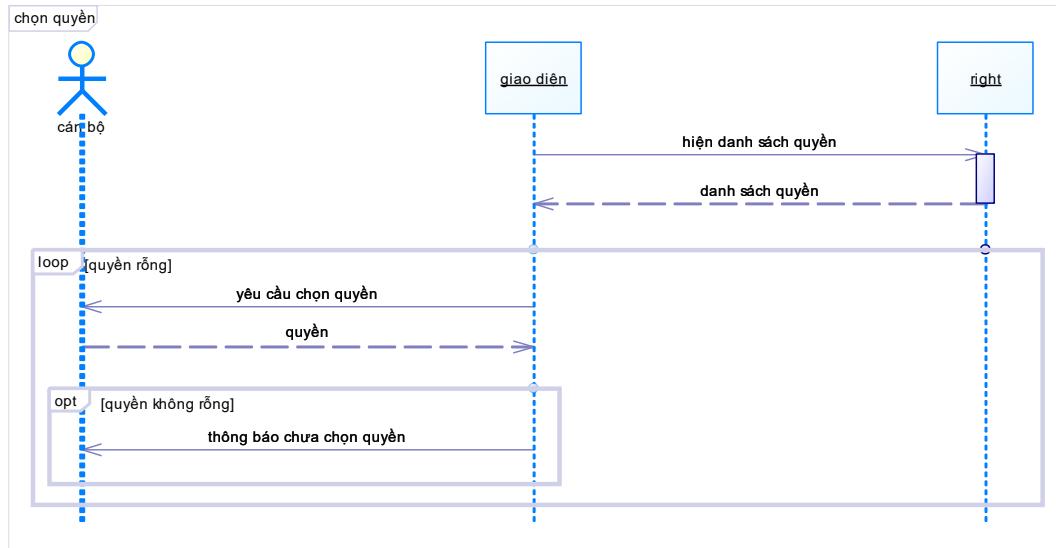


Sơ đồ III.5 Sơ đồ tuần tự nhập số điện thoại

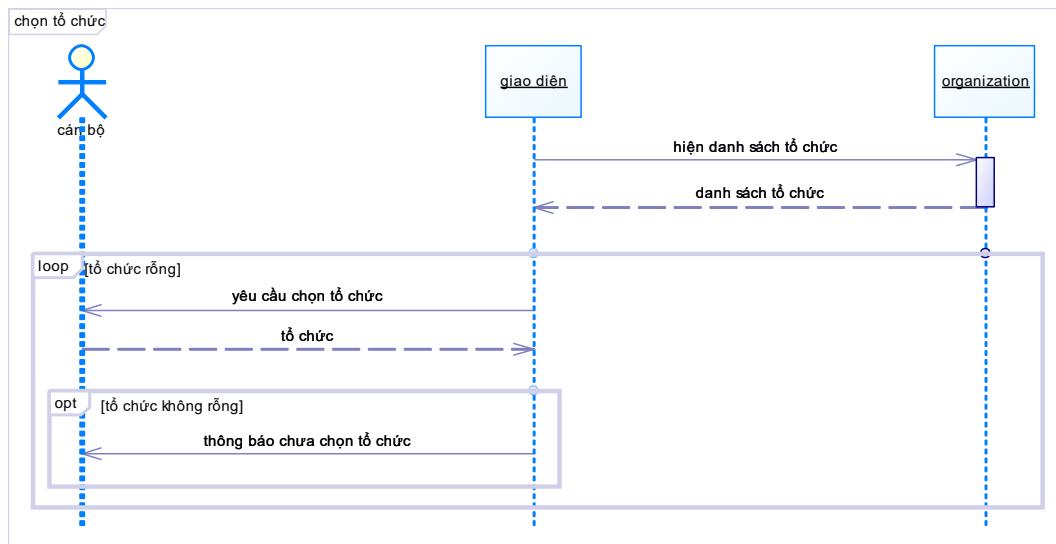


Sơ đồ III.6 Sơ đồ tuần tự nhập chức vụ

Ứng dụng optical character recognition vào hệ thống quản lý công văn

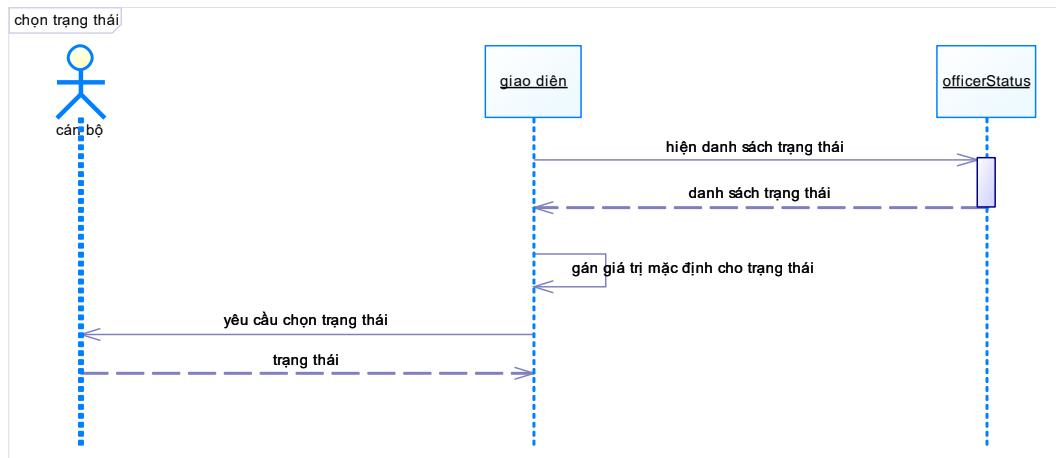


Sơ đồ III.7 Sơ đồ tuần tự chọn quyền

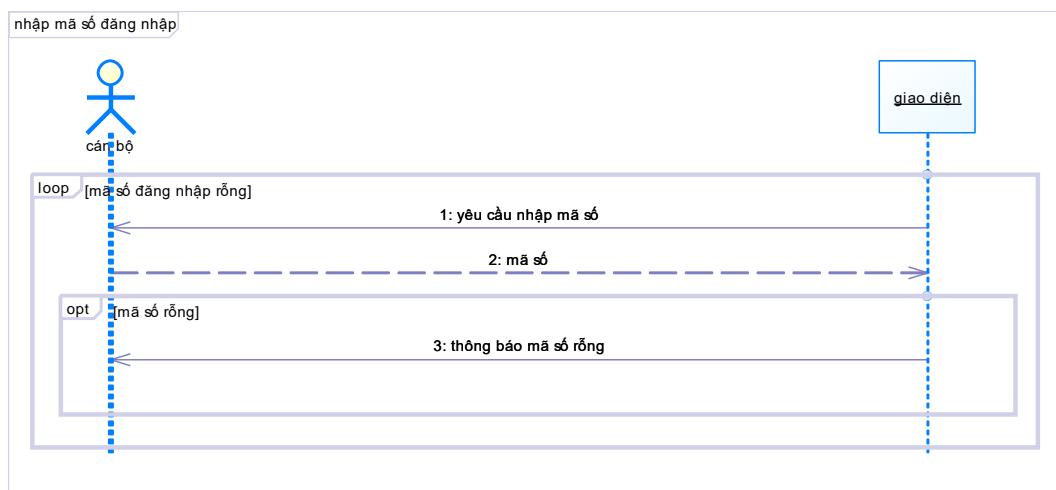


Sơ đồ III.8 Sơ đồ tuần tự chọn tổ chức

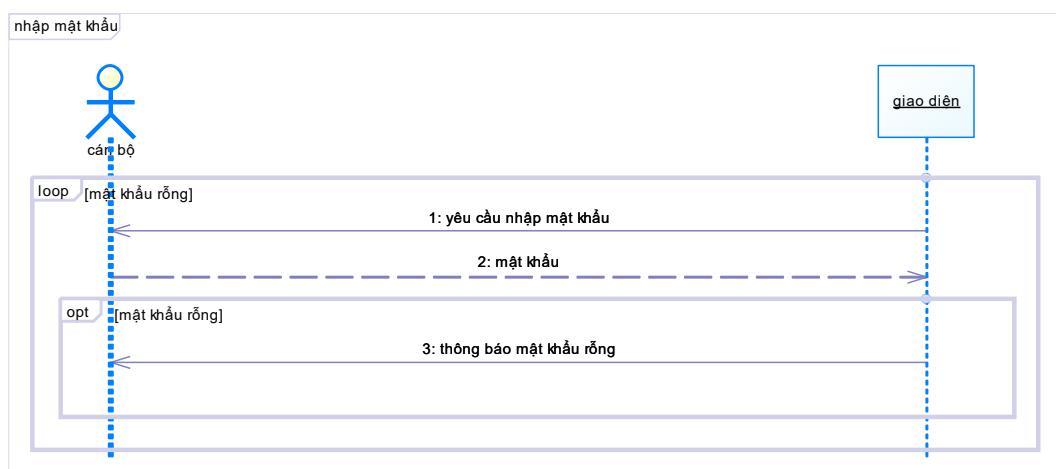
Ứng dụng optical character recognition vào hệ thống quản lý công văn



Sơ đồ III.9 Sơ đồ tuần tự chọn trạng thái



Sơ đồ III.10 Sơ đồ tuần tự nhập mã số đăng nhập



Sơ đồ III.11 Sơ đồ tuần tự nhập mật khẩu