

**ĐẠI HỌC THÁI NGUYÊN**  
**TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN & TRUYỀN THÔNG**

---

**NGÔ MINH HIẾU**

**TÌM HIỂU PHƯƠNG PHÁP ĐÁNH GIÁ ĐỘ CHÍNH XÁC  
CỦA CÁC HỆ THỐNG NHẬN DẠNG CHỮ VIỆT**

**LUẬN VĂN THẠC SĨ KHOA HỌC MÁY TÍNH**

**Thái Nguyên 2015**

**ĐẠI HỌC THÁI NGUYÊN**  
**TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN & TRUYỀN THÔNG**

**NGÔ MINH HIẾU**

**TÌM HIỂU PHƯƠNG PHÁP ĐÁNH GIÁ ĐỘ CHÍNH XÁC  
CỦA CÁC HỆ THỐNG NHẬN DẠNG CHỮ VIỆT**

**Chuyên ngành:** Khoa học máy tính  
**Mã số:** 60 48 01 01

**LUẬN VĂN THẠC SĨ KHOA HỌC MÁY TÍNH**

**NGƯỜI HƯỚNG DẪN KHOA HỌC**  
**TS. NGUYỄN THỊ THANH TÂN**

**Thái Nguyên 2015**

## **LỜI CAM ĐOAN**

Tôi xin cam đoan rằng bản luận văn này là tự thân nghiên cứu và hoàn thành dưới sự hướng dẫn khoa học của TS. Nguyễn Thị Thanh Tân. Nếu có gì vi phạm tôi xin hoàn toàn chịu trách nhiệm.

*Thái Nguyên, ngày tháng năm 2015*

**Ngô Minh Hiếu**

## LỜI CẢM ƠN

Lời đầu tiên tôi xin gửi lời cảm ơn chân thành và lòng biết ơn sâu sắc tới TS Nguyễn Thị Thanh Tân, người đã chỉ bảo và hướng dẫn tận tình cho tôi và đóng góp ý kiến quý báu trong suốt quá trình học tập, nghiên cứu và thực hiện luận văn này.

Tôi xin trân trọng cảm ơn Ban giám hiệu Trường Đại học Công nghệ Thông tin và Truyền thông, Đại học Thái Nguyên, khoa CNTT đã giúp đỡ và tạo các điều kiện cho chúng tôi được học tập và làm khóa luận một cách thuận lợi.

Và cuối cùng tôi xin gửi lời cảm ơn đến gia đình, người thân và bạn bè, những người luôn bên tôi và là chỗ dựa giúp cho tôi vượt qua những khó khăn nhất. Họ luôn động viên tôi khuyến khích và giúp đỡ tôi trong cuộc sống và công việc cho tôi quyết tâm hoàn thành luận văn này.

Tuy nhiên do thời gian có hạn, mặc dù đã nỗ lực cố gắng hết mình nhưng chắc rằng luận văn khó tránh khỏi những thiếu sót. Rất mong được sự chỉ bảo, góp ý tận tình của quý Thầy Cô và các bạn.

*Tôi xin chân thành cảm ơn!*

*Thái Nguyên, ngày tháng năm 2015*

**Ngô Minh Hiếu**

## MỤC LỤC

<b>LỜI CAM ĐOAN</b> .....	<b>1</b>
<b>LỜI CẢM ƠN</b> .....	<b>2</b>
<b>MỤC LỤC</b> .....	<b>3</b>
<b>HÌNH VẼ</b> .....	<b>5</b>
<b>BẢNG</b> .....	<b>6</b>
<b>DANH MỤC CÁC TỪ VIẾT TẮT</b> .....	<b>7</b>
<b>MỞ ĐẦU</b> .....	<b>8</b>
<b>CHƯƠNG 1 - TỔNG QUAN VỀ NHẬN DẠNG CHỮ</b> .....	<b>12</b>
<b>1.1. Qui trình chung của một hệ nhận dạng chữ</b> .....	<b>12</b>
1.1.1. Phân lớp mẫu .....	12
1.1.2. Nhận dạng văn bản .....	13
<b>1.2. Tìm hiểu một số phần mềm nhận dạng chữ</b> .....	<b>16</b>
1.2.1. VnDOCR .....	16
1.2.2. FineReader .....	18
1.2.3. OmniPage.....	20
1.2.4. VietOCR .....	20
<b>1.3. Những vấn đề ảnh hưởng tới chất lượng của một phần mềm nhận dạng</b> .....	<b>22</b>
1.3.1. Chữ bị dính, nhòe .....	23
1.3.2. Văn bản bị đứt hoặc mất nét .....	24
1.3.3. Văn bản bị nhiễu .....	25
1.3.4. Văn bản được in với các kiểu font chữ đặc biệt .....	26
1.3.5. Cỡ chữ quá lớn hoặc quá nhỏ .....	26
<b>1.4. Kết luận</b> .....	<b>27</b>
<b>CHƯƠNG 2 - PHƯƠNG PHÁP ĐÁNH GIÁ HIỆU QUẢ CỦA CÁC THUẬT TOÁN NHẬN DẠNG CHỮ VIỆT</b> .....	<b>28</b>
2.1. Một số khái niệm .....	28
2.2. Bài toán hiệu chỉnh chuỗi ký tự (string editing) .....	29
2.3. Thuật toán Ukkonen.....	34

2.4. Đánh giá độ chính xác mức ký tự.....	40
2.5. Đánh giá độ chính xác mức ký tự theo lớp mẫu .....	44
2.6. Hiệu quả của các ký tự đánh dấu.....	44
2.7. Độ chính xác mức từ.....	46
<b>CHƯƠNG 3 :THỰC NGHIỆM VÀ ĐÁNH GIÁ KẾT QUẢ.....</b>	<b>51</b>
<b>3.1.Phân tích, cài đặt chương trình.....</b>	<b>51</b>
3.1.1. Quy trình thực hiện .....	51
3.1.2. Các cấu trúc dữ liệu.....	52
3.1.3. Danh sách các từ dừng trong tiếng Việt.....	54
3.1.4 Danh sách các ký tự đặc biệt .....	55
3.1.5. Module đánh giá độ chính xác mức ký tự.....	56
3.1.6. Module đánh giá độ chính xác mức từ.....	58
<b>3.2.Đánh giá thực nghiệm.....</b>	<b>65</b>
3.2.1 Dữ liệu thực nghiệm.....	65
3.2.2 Kết quả thực nghiệm .....	68
<b>3.3.Kết luận chương 3.....</b>	<b>70</b>
<b>KẾT LUẬN.....</b>	<b>71</b>
<b>DANH MỤC TÀI LIỆU THAM KHẢO.....</b>	<b>72</b>

## HÌNH VẼ

Hình 1.1: Quy trình chung của một hệ thống nhận dạng chữ .....	15
Hình 1.2. Màn hình làm việc của VnDOCR .....	17
Hình 1.3. Màn hình kết quả phân tích và nhận dạng ảnh hình 1.7 .....	18
Hình 1.4 Màn hình làm việc của OmniPage .....	20
Hình 1.5 Màn hình làm việc của VietOCR .....	21
Hình 1.6 Trường hợp văn bản in đậm.....	23
Hình 1.7: Một số hình ảnh bị biến dạng của các ký tự .....	23
Hình 1.8 Hình ảnh các ký tự tiếng Việt bị nhập nhầm phân dấu .....	24
Hình 1.9 Trường hợp văn bản bị đứt và mất nét .....	24
Hình 1.10 Hình ảnh của ký tự bị biến dạng do lỗi đứt nét.....	24
Hình 1.11 Một số dạng nhiễu thường gặp trên văn bản.....	25
Hình 1.12 Văn bản bị các nhiễu đánh dấu .....	25
Hình 1.13 Văn bản bị nhiễu do bị chồng chữ ký/con dấu .....	26
Hình 1.14 Văn bản được in với kiểu font chữ đặc biệt.....	26
Hình 2.1: Đồ thị $G(A,B)$ , với $A = zxy$ và $B = xyxz$ .....	32
Hình 2.2: Các đường đi trên đồ thị $G(A, B)$ .....	33
Hình 2.3: Sự tương ứng giữa chuỗi văn bản nhận dạng và văn bản mẫu.....	42
Hình 2.4: Độ chính xác mức từ.....	48
Hình 3.1 Quy trình thực hiện của chương trình .....	52
Hình 3.2: Kết quả đánh giá độ chính xác mức ký tự trên một văn bản tiếng Anh .....	61
Hình 3.3: Đánh giá độ chính xác mức từ trên 1 file văn bản tiếng Anh.....	65

## **BẢNG**

Bảng 2.1: Giải thuật cho bài toán chỉnh sửa chuỗi.....	33
Bảng 2.2: Độ chính xác mức ký tự .....	43
Bảng 3.1 Bảng danh sách các từ dùng trong tiếng Việt.....	55
Bảng 3.2 Thông tin các thao tác hiệu chỉnh .....	57
Bảng 3.3 Thông tin về đánh giá độ chính xác mức ký tự .....	57
Bảng 3.4: Các tập dữ liệu tiếng Anh.....	66
Bảng 3.5: Các tập dữ liệu Tiếng Việt.....	67
Bảng 3.6: Độ chính xác mức ký tự trên tập dữ liệu tiếng Anh .....	68
Bảng 3.7: Độ chính xác mức ký tự trên các tập dữ liệu tiếng Việt .....	69
Bảng 3.8: Độ chính xác mức từ trên tập dữ liệu tiếng Anh .....	69
Bảng 3.9: Độ chính xác mức từ tập dữ liệu tiếng Việt .....	69



## DANH MỤC CÁC TỪ VIẾT TẮT

<b>STT</b>	<b>Từ viết tắt</b>	<b>Ý nghĩa</b>	<b>Nội dung</b>
1	NLP	Natural Language Processing	Xử lý ngôn ngữ tự nhiên
2	LCS	Longest common subsequence	Dãy chung dài nhất
3	OCR	Optical Character Recognition	Nhận dạng ký tự quang học

## MỞ ĐẦU

### 1. Tính cấp thiết của luận văn

Nhận dạng mẫu là một ngành khoa học mà vai trò của nó là phân lớp các đối tượng thành một số loại hoặc một số lớp riêng biệt. Tùy thuộc vào lĩnh vực ứng dụng, các đối tượng có thể ở dạng ảnh, dạng tín hiệu sóng hoặc một kiểu dữ liệu bất kỳ nào đó mà cần phải phân lớp. Những đối tượng này được gọi bằng một thuật ngữ chung đó là “mẫu” (pattern). Nhận dạng mẫu đã được biết đến từ rất lâu, nhưng trước những năm 1960 nó hầu như chỉ là kết quả nghiên cứu về mặt lý thuyết trong lĩnh vực thống kê. Tuy nhiên, với sự phát triển không ngừng của khoa học kỹ thuật về phần cứng cũng như phần mềm, các yêu cầu về mặt ứng dụng thực tế của lĩnh vực nhận dạng mẫu ngày càng tăng lên và hiện nay nhận dạng mẫu đã được sử dụng trong rất nhiều lĩnh vực như y học, tự động hoá một số qui trình sản xuất công nghiệp, dự báo thời tiết, dự báo cháy rừng, v.v. Ngoài ra nhận dạng mẫu còn là thành phần quan trọng trong hầu hết các hệ thống máy tính thông minh được xây dựng để thực hiện việc ra quyết định.

Cùng với sự phát triển của nhận dạng mẫu, nhận dạng chữ đã và đang ngày càng trở thành một ứng dụng không thể thiếu được trong đời sống xã hội của con người. Nhận dạng chữ là quá trình chuyển đổi từ dạng hình ảnh của một hay nhiều trang ảnh chứa các thông tin văn bản thành tệp văn bản thực sự có thể soạn thảo được trên máy tính. Ngoài ứng dụng số hóa các trang văn bản, tài liệu, hiện tại nhận dạng chữ còn được ứng dụng rộng rãi trong các hoạt động giao dịch hàng ngày và qui trình tự động hóa các công việc văn phòng, chẳng hạn như nhập liệu tự động phiếu chấm thi trắc nghiệm, phiếu điều tra, nhận dạng các dòng địa chỉ trên phong bì thư, nhận dạng nhãn sản phẩm, nhận dạng thông tin cá nhân trên chứng minh nhân, hộ chiếu, card visit, v.v.