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

HÌNH VẼ

BẢNG

DANH MỤC CÁC TỪ VIẾT TẮT

STT	Từ viết tắt	Ý nghĩa	Nội dung
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 dạngmẫu 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.Tuỳ 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 bảnthành 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.