ĐẠI HỌC THÁI NGUYÊN
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN VÀTRUYỀN THÔNG

TRƯƠNG TUẤN TOÀN

NGHIÊN CỨU PHƯƠNG PHÁP NHẬN DẠNG CHỮ VIỆT IN CHẤT LƯỢNG THẤP

LUẬN VĂN THẠC SĨ KHOA HỌC MÁY TÍNH

Thái Nguyên 2014

ĐẠI HỌC THÁI NGUYÊN
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

TRƯƠNG TUẤN TOÀN

NGHIÊN CỨU PHƯƠNG PHÁP NHẬN DẠNG CHỮ VIỆT IN CHẤT LƯỢNG THẤP

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

LUẬN VĂN THẠC SĨ KHOA HỌC MÁY TÍNH

NGƯỜI HƯỚNG DẪN KHOA HỌC
TS. NGUYỄN THỊ THANH TÂN

Thái Nguyên 2014

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 2014

Trương Tuấn Toàn

LỜI CẢM ƠN

Em xin bày tỏ lòng biết ơn sâu sắc tới TS. Nguyễn Thị Thanh Tân, cô đã hướng dẫn, chỉ dạy tận tình để em hoàn thành luận văn này.

Em xin chân thành cảm ơn các thầy cô giáo trong trường Đại học Công nghệ thông tin và truyền thông – Đại học Thái Nguyên, các thầy cô giáo tại viện CNTT Hà Nội đã truyền thụ kiến thức cho em trong suốt quá trình học tập vừa qua.

Cuối cùng xin cảm ơn gia đình, cảm ơn các bạn đã cùng chia sẻ, giúp đỡ, động viên trong suốt quá trình học tập cũng như trong thời gian thực hiện luận văn.

Thái Nguyên, ngày tháng năm 2014

Trương Tuấn Toàn

MỤC LỤC

HÌNH VẼ

BẢNG

MỞ ĐẦU

1. Tính cấp thiết của luận văn

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. Khi đề cập đến bài toán nhận dạng chữ, người ta thường phân biệt hai loại là chữ in (optical character) và chữ viết tay (handwritten character) [2], [6], [7], [9]. Các kết quả ứng dụng của bài toán nhận dạng chữ in đã và đang được sử dụng rộng rãi trong qui trình tự động hóa các hoạt động văn phòng, mang lại lợi ích thực sự cho con người.

Ngày nay cùng với sự phát triển về mặt lý thuyết, công nghệ, có rất nhiều hướng đi cho việc giải quyết bài toán này như: Hiện tại có rất nhiều phương pháp phân lớp sử dụng trong nhận dạng chữ như: phân loại Bayes, K-láng giềng gần nhất (k-NN), mạng Neural (ANNs), mô hình Markov ẩn (HMM),… Những phương pháp này đã cho kết quả chấp nhận được và có nhiều ứng dụng trong thực tế.

Trên thế giới hiện nay có nhiều chương trình nhận dạng chữ viết (chữ in và viết tay), như các hệ OMNIPAGE, READ-WRITE, WORD-SCAN,... Ở Việt Nam cũng có một số hệ như WORC của công ty 3C, VIET-IN của công ty SEATIC, VNDOCR của Viện Công Nghệ Thông Tin, Image Scon của Trung Tâm Tự Động Hóa Thiết Kế, hệ WINGIS của công ty DolfSoft [2].

Đối với bài toán nhận dạng văn bản tiếng Việt, có thể nói cho đến thời điểm hiện tại, việc nhận dạng các văn chữ in bản chất lượng thấp vẫn là một vấn đề thách thức. Vì lý do đó, luận văn này sẽ tập trung nghiên cứu một số phương pháp phân lớp mẫu và trích chọn đặc trưng nhằm lựa chọn được một phương pháp thích hợp cho việc nhận dạng chữ Việt in chất lượng thấp, thử

nghiệm xây dựng chương trình nhận dạng ký tự chữ Việt một văn bản mà trong văn bản đó xuất hiện nhiều ký tự bị dính, biến dạng, bị đứt hay mất nét... với mong muốn sẽ làm ra một sản phẩm nhận dạng văn bản chữ in tiếng Việt chất lượng thấp hoàn chỉnh trong tương lai.

2. Mục tiêu của luận văn

Luận văn tập trung nghiên cứu một số phương pháp phân lớp mẫu và trích chọn đặc trưng nhằm lựa chọn được một phương pháp thích hợp cho việc nhận dạng các ảnh chữ cái và chữ số tiếng Việt chất lượng thấp.

Ảnh chất lượng thấp ở đây bao gồm các ảnh kí tự bị lỗi do nhiễu, do bị đứt nét, bị thiếu hoặc thừa ra một phần nào đó do dính vào kí tự bên cạnh, do các thành phần của kí tự bị dính với nhau chẳng hạn như phần mũ, dấu dính với phần chữ đối với kí tự tiếng Việt.

3. Bố cục của luận văn

Các nội dung trình bày trong luận văn được chia thành 3 chương:

Chương I: Tổng quan về bài toán nhận dạng chữ Việt.

Chương này trình bày tổng quan về các vấn đề liên quan đến nhận dạng, các đặc trưng của chữ Việt và chữ Việt chất lượng thấp, những vấn đề tồn tại trong bài toán nhận dạng ảnh văn bản chất lượng thấp, đưa ra mô hình chung của hệ thống nhận dạng, các hướng tiếp cận nhận dạng, các yếu tố ảnh hưởng đến hệ thống nhận dạng.

Chương II: Một số vấn đề trong nhận dạng ký tự chất lượng thấp

Chương này trình bày những khái niệm cơ bản về ảnh ký tự chất lượng thấp, một số hướng tiếp cận trong phân lớp và trích chọn đặc trưng ký tự và lựa chọn một phương pháp nhận dạng ảnh ký tự chất lượng thấp.

Chương III: Thực nghiệm và đánh giá kết quả