

I H C THÁI NGUYÊN
TR NG I H C CÔNG NGH THÔNG TIN VÀ TRUY N THÔNG

TR NG TU N TOÀN

NGHIÊN C U PH NG PHÁP NH N D NG
CH VI TIN CH TL NG TH P

LU N V N TH C S KHOA H C MÁY TÍNH

Thái Nguyên 2014

I H C THÁI NGUYÊN
TR NG I H C CÔNG NGH THÔNG TIN VÀ TRUY N THÔNG

TR NG TU N TOÀN

NGHIÊN C U PH NG PHÁP NH N D NG CH VI T
IN CH T L NG TH P

Chuyên ngành: Khoa h c máy tính
Mã s : 60 48 01

LU N V N TH C S KHOA H C MÁY TÍNH

NG I H NG D N KHOA H C
TS. NGUY N TH THANH TÂN

Thái Nguyên 2014

L I CAM OAN

Tôi xin cam oan r ng b n lu n v n này là t thân nghiên c u và hoàn thành d i s h ng d n khoa h c c a TS. Nguy n Th Thanh Tân. N u có gì vi ph m tôi xin hoàn toàn ch u trách nhi m.

Thái Nguyên, ngày tháng n m 2014

Tr ng Tu n Toàn

L I C M N

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, đồng hành trong suốt quá trình học tập cũng như trong thời gian thực hiện luận văn.

Thái Nguyên, ngày tháng năm 2014

Trưởng Luận Văn

M C L C

L I C A M O A N.....	i
L I C M N.....	ii
M C L C	iii
H Ì N H V	v
B Ñ G.....	vi
M U	1
CH Ñ G 1 - T Ñ G Q U A N V B À I T O Á N N H Ñ D Ñ G CH V I T	4
1.1 Qui trình chung c a m t h ñ h ñ d ñ g ch	4
1.1.1 Phân l p m u.....	4
1.1.2 Nh ñ d ñ g v ñ b ñ	11
1.2 Ch V i t và các c tr ñ g c a ch V i t.....	14
1.2.1 B ñ g ch cái t i ñ g V i t.....	14
1.2.2 Các nguyên âm trong t i ñ g V i t.....	14
1.2.3 C u trúc thanh i u.....	15
1.3 Nh ñ g t ñ t i trong nh ñ d ñ g v ñ b ñ ch t l ñ g th p.....	16
1.3.1 Ch b d ñnh, ñhøe	17
1.3.2 V ñ b ñ b t h o c m t nét.....	18
1.3.3 V ñ b ñ b ñ h i u.....	19
1.3.4 V ñ b ñ c i n v i các k i u font ch c b i t.....	20
1.3.5 C ch quá l ñ h o c quá ñh	21
1.4 K t lu ñ	22
CH Ñ G 2 - M T S V Ñ TR O N G N H Ñ D Ñ G K Ý T CH T L Ñ G	
TH P	23
2.1 Trích ch ñ c tr ñ g.....	24
2.1.1 Các c tr ñ g s d ñ g trong hu ñ l u y ñ ñ h ñnh	26
2.1.2 Các c tr ñ g s d ñ g trong quá trình ñ h ñ d ñ g	28
2.2 Nh ñ d ñ g ký t d a vào c tr ñ g trích ch ñ	29

2.2.1	Phân c m t p c tr ng.....	30
2.2.2	Thu t toán phân l p ký t	44
2.3	K t lu n	50
CH	NG 3 - TH C NGHI M VÀ ÁNH GIÁ K T QU	51
3.1	Bài toán	51
3.2	Cài t ch ng trình th nghi m.....	51
3.3	ánh giá th c nghi m	60
3.3.1	o ánh giá.....	60
3.3.2	D li u th c nghi m	61
3.3.3	K t qu th c nghi m.....	62
3.4	K t lu n	65
K	TLU N.....	67
I.	TÓM T T CÁC K T QU T CC ALU NV N.....	67
II.	NH NG V N CH A CGI IQUY TB ILU NV N.....	67
III.	H NG PHÁT TRI N.....	68
DANH M	C TÀI LI U THAM KH O	69

HÌNH V

Hình 1.1: Quy trình chung c a m t h th ng nh n d ng ch	11
Hình 1.2: Tr ng h p v n b n in m	17
Hình 1.3: M t s hình nh b bi n d ng c a các ký t	18
Hình 1.4: Hình nh các ký t ti ng Vi t b nh p nh ng ph n d u.....	18
Hình 1.5: Tr ng h p v n b n b t và m t nét.....	19
Hình 1.6: Hình nh c a ký t b bi n d ng do l i t nét.....	19
Hình 1.7: M t s d ng nhi u th ng g p trên v n b n.....	20
Hình 1.8: V n b n b các nhi u ánh d u.....	20
Hình 1.9: V n b n b nhi u do b ch ng ch ký/con d u.....	20
Hình 1.10: V n b n c in v i ki u font ch c bi t.....	21
Hình 2. 1: Các c tr ng hu n luy n mô hình	27
Hình 2.2: Trích ch n các c tr ng nh n d ng.....	29
Hình 2.3: c tr ng c a m t dòng nh	29
Hình 2.4: M t c u trúc cây K-D	33
Hình 2.5: C u trúc d li u l u các c tr ng u vào	34
Hình 2.6: C u trúc d li u cây K-D.....	35
Hình 2.7: C u trúc CLUSTER	36
Hình 2.8: C u trúc DIM_DESC mô t m i chi u c a cây K-D	37
Hình 2.9: M t s m u i di n cho l p ký t ‘ ’	44
Hình 2.10: Thu t toán phân l p ký t	46
Hình 2.11: K t qu th c hi n c a thu t toán	49
Hình 3.1: Quy trình th c hi n c a ch ng trình th nghi m	52
Hình 3.2: Các t p d li u th nghi m	62

B NG

B ng 1.1: C u trúc thanh i u trong ti ng Vi t	16
B ng 3.1: Các l p ký t hu n luy n thu t toán.....	53
B ng 3.2: K t qu th c nghi m	63

M U

1. Tính cấp thiết của luận văn

Nhận dạng chữ là quá trình chuyển đổi dạng hình ảnh của một hay nhiều trang chữ của các thông tin văn bản thành tập văn bản số có thể số hóa trên máy tính. Khi thực hiện 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 nghiên cứu của bài toán nhận dạng chữ in đã và đang được ứng dụng rộng rãi trong quá 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 nghiên cứu về các giải pháp bài toán này như: Hiện tại có rất nhiều phương pháp phân loại số dạng trong nhận dạng chữ như: phân loại Bayes, K-láng giêng 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,... 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 Scan của Trung Tâm Tự động Hóa Thiệt Kế, hệ WINGIS của công ty DolfSoft [2].

Trong bài toán nhận dạng văn bản tiếng Việt, có thể nói cho đến thời điểm hiện tại, việc nhận dạng các văn bản chữ in bằng thuật toán tập văn bản là một thách thức. Vì lý do đó, luận văn này sẽ tập trung nghiên cứu một số phương pháp phân loại mô-đun và trích chọn các từ ngữ nhằm lựa chọn một phương pháp thích hợp cho việc nhận dạng chữ viết in bằng thuật toán, th

nghi m xây d ñg ch ñg trình nh ñ ñg ký t ch Vi t m t v n b n mà trong v n b n ó xu t hi n nhi u ký t b ñính, bi ñ ñng, b t hay m t nét... v ñ i mong mu n s ñ làm ra m t s ñ ph m nh ñ ñg v n b n ch ñ ñn ti ñg Vi t ch t l ñg th p hoàn ch ñnh trong t ñg lai.

2. M c tiêu c a lu n v n

Lu n v n t p trung nghiên c u m t s ñ ph ñg pháp phân l p m u và trích ch ñ c tr ñg nh m l a ch ñ c m t ph ñg pháp thích h p cho vi c nh ñ ñg các nh ch cái và ch s ñ ti ñg Vi t ch t l ñg th p.

nh ch t l ñg th p ñây bao g m các nh ký t b l i do nhi u, do b t nét, b thi u ho c th a ra m t ph ñn nào ó do ñính vào ký t bên c ñnh, do các thành ph ñn c a ký t b ñính v ñ nhau ch ñg h n nh ph ñn m , ñ u ñính v ñ ph ñn ch ñ ñ v ñ ký t ti ñg Vi t.

3. B c c c a lu n v n

Các n ñ ñng trình bày trong lu n v n c ñ chia thành 3 ch ñg:

Ch ñg I: T ñg quan v bài toán nh ñ ñg ch Vi t.

Ch ñg này trình bày t ñg quan v các v ñ ñ liên quan ñ nh ñ ñg, các c tr ñg c a ch Vi t và ch Vi t ch t l ñg th p, nh ñg v n t ñn t ñi trong bài toán nh ñ ñg nh v n b n ch t l ñg th p, ñ ra mô hình chung c a h ñ th ñg nh ñ ñg, các h ñg tí p c ñnh ñ ñg, các y u t ñh h ñg ñh ñ th ñg nh ñ ñg.

Ch ñg II: M t s v n trong nh ñ ñg ký t ch t l ñg th p

Ch ñg này trình bày nh ñg khái ni m c b n v ñnh ký t ch t l ñg th p, m t s ñ h ñg tí p c ñn trong phân l p và trích ch ñ c tr ñg ký t và l a ch ñn m t ph ñg pháp nh ñ ñg nh ký t ch t l ñg th p.

Ch ñg III: Th c nghi m và ánh giá k t qu