ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

KHOA KHOA HỌC MÁY TÍNH

ĐỒ ÁN CUỐI KỲ MÔN MÁY HỌC

NHẬN DIỆN CHỮ CÁI VIẾT TAY TIẾNG VIỆT

Nguyễn Quốc Cường - 18520206

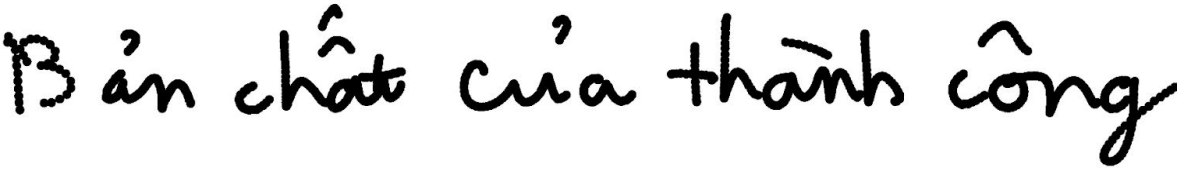
Ho Chi Minh City, 07/2020

I. GIỚI THIỆU BÀI TOÁN:

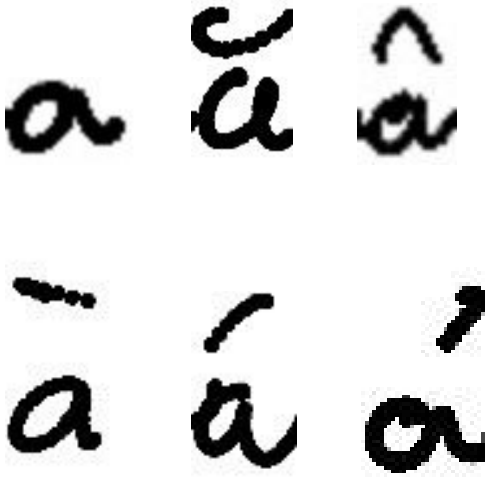
Bài toán nhận diện ký tự quang học(OCR) là một trong các bài toán Computer Vision có rất nhiều ứng dụng trong thực tế. Nhận diện ký tự scene text giúp trích xuất thông tin từ ảnh, nhận diện chữ viết tay giúp tự động hóa quá trình chuyển văn bản viết tay sang chữ in,... Trong đồ án này, sinh viên tìm hiểu về một bước trong bài toán con của OCR là nhận diện văn bản viết tay tiếng Việt. Có nhiều hướng tiếp cận cho bài toán này. Trong phạm vi đề tài, sinh viên tìm hiểu hướng tiếp cận theo cách chia nhỏ ảnh chứa văn bản thành các ảnh nhỏ chứa ký tự và nhận diện các ký tự này. Sinh viên tập trung tìm hiểu bước nhận diện ký tự trong quy trình trên.

II. MÔ TẢ DỮ LIỆU:

- Tập dữ liệu được xây dựng dựa trên tập dữ liệu của cuộc thi ICFHR 2018 Competition. Tập dữ liệu gốc có dạng ảnh chứa 1 dòng văn bản như sau:



- Tập dữ liệu của đồ án được xây dựng bằng cách cắt các vùng chứa ký tự trong ảnh gốc



- Dữ liệu có tổng cộng 3115 ảnh, được chia thành:
    + Tập train: 2180 ảnh
    + Tập test: 935 ảnh
- Đối với chữ viết tay có dấu, dấu thường sẽ nằm lệch nên khó có thể cắt được các chữ cái có dấu (cũng sẽ gây khó khăn đối với việc nhận diện từng ký tự và ghép lại). Có nhiều ký tự chứa dấu như trên chứa ít sample (khoảng 3-4 ảnh) nên dataset đang bị lệch(skew)

III. TIỀN XỬ LÝ VÀ RÚT TRÍCH ĐẶC TRƯNG:

- Đối với các ảnh quá nhỏ, cần chỉnh ảnh về kích cỡ đủ để các thuật toán rút trích đặc trưng ảnh hoạt động (width * height > 800)
- Sử dụng một số đặc trưng ảnh dựa trên hình dạng[2]: Aspect ratio, Rectangularity, Circularity, Equivalent diameter, Orientation, Minor,major axis, Contrast, Correlation, Entropy, Inverse difference moment

IV. MÔ HÌNH MACHINE LEARNING:

- Thử nghiệm 2 mô hình SVM(dùng poly kernel) và Decision tree (của thư viện Scikit-learn)

- Đối với SVM, thử nghiệm chỉnh hyper parameter degree
- Đối với Decision tree, thử nghiệm chỉnh hyper parameter depth

V. KẾT QUẢ:

- Model SVM(poly kernel) bị underfitting đối với dữ liệu và cách trích xuất đặc trưng như trên
- Model Decision Tree cho kết quả accuracy 60% trên tập test. Nhưng model vẫn đang overfit (accuracy trên tập train là 99%)

	precision	recall	f1-score	support
0	0.00	0.00	0.00	12
1	0.07	0.88	0.14	59
2	0.00	0.00	0.00	58
3	0.12	0.24	0.16	74
4	0.00	0.00	0.00	2
5	0.00	0.00	0.00	35
6	0.00	0.00	0.00	5
7	0.00	0.00	0.00	31
8	0.00	0.00	0.00	31
9	0.00	0.00	0.00	14
10	0.14	0.20	0.16	50
11	0.00	0.00	0.00	8
12	0.00	0.00	0.00	28
13	0.00	0.00	0.00	32
14	0.00	0.00	0.00	41
15	0.00	0.00	0.00	9
16	0.00	0.00	0.00	19
17	0.00	0.00	0.00	9
18	0.00	0.00	0.00	4
...				
accuracy			0.09	935
macro avg	0.01	0.02	0.01	935
weighted avg	0.02	0.09	0.03	935

SVC report

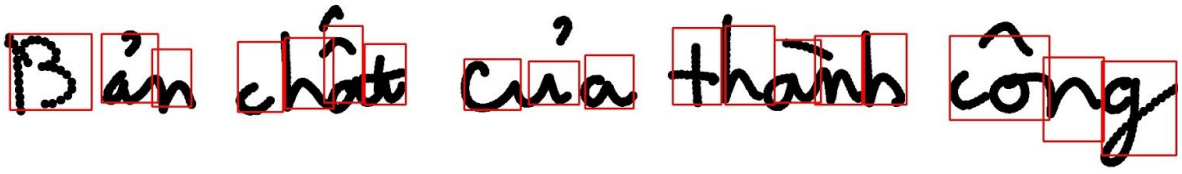
	precision	recall	f1-score	support
no0	0.36	0.67	0.47	12
no1	0.52	0.85	0.64	59
no2	0.55	0.60	0.57	58
no3	0.69	0.85	0.76	74
no4	0.00	0.00	0.00	2
no5	0.55	0.83	0.66	35
no6	0.33	0.20	0.25	5
no7	0.65	0.55	0.60	31
no8	0.54	0.61	0.58	31
no9	0.35	0.43	0.39	14
no10	0.59	0.80	0.68	50
no11	0.50	0.38	0.43	8
no12	0.69	0.71	0.70	28
no13	0.84	0.81	0.83	32
no14	0.70	0.90	0.79	41
no15	0.50	0.67	0.57	9
no16	0.83	0.26	0.40	19
no17	0.14	0.11	0.12	9
no18	0.50	0.50	0.50	4
...				
accuracy			0.60	935
macro avg	0.50	0.42	0.43	935
weighted avg	0.61	0.60	0.58	935

Decision tree report

- Đối với model SVM, có thể chọn các loại kernel khác hoặc tìm thêm feature
- Đối với model decision tree, có thể mở rộng tập dataset nhiều hơn để tránh overfit

VI. DEMO:

Sinh viên thử xây dựng ứng dụng demo áp dụng module nhận diện ký tự vừa train để nhận diện một đoạn văn bản. Ở bước tách riêng từng ký tự, sinh viên sử dụng model pretrained CRAFT[1].



Link video demo: link

VII. HƯỚNG CẢI TIẾN ĐỂ CÓ THỂ ỨNG DỤNG NHẬN DIỆN ĐOẠN VĂN:

- Mở rộng tập dataset cho bước nhận diện chữ cái
- Train lại model detection dành cho tiếng Việt thay cho model CRAFT ở trên
- Xây dựng module tiền xử lý tự động sửa vị trí của dấu

Tài liệu tham khảo:

[1] "Character Region Awareness for Text Detection"; Youngmin Baek, Bado Lee, Dongyoon Han, Sangdoo Yun, and Hwalsuk Lee* Clova AI Research, NAVER Corp; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019

[2] "Shape analysis and Measurement", Michael A. Wirth