A Dataset of Chinese Calligraphy Characters

Bai Bing, Hsu ChingWei

Department of Computer Science
Tokyo Institute of Technology

2022/02/03

Overview

1. Introduction

2. Dataset details

2.1 Data collection

2.2 Data annotation

2.3 Content overview

3. Experiments

3.1 Font classification

3.2 Character Classification

4. Conclusion

Introduction

- We made a dataset of **Chinese calligraphy characters**
  - So-called ”shufa”, are stylized artistic writings of Chinese characters
  - Have great variety of styles and appearances

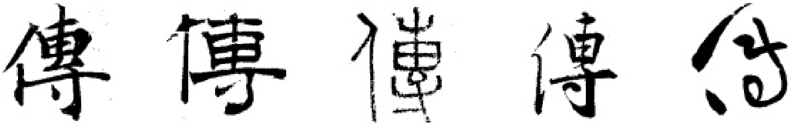


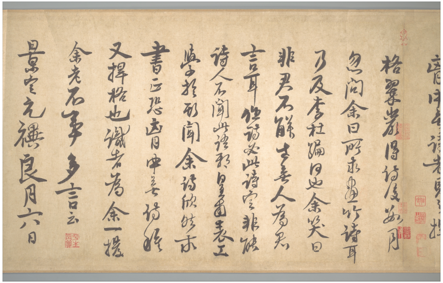
Figure: Different fonts of the same word ”chuan”

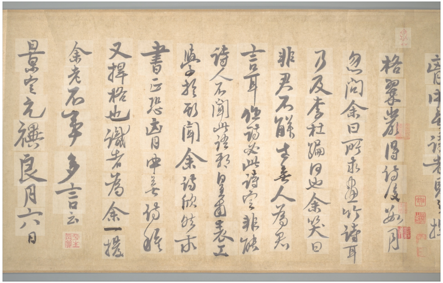
Related Work

- Handwritten Chinese character dataset
  - HCL2000 [1], SCUT-COUCH [2], CASIA [3]
  - Contain only modern handwrittings of Chinese characters
- Handwritten Chinese character dataset
  - Cursive Chinese Calligraphy Dataset [4]: only contains cursive characters
  - Cadal calligraphic database [5]: most similar to our work
- The lack of dataset make the challenging Chinese calligraphies recognition / classification tasks even more difficult.

Data Collection(1)

- Image source: The Metropolitan Museum of Art website (public domain)
- Applied a Chinese character detection network based on YoloV5 to crop into 4822 images with isolated characters

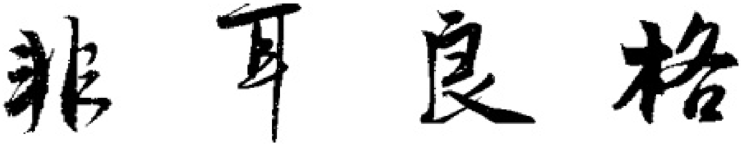




Data Collection(2)

Processed the cropped images to get neat backgrounds and centered characters

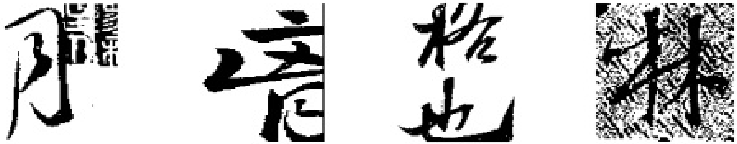
- Threshold the pixel values
- Add a 10% padding to all sides
- Resizing to 100×100pixels,



Data Collection(3)

Discard images that have either of the following problems:

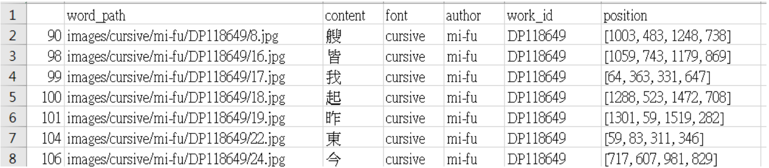
- The image overlaps with non-character such as stamps.
- The character is out of the image's border.
- The number of characters in the image is not one.
- The background is noisy due to inappropriate thresholding.



A total of 2896 images remaining in the dataset.

Data Annotation

- Type of font: regular / clerical / cursive / semi-cursive / seal
- Author
- Textual content
  - Using traditional Chinese characters
  - Encoded by UTF-8 BOM



1		word_path	content	font	author	work_id	position
2	90	images/cursive/mi-fu/DP118649/8.jpg	艘	cursive	mi-fu	DP118649	[1003, 483, 1248, 738]
3	98	images/cursive/mi-fu/DP118649/16.jpg	皆	cursive	mi-fu	DP118649	[1059, 743, 1179, 869]
4	99	images/cursive/mi-fu/DP118649/17.jpg	我	cursive	mi-fu	DP118649	[64, 363, 331, 647]
5	100	images/cursive/mi-fu/DP118649/18.jpg	起	cursive	mi-fu	DP118649	[1288, 523, 1472, 708]
6	101	images/cursive/mi-fu/DP118649/19.jpg	昨	cursive	mi-fu	DP118649	[1301, 59, 1519, 282]
7	104	images/cursive/mi-fu/DP118649/22.jpg	東	cursive	mi-fu	DP118649	[59, 83, 311, 346]
8	106	images/cursive/mi-fu/DP118649/24.jpg	今	cursive	mi-fu	DP118649	[717, 607, 981, 829]

Content overview(1)

Includes calligraphy works of different fonts from 6 famous calligraphers.

Font	Author	# of Data
Semi-cursive	Zhao MengJian	1375
Cursive	Huang TingJian	541
	Mi Fu	183
Regular	Zhong ShaoJing	653
Seal	Wu XiZai	52
	Yuan YuHe	47
Clerical	Yuan YuHe	45

Content overview(2)

Contains a broad variety of characters with 963 different textual contents.

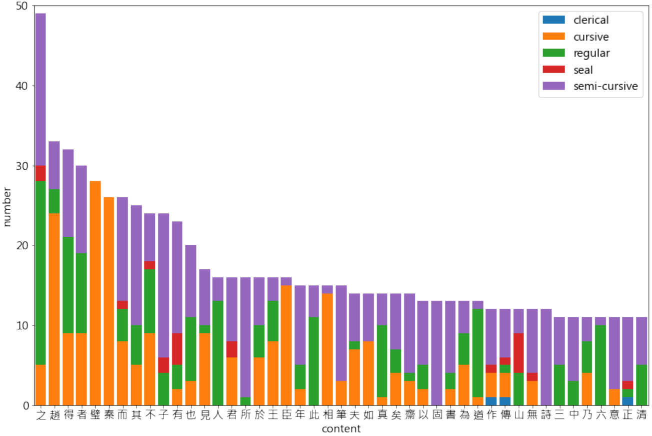


Figure: Distribution of the top frequency characters

Font Classification: Implementation details

- Dataset: split into train/valid/test set using a 80:10:10 ratio, batch size = 16
- Network Architecture: ResNet, SE-ResNet, Se-ResNeXt
- Optimizer: SGD, momentum 0.9, initial learning rate 0.1, weight decay 1e-4
- Criterion: Cross entropy loss
- Training time: 20 Epochs
- Machines: Google Colab, Tsubame

Font Classification: Results

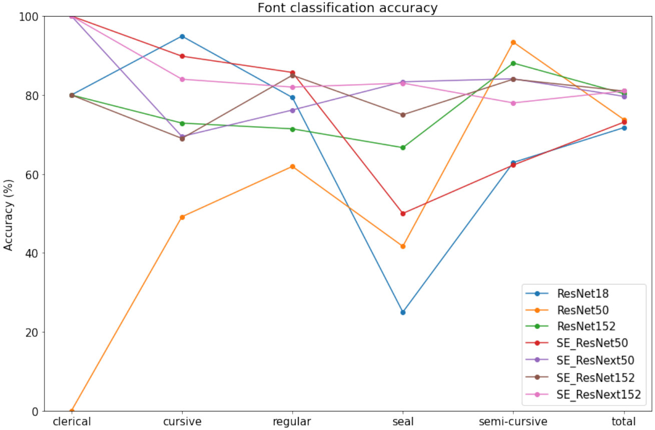


Figure: Font classification accuracy

Character Classification: Implementation details

- Dataset: Construct dataset from ”semi-cursive” images, batch size =10
- Network Architecture: , Siamese Network
- Optimizer: Adam, initial learning rate 1e-4
- Criterion: Contrastive Loss
- Training time: 50 Epochs
- Machines: Google Colab

Character Classification: Training Result

- train loss

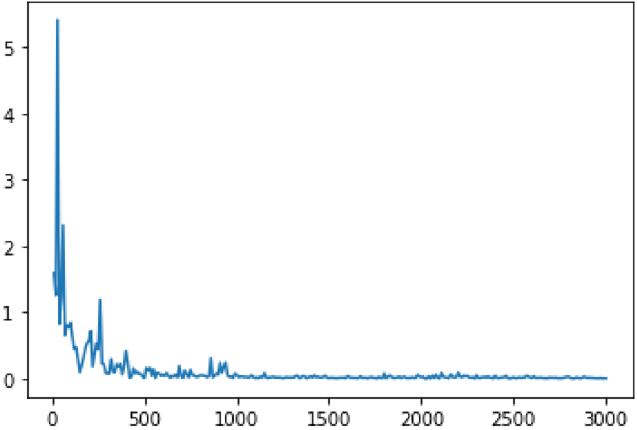
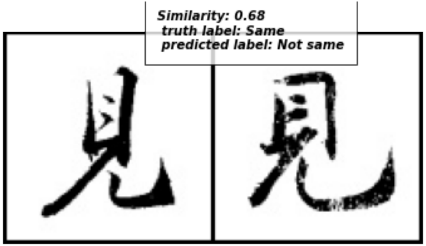
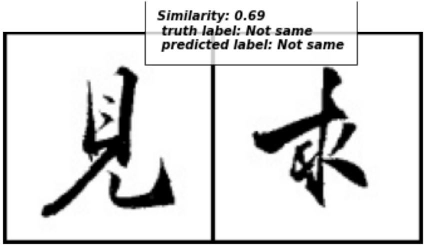


Figure: Training Loss

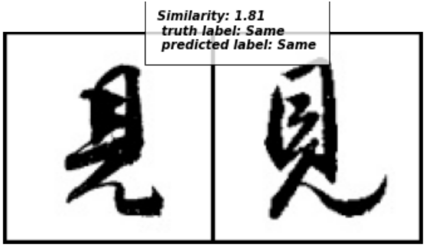
Character Classification: Similarity Comparison



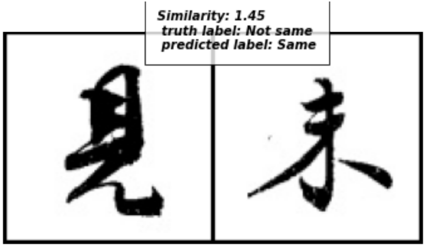
Similarity: 0.68
truth label: Same
predicted label: Not same



Similarity: 0.69
truth label: Not same
predicted label: Not same



Similarity: 1.81
truth label: Same
predicted label: Same



Similarity: 1.45
truth label: Not same
predicted label: Same

Conclusion

Construct Dataset

- Collect calligraphy works (Bai)
- Extract character images (Hsu)
- Annotation (Hsu, Bai)

Load Data and Train Neural Networks

- Create data loader (Hsu)
- Font classification on the ResNets (Hsu)
- Character recognition on a Siamese net (Bai)

References

Zhang, Honggang and Guo, Jun and Chen, Guang and Li, Chunguang (2000)
HCL2000-A large-scale handwritten Chinese character database for handwritten character recognition
International Conference on Document Analysis and Recognition 286–290.

Li, Yunyang and Jin, Lianwen and Zhu, Xinghua and Long, Teng (2008)
SCUT-COUCH2008: A comprehensive online unconstrained Chinese handwriting dataset
ICFHR 37 – 41.

Liu, Cheng-Lin and Yin, Fei and Wang, Da-Han and Wang, Qiu-Feng (2011)
CASIA online and offline Chinese handwriting databases
International Conference on Document Analysis and Recognition 37 – 41.

Liang, Jung and Liao, Wen-Hung and Wu, Yi-Chieh (2020)
Toward Automatic Recognition of Cursive Chinese Calligraphy: An Open Dataset For Cursive Chinese Calligraphy Text
International Conference on Ubiquitous Information Management and Communication (IMCOM) 1 – 5.

Zhang, Xiafen and Nagy, George (2011)
The CADAL calligraphic database
Proceedings of the 2011 Workshop on Historical Document Imaging and Processing 37 – 42.

The End