CS 229 Project Proposal

Chinese Ancient Calligraphy Character Reconstruction

Li Deng **Liyi Wang**

dengl11 liyiw

1. Motivation

Calligraphy is a piece of gem in Chinese splendid ancient culture, endowed with various font families and multiple factions within each font family in the evolution of thousand of years. However, most ancient calligraphy works are carved onto stones. After years of weathering, many of the characters are incomplete, which is a huge loss to the cultural heritage of the world. We both are Chinese calligraphy lovers since childhood, and feel responsible and passionate to restore and protect them. By using Machine Learning, we are hoping to teach the machine Chinese calligraphy, and to reconstruct those incomplete calligraphy characters of different font families, by learning from the style of other calligraphy characters of the same font family in existence.

Input Expected Output

2. **Expected Target**
    - **Prediction** of various font families: which font family?
    - **Reconstruction** of characters to complete character to specified font family

3. **Relevant Dataset**
    - ***Microsoft Word* Chinese character generation**
    - **Chinese calligraphy dictionary website:**

      (Ex: http://m.cidianwang.com/shufa/shu3304_ks.htm)
    - **Chinese calligraphy font library:**

      (Ex: http://www.diyiziti.com)
    - Chinese ancient Documents
    - Chinese Calligraphy copybooks

4. **Methodology**
    - **Character Recognition:**
        i. Image recognition with supervised learning from database of chinese character
        ii. Inference from context of the documents
    - **Font Reconstruction:**

i. Font recognition with supervised learning from database of various font families (Naive Bayes? Support Vector Machine?)

ii. Try perceptron algorithm and KNN algorithm and compare their results

5. Milestones

1. **Data Collection**: 3rd Week (Done) Image segment by python

2. **Image Parsing and Feature Extraction**: 4-5th Week
3. **Character Recognition**: 6-7th week
4. **Font Reconstruction**: 8-10th week
5. **Documentation** 11th week

6. Detailed Tasks

- **Image Preprocessing:** reduce noise in image(background color, grid line, etc.), decompose characters into strokes.
- **Image Conversion:** image to matrix
- **Image Generation**: generate matrix or any other mathematical information to images
- **Feature Extraction:**
  - **i.** Calligraphy Theory Reading
  - **ii.** Define a set of features corresponding to different font styles, and convert matrix to feature vectors
  - **iii.** Find width-height-ratio according to characters in a whole, find incline angle of a stroke(horizontal, vertical, press down, dot, hook, turning) and the variance of thickness in it.
- **Machine Learning Model Choice**: choose appropriate learning algorithm (supervised learning) to build feature-to-output model

7. Learning Resources

- Neural Networks for Machine Learning: https://www.youtube.com/playlist?list=PLoRl3Ht4JOcdU872GhiYWf6jwrk_SNhz9

8. Recent Schedule

Week	Li	Liyi
3-4	Feature Definition	Image conversion
	Possible Model study	
4-5	Character Decomposition	Overall Features Extraction

9. Literature Review

Relevant Previous Year's CS229 Projects

- **Automatic Image Colorization:**
    - Support Vector Regression and Markov Random Fields: train image features
    - Scikit-image for image segment
- **Object recognition and image reconstruction**
    - Extract HOG features and linear SVM classifier on a set of features
    - Pairwise Dictionary
    - Convolutional Neural Network
- **Human Relation Match**
    - Decision Tree and a Naive Bayes

10. Github

Our project is hosted at Github:

https://github.com/dengl11/Chinese-Calligraphy-Character-Reconstruction