Identifying Chinese Calligraphers from their Characters Using Diffusion Maps

Evan Gerritz, Yale College '24
Advisor: Professor Steven Zucker

Overview

We will attempt to learn the natural coordinates of the manifold of Chinese calligraphic styles by applying the nonlinear dimensionality reduction algorithm *diffusion maps* to various writing samples of famous calligraphers. Through this we will write a final report discussing the key differentiators of calligraphic style and the feasibility of computer and human differentiation of calligraphers based on style. We will also apply various other machine-learning techniques to compare their efficacy for this task.

Background and Motivation

Chinese calligraphy (in Chinese, 书法—literally 'rules/methods of writing') is one of the oldest and deepest artforms in the world. While the original paper on which the most famous calligraphers wrote has long since disintegrated, their works have been preserved in the form of steles—large stones in which master carvers painstakingly transferred the ink characters on paper to engravings in a rock, a significantly more durable material.

To become a proficient calligrapher, one begins by reproducing as exactly as possible the characters written by famous calligraphers (a process known as 临帖) before attempting to create anything new themselves. Many of the most famous Chinese calligraphers have highly distinctive styles easily recognized by those familiar with the millennia-old artform. A comparison of the "four great masters of regular script" (楷书四大家), who each have calligraphic styles named after them, can be seen in Figure 1.

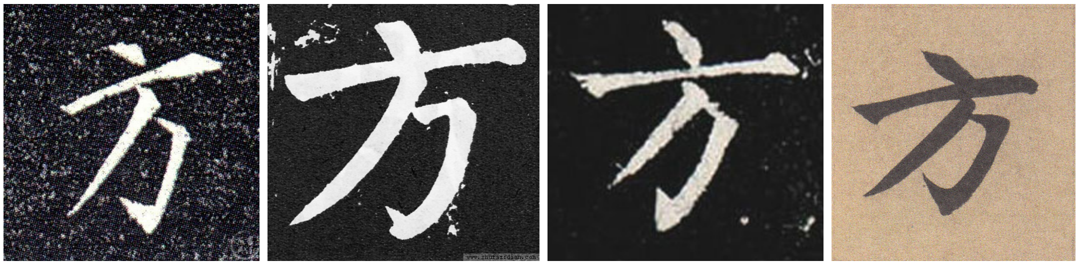

Figure 1: Four instances of the character 方 (place) written by (from left to right): Ouyang Xun (欧阳询, 557-641 CE) Yan Zhenqing (颜真卿, 709-785 CE), Liu Gongquan (柳公权, 778-865 CE), Zhao Mengfu (赵孟□, 1254-1322 CE) (Gao, 2007).

Our question is as follows: can a computer be trained to recognize a calligrapher from an image of a single character? To do this, we will use a nonlinear dimensionality reduction technique called diffusion maps. Dimensionality reduction techniques are useful for finding patterns in data that are high dimensional (such as a 128x128, 8-bit image). Linear algorithms, such as Principal Component Analysis, are useful as they find the “best” lines onto which to project the data, according to some criterion (maximum variance while maintaining orthogonality, in the case of PCA); however, the linearity restriction will fail to capture a potentially rich manifold on which the data truly lie. Figure 2 demonstrates the necessity for nonlinear techniques.

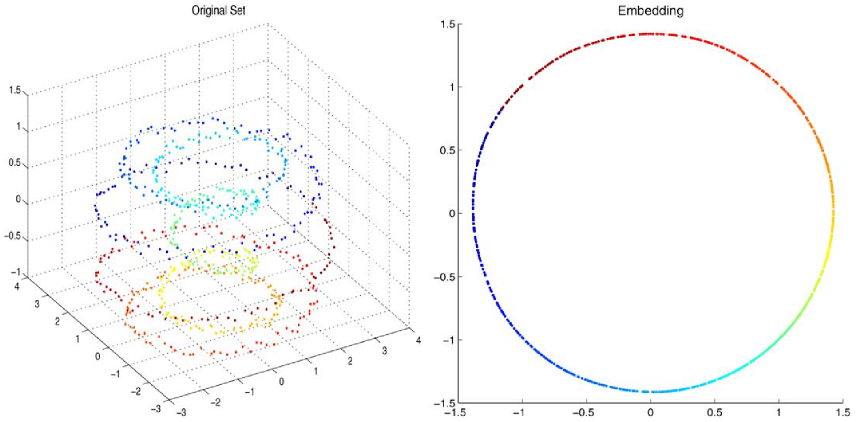

Figure 2 (from Coifman et. al., 2006): notice how the colors on the helix change according to one’s location on the helix and not in Euclidean space itself. Only a nonlinear dimensionality reduction can express this relationship; the result of a diffusion map embedding is on the right.

The diffusion maps algorithm solves this linearity constraint by considering each datapoint as a node in a graph, with connections to each other node weighted according to some similarity measure. If one now considers taking random walks on this graph using the similarity measure between two nodes as the transition probability (after rescaling such that the values are between 0 and 1), one can create a notion of how “close” two nodes are—not in the Euclidean sense, but instead how close they are on the manifold on which we believe the data live. To actually compute these manifold embeddings, one computes the singular value decomposition of the similarity matrix, after rescaling by the row sums, to get the coordinate functions of the learned manifolds.

Using these coordinate functions, we can compute the coordinates on which each artist lives on a calligraphic style manifold. This will provide us with new insights into the latent, intrinsic structure of calligraphic style.

Goals and Methods

To begin, we will need to find or create a dataset of several examples of various characters drawn by famous calligraphers. One useful resource for this purpose is https://www.shufazidian.com/, which is a huge database of characters drawn by famous calligraphers searchable by character. Some other preprocessing and data cleaning will likely also need to be performed in order to improve the results. Examples of such preprocessing include normalizing contrast or ink/paper colors, centering/cropping images, and choosing between shallow bit-depth and high-resolution (e.g., a 256x256 binary image) and high bit-depth, low-resolution (e.g., an 8x8 16-bit image).

We will then try to find the best similarity kernel for producing the similarity matrix, as this choice is particularly critical in the case of diffusion maps; one option is cosine similarity, and another is Earthmover's distance. To do this, we will examine the resulting embeddings using various similarity kernels.

Once we have a suitable dataset and similarity matrix, we need to decide which eigenfunctions on the manifold are best as a calligrapher coordinate-system. This can be done by plotting the projections of data points onto 2-3 eigenfunctions and color-coding the points by calligrapher. The ideal manifold will have these points placed such that all points of one color are adjacent. The final report will include an analysis of what structure our diffusion maps embedding reveals by looking at actual examples of characters clustered together.

To get a sense of the success of the diffusion maps approach, we will then compare the embeddings to results using, e.g., PCA (a linear algorithm), t-SNE (another popular nonlinear algorithm), and a deep neural network trained purely for classification.

The final report and poster will include the results of the diffusion maps algorithm, a discussion of the latent structure revealed by its embedding, and a comparison to results from other algorithms.

Timeline[1]

- September:
  - Preliminary research on diffusion maps (2 weeks)
  - Creation and preparation of dataset (2-3 weeks)
- October:
  - Main experiments performed using diffusion maps and other techniques (4 weeks)
- November:
  - Additional experiments (1-2 weeks)
  - Possibly some dataset improvements (1-2 weeks)
  - Preparation of final deliverables (2 weeks)

[1] N.B.: the number-of-week time estimates are not disjoint.

Deliverables

We will create a final report and a poster for this project, each to be submitted before the given deadlines. Additionally, all code used for this project will be made available.

References

Coifman, R. R., Lafon, S., Lee, A. B., Maggioni, M., Nadler, B., Warner, F., and Zucker, S. W. (2005). Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps. Proceedings of the National Academy of Sciences. 102, 7426–7431. https://doi.org/10.1073/pnas.0500334102.

Coifman, R. R., & Lafon, S. (2006). Diffusion maps. *Applied and Computational Harmonic Analysis*, *21*(1), 5–30. https://doi.org/10.1016/j.acha.2006.04.006.

Gao, C. (2007). *China's Calligraphy Art Through the Ages*. China Intercontinental Press (五洲传播出版社).