北京大学

A Method of Density Analysis for Chinese Characters

Jingwei Qu, Xiaoqing Lu, Lu Liu, Zhi Tang, and Yongtao Wang

Institute of Computer Science & Technology of Peking University, Beijing, China

2014-12-09

Outline

- Introduction
- Related Work
- Density Analysis
  - Center-to-Center Distance of Connected Components
  - Gap between Connected Components
  - Ratio of Perimeter and Area
  - Connected Components Area Ratio
  - Area Ratio of Holes
  - Overall Density Metric
- Experiments
- Conclusion & Future Work

Outline

- Introduction
- Related Work
- Density Analysis
  - Center-to-Center Distance of Connected Components
  - Gap between Connected Components
  - Ratio of Perimeter and Area
  - Connected Components Area Ratio
  - Area Ratio of Holes
  - Overall Density Metric
- Experiments
- Conclusion & Future Work

Introduction

- **Density is a significant factor in the design, recognition, and other applications of Chinese characters as fonts**
  - **The evaluation of font beauty**
  - **The aesthetic quality of a page**
  - **The consistent overall layout effect of different languages and different characters**

Introduction

- **The challenges density analysis face**

  ◆Density degree is basically evaluated by human visual perception, which lacks reasonable visual models

  ◆Various factors influence shape density, thus requiring deep-seated shape analysis

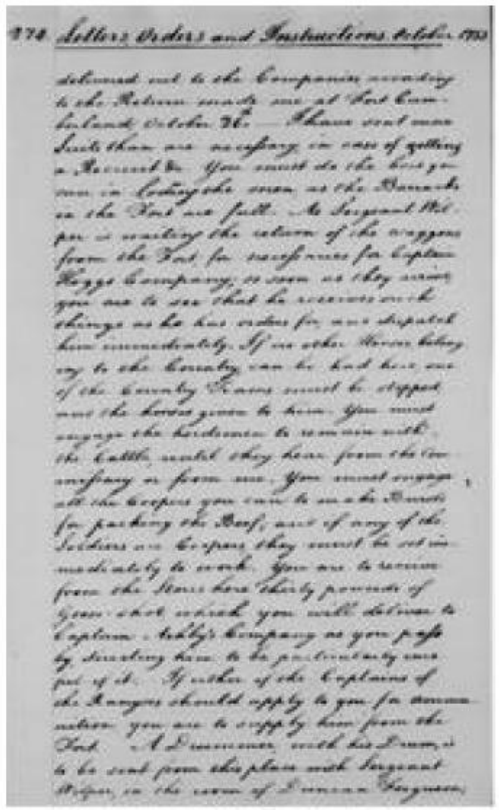
  ◆No evaluation method or common dataset is authorized to judge the validity of the density metric

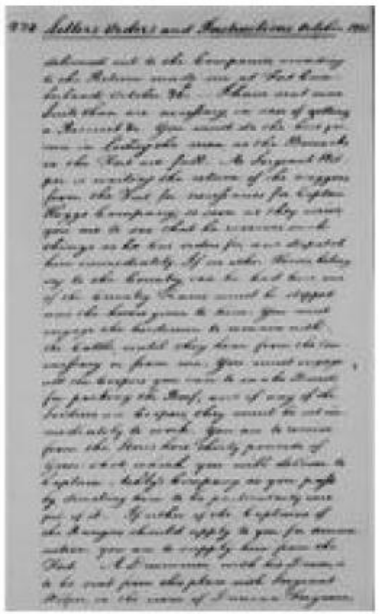
Outline

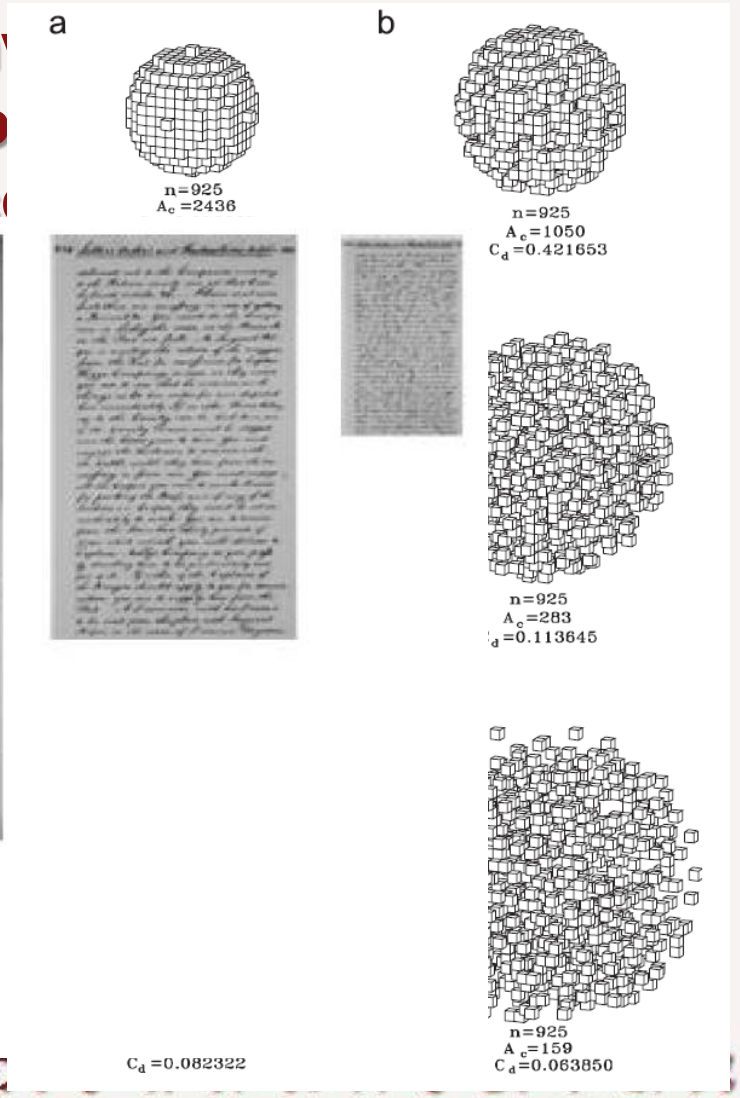
- Introduction
- Related Work
- Density Analysis
  - Center-to-Center Distance of Connected Components
  - Gap between Connected Components
  - Ratio of Perimeter and Area
  - Connected Components Area Ratio
  - Area Ratio of Holes
  - Overall Density Metric
- Experiments
- Conclusion & Future Work

Related Work

- Density analysis has been in
  studies on pattern recognitio
  retrieval, which use similar c







Outline

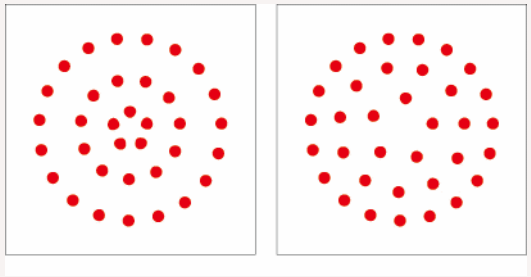
- Introduction
- Related Work
- Density Analysis
  - Center-to-Center Distance of Connected Components
  - Gap between Connected Components
  - Ratio of Perimeter and Area
  - Connected Components Area Ratio
  - Area Ratio of Holes
  - Overall Density Metric
- Experiments
- Conclusion & Future Work

Density Analysis

- Density is one of the important properties for shape analysis
- A character → A shape composed of multiple connected components. Five dominant metrics for a more in-depth analysis and effective description of density:
  - Center-to-center distance of connected components
  - Gap between connected components
  - Ratio of perimeter and area
  - Connected components area ratio
  - Area ratio of holes
- Combining five features → Overall Density Metric

CCDCC

- **CCDCC**
  - **Center-to-Center Distance of Connected Components**
  - **Definition: The distances between geometric centers of any two connected components**
  - **Description: The layout feature of all connected components in a Chinese character**



Same character with tight ZhongGong (left) and loose ZhongGong (right)

CCDCC

- **Formula:**

$$d_{ij} = \sqrt{(\bar{x}_i - \bar{x}_j)^2 + (y_i - \bar{y}_j)^2}$$

$$D = \{d_{12}, d_{13}, \cdots, d_{ij}, \cdots, d_{n(n-1)}\}, |D| = C(n,2)$$

- **WCCDCC:**

$$d_{avg} = \frac{\sum d_{ij} - d_{max} - d_{min}}{|D| - 2}$$

$$WCCDCC = \frac{d_{max} * w_{min} + d_{min} * w_{max} + d_{avg} * w_{avg}}{L_{diagonal}}$$

GCC

- GCC
  - Gap between Connected Components
  - Definition: The gaps between any two connected components
  - Description: The layout feature of all connected components in a Chinese character

An example of Gap

GCC

- **Formula:**

$$gap_{ij} = 2 * min\{s | Compn(A_{ij} \cdot B_s) = 1\}$$

$$GAP_{SET} = \{gap_{12}, gap_{13}, \cdots, gap_{ij}, \cdots, gap_{n(n-1)}\}, |G| = C(n,2)$$

- **An average value of gaps：**

$$gap_{avg} = \frac{\sum gap_{ij} - gap_{most} * n_{most} - gap_{max} * n_{max} - gap_{min} * n_{min}}{|G| - n_{most} - n_{max} - n_{min}}$$

RPA

- RPA
  - Ratio of Perimeter and Area
  - Definition: The ratio (perimeter^2) divided by area
  - Description: The influence of the appearance of the outline on character density
- Formula:

$$C = \frac{n - P/4}{n - \sqrt{n}}$$

RPA

- RPA can be applied not only to Chinese characters with multiple connected components, but also to those with single connected component
- RPA reveals the distribution of all pixels of a character and partially describes its density
- A higher pixel distribution results in a more concentrated image

(a) (b)

An example of the RPA: (a) RPA = 0.9922 (b) RPA = 0.9514

CCAR

- **CCAR**
  - **Connected Components Area Ratio**
  - **Definition: The ratio of the sum area of all connected components in a Chinese character with respect to the area of the region enclosed by the entire convex hull**
  - **Description: The influence of the scale of connected components on character density**
- **Formula:**

$$CCAR = \frac{s_1 + s_2 + \cdots + s_i + \cdots + s_n}{S_{con}}$$

CCAR

- Both characters with multiple connected components and characters with single connected components
- Revealing the fullness by which the strokes of a Chinese character cover the entire region.
- The higher the CCAR of a Chinese character, the denser is the image being perceived by human visual perception

(a) (b)

An example of CCAR: (a) CCAR = 0.2979 (b) CCAR = 0.4122

ARH

- ARH
  - Area Ratio of Holes, ARH
  - Definition: The ratio of the sum area of all holes in a Chinese character with respect to the area of the character whose holes have been filled
  - Description: The influence of the structure of strokes on character density
- Formula:

$$ARH = \frac{S_{hole}}{S_{com}}$$

ARH

- We consider the scale of the holes
- ARH reveals the relative size of the holes in a Chinese character
- When ARH is high, the hole is large with respect to the character, and the character will seem loose and expanded to human vision
- Enhancement of CCAR descriptive power

(a)

(b)

An example of CCAR and ARH: (a) CCAR = 0.4171, ARH = 0.2195
(b)CCAR = 0.4138, ARH = 0.0915

Overall Density Metric

- **With the above five density features, a density descriptor can be directly obtained by the five-dimension vector:**

$$V_{density} = (WCCDCC, GAP, RPA, CCAR, ARH)$$

Overall Density Metric

- The five factors have varying importance
- CCAR and RPA are adopted to calculate weights to enhance the gap value

$$GAP = gap_{most} * w_{most} + gap_{max} * w_{max} + gap_{min} * w_{min} + gap_{avg} * w_{avg}$$

(a) (b)

An example of GAP : (a) GAP = 22 (b) GAP = 64

Overall Density Metric

- **The overall density metric:**

$$GAP_{norm} = \frac{GAP}{L_{diagonal}}$$

$$Density = (WCCDCC + GAP_{norm} + ARH) \,/\, 3$$

Outline

- Introduction
- Related Work
- Density Analysis
  - Center-to-Center Distance of Connected Components
  - Gap between Connected Components
  - Ratio of Perimeter and Area
  - Connected Components Area Ratio
  - Area Ratio of Holes
  - Overall Density Metric
- Experiments
- Conclusion & Future Work

Experiments

- **Datasets**
  - 4 typefaces: Song, Fangsong, Boldface, and Regular Script
  - Each: 6715 128*128

Experiments

- **Comparsion R**
  - **We calculate t** **sity of each Chinese chara** **nd compare them**

Table 1. The samples of comparison on the five features of four different typefaces.

	Song	Fangsong	Boldface	Regular Script
(a)				
WCCDCC	0	0	0	0
GAP	2.5	2.5	2.5	2.5
RPA	0.8943	0.9306	0.9766	0.9619
CCAR	0.5399	0.9159	0.9690	0.8449
ARH	0	0	0	0
Density	0.0072	0.0079	0.0075	0.0073
(b)				
WCCDCC	0.3154	0.2938	0.3329	0.3150
GAP	8	8	8	12
RPA	0.9333	0.9111	0.9632	0.9353
CCAR	0.2874	0.2459	0.4413	0.3062
ARH	0.2595	0.2207	0.1381	0.1305
Density	0.2077	0.1883	0.1734	0.1734
(c)				
WCCDCC	0.3443	0.3017	0.3457	0.3295
GAP	57.7750	21.7273	59.1364	77.8000
RPA	0.9125	0.8940	0.9469	0.9238
CCAR	0.2845	0.2605	0.3913	0.3035
ARH	0	0	0	0
Density	0.2312	0.1456	0.2375	0.2751
(d)				
WCCDCC	0.2765	0.2792	0.2871	0.2895
GAP	23.7000	21.7000	26.4000	19.8000
RPA	0.9456	0.9352	0.9701	0.9564
CCAR	0.2888	0.2704	0.4118	0.3695
ARH	0.5340	0.5122	0.4197	0.4067
Density	0.3185	0.3113	0.2907	0.2764
(e)				
WCCDCC	0.3290	0.3129	0.3347	0.3096
GAP	49.0909	17.2364	55.2000	47
RPA	0.9211	0.8932	0.9491	0.9252
CCAR	0.3349	0.2620	0.4601	0.3263
ARH	0	0	0	0
Density	0.2715	0.1438	0.2346	0.143

Experiments

- **Clustering Results**
  - **We use the K-means algorithm to cluster Chinese characters with boldface**
    - **Feature Vector——WCCDCC, GAP_{norm}, RPA, CCAR and ARH**

Table 2. The clustering results based on the feature vectors.

	WCCDCC	GAP_{norm}	RPA	CCAR	ARH
(a)	0.3383	0.4808	0.9519	0.4136	0.0732
(b)	0.3161	0.2580	0.9543	0.4160	0.0776
(c)	0.2945	0.0725	0.9606	0.4276	0.1314

Table 3. Some examples in each cluster.

(a)	二	三	为	心
	六	办	示	兰
(b)	业	代	泯	粉
	参	悴	臂	战
(c)	一	上	片	腱
	髑	麈	幢	篝

Outline

- Introduction
- Related Work
- Density Analysis
  - Center-to-Center Distance of Connected Components
  - Gap between Connected Components
  - Ratio of Perimeter and Area
  - Connected Components Area Ratio
  - Area Ratio of Holes
  - Overall Density Metric
- Experiments
- Conclusion & Future Work

Conclusion&Future Work

- Conclusion
  - We propose five density metrics from the pixel, outline, and component levels
    - CDCC、GAP、RPA、CCAR和ARH
  - Both local and global information of the connected components are considered
  - Our method can not only discriminate density differences between different typefaces for the same Chinese character, but also depict density differences between characters with the same typeface
- Future Work
  - Exploring more density factors to describe density
  - Incorporating related knowledge such as psychology
  - Applying the density analysis to benefit more research fields

北京大學

Thanks!
Q&A

qujingwei@pku.edu.cn