

### 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





- 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
- 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



PostScript 字形描述技术是用美国 Adobe 公司的 PostScript 页面描 述语言来描述字形的一种技术。 CID 字库是美国 Adobe 公司发表 的最新字库格式,所有字形描述都 采用 PostScript Type 1 格式,它 具有易扩充、速度快、兼容性好、 简便、灵活等特点,这种标准格式 保证了跨平台的高质量输出。

技术研究所



### 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





- 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

#### 974 Setters Orders and Instructions totale mo

determent wat to the Companies according

to che Robert mate me at that lan - local man Suits than are necessary in case of getting a Recount to you must do the bourge in loting the men as the Damash an the Tint are full . As Sugart the from the Fat for surprises for lapter Rogge Company; 11 voor as they aring no are to see that he receives much things as he has orders for and depated here comments atily. If we other Marine heling my to the bounty can be had beer our of the boundry Trains sured to stopped. and the hordes given to here. you must upage the herderness to remain with the Callle would shay hear from the low infrary as from sure . you count on your , all the beopers you can to make Burch In packing the Berf; and if any of the loidiers are beepens they event be set in mediately to work. you are to receive from the Sense hore thereby pounds of Geon abort which you will delive to laptom - tally's bimpany as you pop per if it. If wither of eds Coplains of che Annyous should apply to you for some nation you are to supply him from the and A Dummenter with his Dum, is to be sent from this place with Sergeant Steper, in the cover of Suman Geogueses,

#### 974 Letters inders and Inchastions able me

abound int to the bourpanne - - - a de Retern marte un at het lan its than are accepting in can the court to gove much do che use faith to began ing the estare of the way she have for northands for the Kogge hamping a care as the ups as he time orders for and departed deately of an actor the to cattle matil they down from the participate the Stort, and if any of the Hours have thinky proved of telly brokeny as you page Gruther y de bapt. a to must prove also place with how

北泉



C<sub>d</sub> = 0.082322

 $A_{c} = 1050$  $C_{d} = 0.421653$ 

b



A\_=283 = 0.113645



6





- 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
  - 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)





• 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
  - 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





#### • Formula:

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

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

#### • 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

- Ratio of Perimeter and Area
- Definition: The ratio (perimeter<sup>2</sup>) 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 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
  - 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:

$$CCAR = \frac{s_1 + s_2 + \dots + s_i + \dots + s_n}{s_{con}}$$





- 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







#### • 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:

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





- 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)** 

北京大学计算机科学技术研究所

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 fivedimension 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

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



### **Overall Density Metric**

#### • The overall density metric:

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

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





- 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





#### • Datasets

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











### Experiments

Boldface

Regular Script

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

Song

Comparison R

We calculate **Chinese chara** 

(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	sity of each
Density	0.0072	0.0079	0.0075	0.0073	ony of each
(b)	保	保	保	保	nd compare them
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)	游	游	游	游	
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)	茴	茴		茴	_
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)	阀	阀	阀	阀	
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	<b>耳 %1 科 子 仅 不 研 咒 所</b>

#### sity of each nd compare them





- Clustering Results
  - We use the K-means algorithm to cluster Chinese characters with boldface
    - ◆ Feature Vector—WCCDCC, GAP<sub>norm</sub>, RPA, CCAR and



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

	WCCDCC	GAP <sub>norm</sub>	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.



技术研究所





- 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



27

# **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
  - 北京大学计算机科学技术研究所



# Thanks! Q&A

qujingwei@pku.edu.cn