Implementation and Evaluation of Document Retrieval for the PC Notes Taker (PCNT) Handwriting Device

Nasir Mahmood

Otto - von - Guericke University, Magdeburg

November 1, 2007

Outline

- Introduction
- Method
- Testing and Evaluation
- Results and Discussions
- Summary

Handwriting

- **Handwriting** is used for
  - literary writing
  - correspondence
  - advertisement
  - ...
- its **electronic articulation** are
  - typewriter
  - computer
- hasn't lost **importance** due to claims of
  1. authenticity
  2. (inter-)mediality
  3. coporeality

Digital Handwriting

- Digital representation of the information of a user 's handwriting
- A way to convert written words from the ink on paper to digital format
- **Acquisition Approaches**
  1. Offline handwriting acuquisition
  2. Online handwriting acquisition

Offline Handwriting Acquisition

- Visual representation of a text
- No dynamic information
- A scanned image of handwriting is digitally read in
- **Applications**
  - Optical Character Recognition (OCR)
  - Intelligent Character Recognition (ICR)
- **Disadvantages**
  - noise from scanning the text
  - lines or patterns
  - extra marks due to dust or scratches

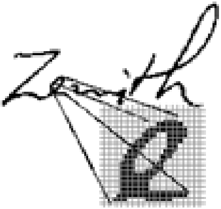


Image by visoinobjects

Online Handwriting Acquisition

- The way a text is written is important
- **Digital ink signal:**
  - a sequence of 2D points over time
  - information of strokes & trajectories
- **Devices** to capture digital ink
  1. digital pen on a patterned paper
  2. paper-based capture device
  3. pen-sensitive surface (touch screen)
- **Advantages**
  - No optical noise, easier data processing
  - Broad range of applications

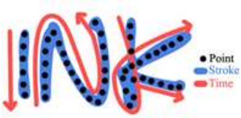


Image by visoinobjects

Document Retrieval

Given a set D of documents and a query word q, find a list $D^{'}$ of documents where q occurs at least once

$$D = \{d_1, d_2, d_3, ..., d_n\}$$
$$D^{'} = \{d_1^{'}, d_2^{'}, d_3^{'}, ..., d_n^{'}\}$$

query q and documents D are handwritten

Document Retrieval Methods

1. Textual recognition
   - most intuitive, simple string search on textual features
   - fails in most searches, not suitable for hand-drawn images
2. Direct handwriting matching
   - works for all kinds of scripts and images
   - our method lives under this catergory

Aim of the Work

- **Part A:**
  1. Introduction of sub-features used for document retrieval
  2. Comparision of existing and newly introduced features
- **Part B:**
  1. Implementation of document retrieval system for PC Notes Taker (PCNT) device
  2. Evaluation of our method against PCNT and its comparsion with those already tested

Document Retrieval Algorithms

Ideal Retrieval Algorithm

1. low complexity
2. works with simple features
3. faster and accurate

Existing Approaches

- Image feature indexing (Srihari et al.)
- Pattern recognition & maching learning (Schomaker et al.)
- Graph matching approach (Fonseca et al.)
- String matching (Sun et al.)

String Algorithms - Approximate String Search

- How close two strings (query & its instance in document) are.
- **Edit distance**, most common similarity measure
- **Approximate String Search - Local Alignment**
  - fuzzy search of short string (q) within a longer one (d)
  - a matrix D of dimension $(m+1)x(n+1)$
  - m and n are length of q and d
  - for a match $D(m,j) < \tau$, τ is a threshold

$$D(i,j) = \left\{ \begin{array}{ll} 0 & if \;\; i = 0, \\ D(i-1,0)+1 & if \;\; i > 0 \; and \; j = 0, \\ min \left\{ \begin{array}{l} D(i,j-1)+1 \\ D(i-1,j)+1 \\ D(i-1,j-1)+\delta(i,j) \end{array} \right\} & else, \end{array} \right\}$$

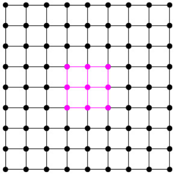
$$\delta(i,j) = \left\{ \begin{array}{ll} 0 & if \; q[i] = d[j], \\ 1 & else, \end{array} \right\}$$

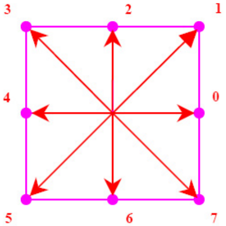
Handwriting Features

- Feature data represent ink traces of a writing process
- Approximate string search works with string features
- String features: x_t, y_t position of pen tip over time t
- **Freeman grid codes** - string features
  1. discretization of data with a grid
  2. assignment of codes to discretized data
- **Square** Freeman grid codes
- **Triangular** Freeman grid codes

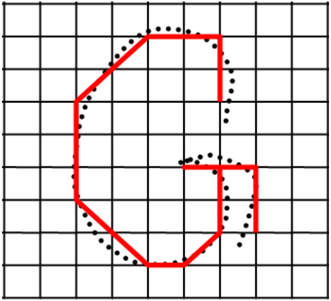
Square Freeman Grid Codes

- Two sets of evenly spaced lines perpendicular to each other - **square grid**
- Handwritten input is superimposed on the grid
- **Sample points** x_t, y_t are mapped to **nodes**
- Each sample point gets one of **8 symbols**
- **9th symbol** to encode a **gap**
- Encoded ink shape is sequence of symbols rather than sample points x_t, y_t





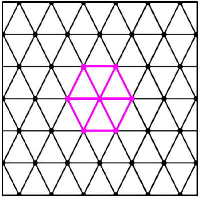
Square Freeman Grid Codes

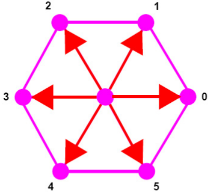


2224455666770122440066

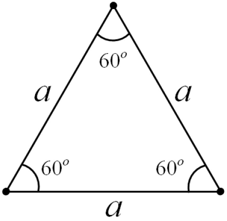
Triangular Freeman Grid Codes

- Three sets of lines at 60-degree angle to each other - **triangular grid**
- Quantization and assignment of codes for feature extraction is similar to square grid type except
  1. **Triangular grid** to sumperimpose handwriting onto
  2. **Six node symbols** to map sample points

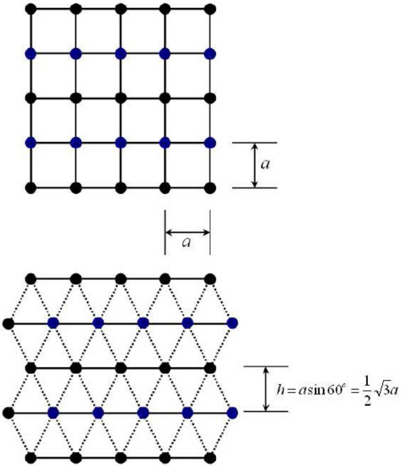




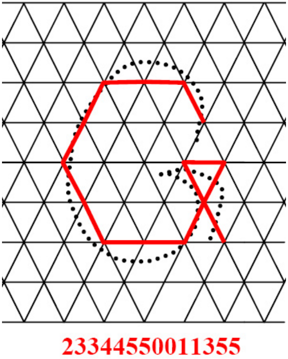
Triangular Freeman Grid Codes



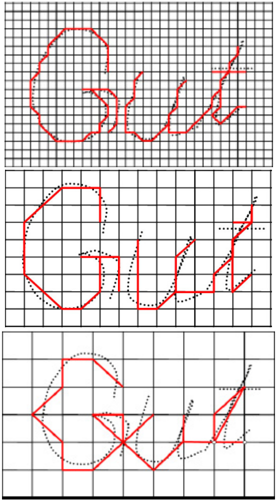
$$h = a\ sin60^o = \frac{1}{2}\sqrt{3}\ a$$

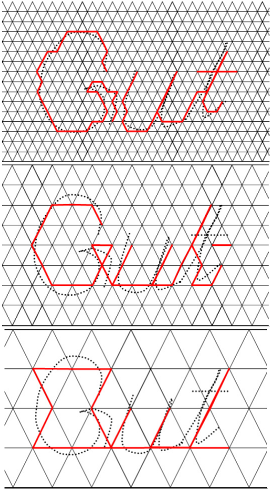


Triangular Freeman Grid Codes



Freeman Codes



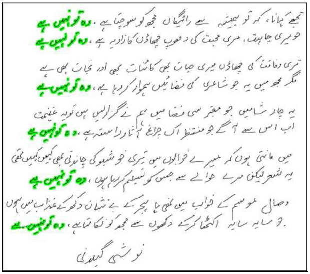


Pegasus PC Notes Taker Device (PCNT)

- **PCNT** captures handwriting online
- Its package comes with
  1. a cordless electronic pen
  2. a detachable base with USB cable
- For applications, its SDK is available to
  1. to capture data from device
  2. to process it accordingly
- Coverage area: A4 size paper
- Resolution: 1200 DPI

Data Collection

- No suitable testset database available
- Built our own **database**
  - in **English** and **Urdu** scripts
  - documents written with PCNT
  - documents read in with SDK
- **Database**
  - 80 documents by 8 persons
  - 5 documents per person in each script
  - documents contents - repetitive words/phrases
  - 29 queries manually selected & tagged
  - 804 true matches selected & tagged



تجھے بتانا، کہ تو ہمیشہ سے رائیگاں مجھ کو سوچتا ہے، وہ تو نہیں ہے
جو میری چاہت، مری محبت کی دھوپ چھاؤں کا زاویہ ہے، وہ تو نہیں ہے

تری رفاقت کی چھاؤں میری حیات بھی کائنات بھی اور نجات بھی ہے
مگر مجھ میں یہ جو شاعری کی فضائیں ہموار کر رہا ہے، وہ تو نہیں ہے

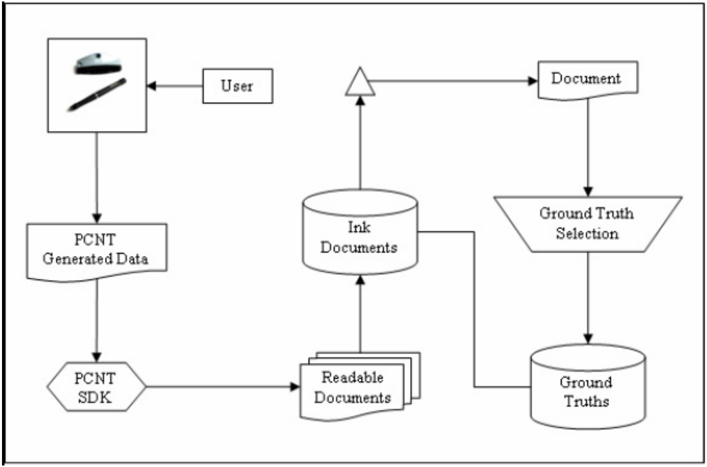
یہ چار شامیں جو معتبر سی فضا میں ہم نے گزار لیں ہیں تو یہ غنیمت
اب اس سے آگے جو منتظر اک چراغ آثار راستہ ہے، وہ تو نہیں ہے

میں جانتی ہوں کہ میرے خوابوں میں تیری خوشبو کی چاندنی بھی کہیں کہیں تھی
یہ شعر لیکن مرے حوالے سے جس کو تسلیم کر رہا ہوں، وہ تو نہیں ہے

وصال موسم کے خواب میں تھی یا ہجر کے بے نشان دکھ کے عذاب میں ہوں
جو سایہ سایہ اکٹھا کر کے دکھوں سے مجھ کو لگاتا ہے، وہ تو نہیں ہے

نوشی گیلانی

Data Collection



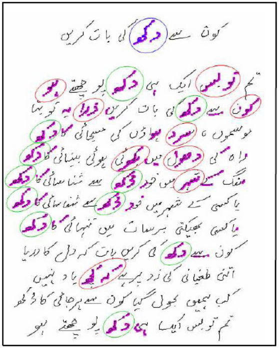
Performance Measures

- Search operation results in
  - matches,
  - mismatches and
  - missed instances
- Retrieval measures:

$$\textbf{Precision} = \frac{matches}{matches + mismatches}$$

$$\textbf{Recall rate} = \frac{matches}{matches + missings}$$

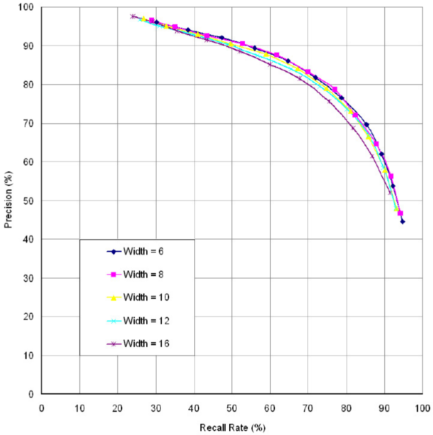
$$F_1 \ \textbf{measure} = \frac{2 \times precision \times recall}{precision + recall}$$

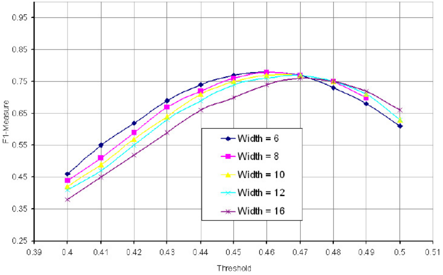


Freeman Grid Codes

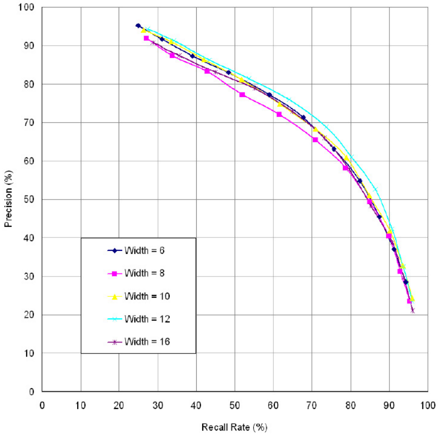
- Square Freeman codes
- Triangular Freeman codes
- Square vs. Triangulare Freeman codes
- Freeman codes: PCNT vs. ioPen

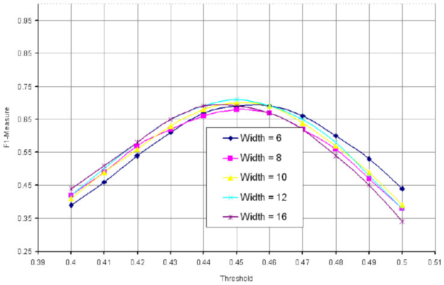
Square Freeman Grid Codes



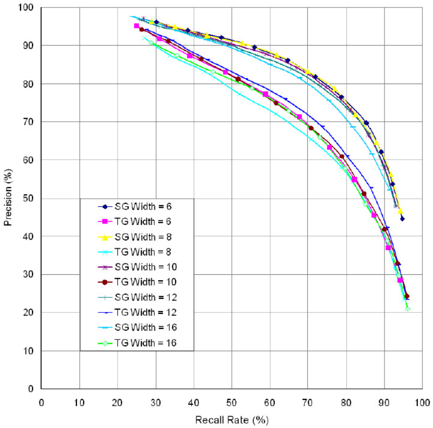


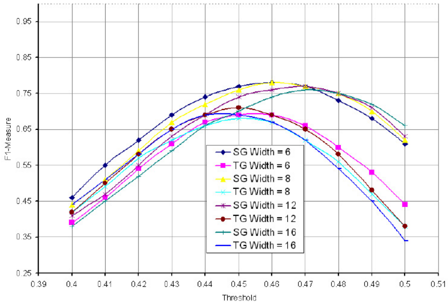
Triangular Freeman Grid Codes





Square Vs. Triangular Grid Codes





Performance with PC Notes Taker Device (PCNT)

	PCNT Device				ioPen Device			
GS	**P**	**R**	F_1	**T**	**P**	**R**	F_1	**T**
6	76.51	78.78	0.78	8458	81.50	81.50	0.81	1555
8	78.68	76.97	0.78	4644	82.30	78.90	0.80	1607
10	78.98	74.80	0.77	2810	78.30	78.80	0.78	572
12	79.47	73.10	0.76	2007	77.10	73.90	0.75	451
16	81.49	67.74	0.74	1326	73.80	71.60	0.72	284

GS = Grid size, P = Precision (%)
R = Recall rate (%), T = Time (milliseconds)

Summary

- **Retrieval System**
  - Approximate string search - retrieval algorithm
  - It works with all kinds of scripts/figures
- **Handwriting Features**
  - **Freeman** to convert handwriting signals to code string
  - Introduced triangular Freeman features: 6 equidistant directions rather than 8 directions of square Freeman features
  - Little performance difference with both types of features
- **PC Notes Taker**
  - To build database, documents written in Urdu & English
  - benchmark: using triangluar and square Freeman features
  - No performace difference from earlier tests with **ioPen**