Technical Presentations

- Page Segmentation (and rule line separation)
- Page Layout Similarity
- Document ID/Script ID

This afternoon

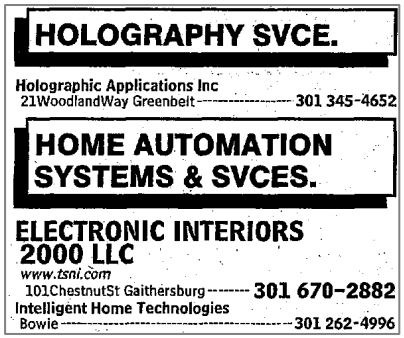
- Signature Detection
- Logo Detection and Recognition
- Stamp Detection
- Font OCR

Observations

- Many documents use structure to convey function
- Salient structure of entries is significantly important to the content parsing
- Traditional document analysis approaches don't capture the implicit repetitive structures that publishers use to convey information

【大冤案】gross injustice
【大元帅】generalissimo
【大员】[旧] high-ranking official: 委派～appoint high-ranking officials
【大圆航向】[航空] great-circle course
【大院】courtyard; compound: 居民～ residential compound
【大约】①(约略) approximately; about ②(很可能) probably

Bilingual Dictionary

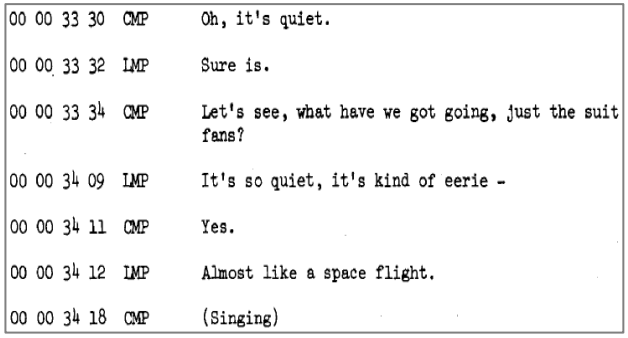


HOLOGRAPHY SVCE.
Holographic Applications Inc
21WoodlandWay Greenbelt ---------- 301 345-4652

HOME AUTOMATION SYSTEMS & SVCES.

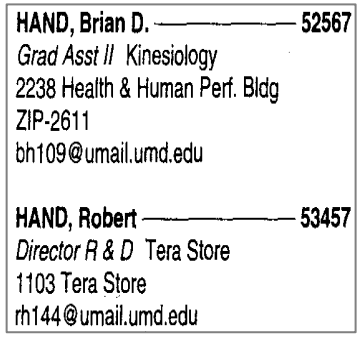
ELECTRONIC INTERIORS 2000 LLC
www.tsni.com
101ChestnutSt Gaithersburg -------- 301 670-2882
Intelligent Home Technologies
Bowie ---------- 301 262-4996

Yellow Book



00 00 33 30	CMP	Oh, it's quiet.
00 00 33 32	LMP	Sure is.
00 00 33 34	CMP	Let's see, what have we got going, just the suit fans?
00 00 34 09	LMP	It's so quiet, it's kind of eerie -
00 00 34 11	CMP	Yes.
00 00 34 12	LMP	Almost like a space flight.
00 00 34 18	CMP	(Singing)

Speaking Transcript



HAND, Brian D. ———— **52567**
Grad Asst II Kinesiology
2238 Health & Human Perf. Bldg
ZIP-2611
bh109@umail.umd.edu

HAND, Robert ———— **53457**
Director R & D Tera Store
1103 Tera Store
rh144@umail.umd.edu

Phone Book

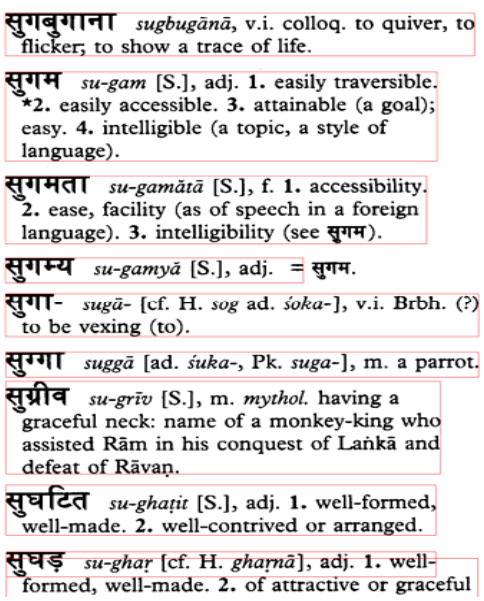
Examples of structured documents

Definition of Success

- **Provide retargetable document analysis capabilities**
  - To improve performance through optimization
  - To provide new capabilities – Scripts, languages, layouts, etc
- **Provide solutions for the desktop**
  - Make use of “user in the loop” to make key decisions in the process
  - Bootstrap training and minimize tedious feedback from user
- **Provide rapid solutions, consistent with task**
  - Layout capabilities in minutes
  - Font (faces and styles) in hours
  - Language capabilities in days

Target Application: Bilingual Dictionaries

- Key source of lexical knowledge for language systems, useful for CLIR, MT and new language learning
- Primary example of structured content
  - Layout
  - Language
  - Tagging



सुगबुगाना *sugbugānā*, v.i. colloq. to quiver, to flicker; to show a trace of life.

सुगम *su-gam* [S.], adj. **1.** easily traversible. ***2.** easily accessible. **3.** attainable (a goal); easy. **4.** intelligible (a topic, a style of language).

सुगमता *su-gamătā* [S.], f. **1.** accessibility. **2.** ease, facility (as of speech in a foreign language). **3.** intelligibility (see **सुगम**).

सुगम्य *su-gamyă* [S.], adj. = **सुगम**.

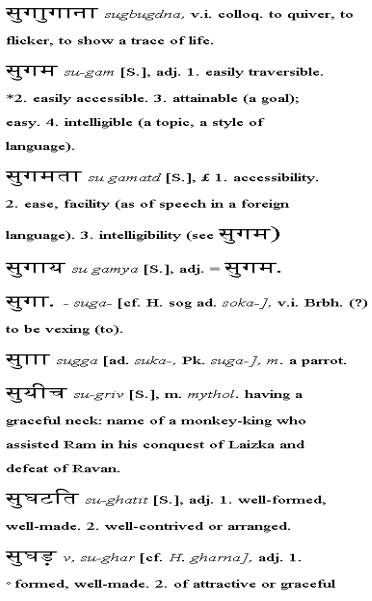
सुगा- *sugā-* [cf. H. *sog* ad. *śoka-*], v.i. Brbh. (?) to be vexing (to).

सुग्गा *suggā* [ad. *śuka-*, Pk. *suga-*], m. a parrot.

सुग्रीव *su-grīv* [S.], m. *mythol.* having a graceful neck: name of a monkey-king who assisted Rām in his conquest of Laṅkā and defeat of Rāvaṇ.

सुघटित *su-ghaṭit* [S.], adj. **1.** well-formed, well-made. **2.** well-contrived or arranged.

सुघड़ *su-ghaṛ* [cf. H. *ghaṛnā*], adj. **1.** well-formed, well-made. **2.** of attractive or graceful



सुगुगाना *sugbugdna*, v.i. colloq. to quiver, to flicker, to show a trace of life.

सुगम *su-gam* [S.], adj. 1. easily traversible. *2. easily accessible. 3. attainable (a goal); easy. 4. intelligible (a topic, a style of language).

सुगमता *su gamatd* [S.], £ 1. accessibility. 2. ease, facility (as of speech in a foreign language). 3. intelligibility (see सुगम)

सुगाय *su gamya* [S.], adj. = सुगम.

सुगा. *- suga-* [cf. H. sog ad. *soka-J*, v.i. Brbh. (?) to be vexing (to).

सुगा *sugga* [ad. *suka-*, Pk. *suga-J*, *m.* a parrot.

सुयीच *su-griv* [S.], m. *mythol.* having a graceful neck: name of a monkey-king who assisted Ram in his conquest of Laizka and defeat of Ravan.

सुघटति *su-ghatit* [S.], adj. 1. well-formed, well-made. 2. well-contrived or arranged.

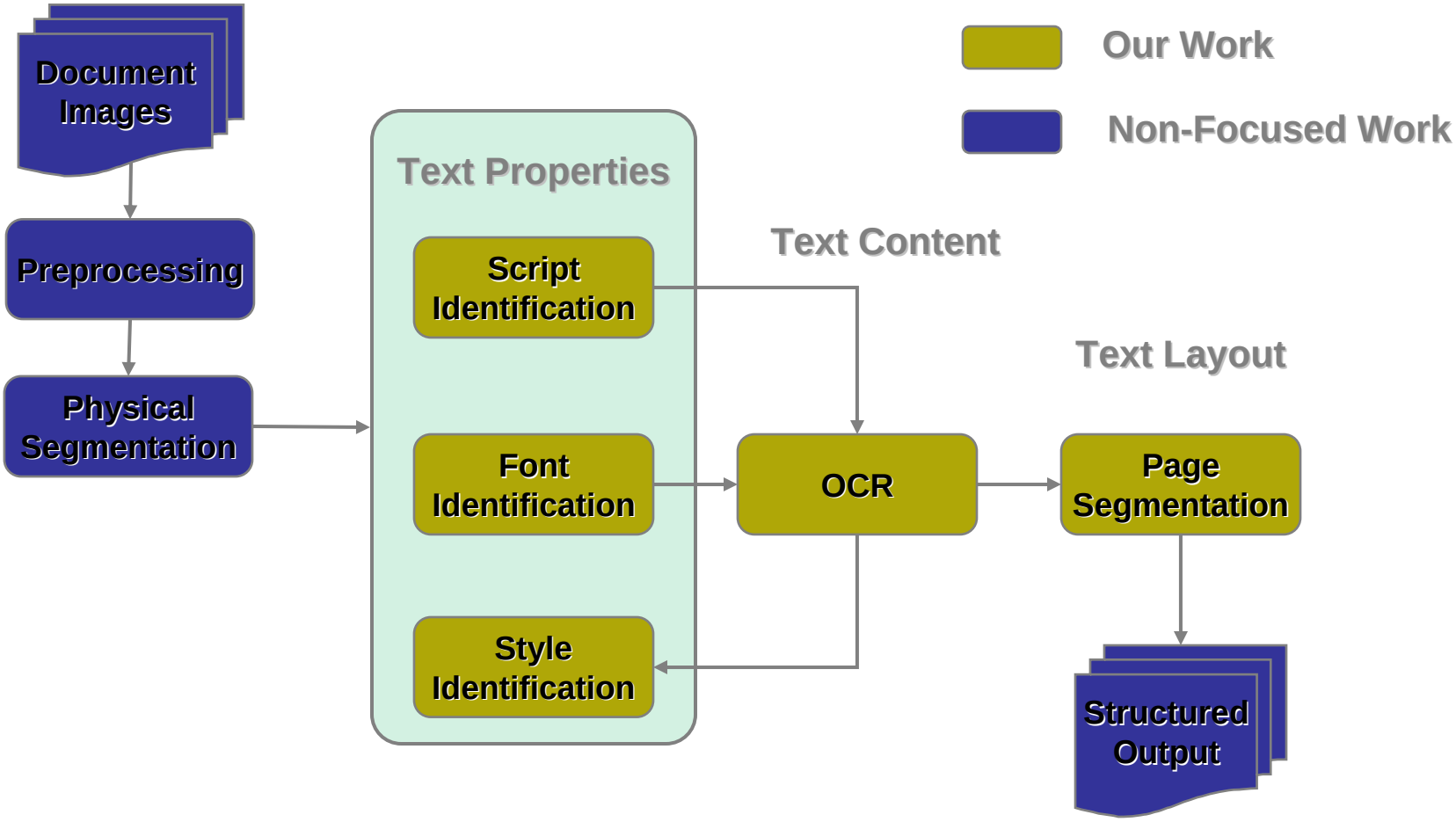
सुघड़ *v, su-ghar* [cf. *H. gharna J*, adj. 1. ◦ formed, well-made. 2. of attractive or graceful

Challenges

- **Layout varies between source**
- **Multiple scripts appear on a single page**
- **Lack of OCR for some languages**
- **Available OCR not optimized**
- **Unreliable font and style information**

- **Goals**
  - **Limited amount of training data**
  - **Make efficient use of user in the loop**
  - **Get it done quickly—new capabilities in 24 hours**

Presented Framework



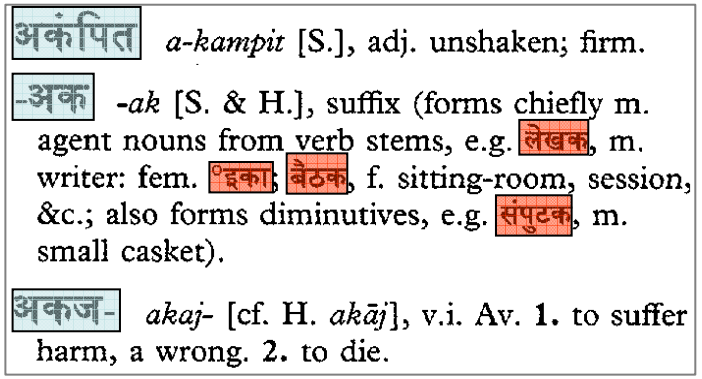
Script Identification

- **Goals**
  - Improve the performance of OCR
  - Useful feature for logical and semantic segmentation
  - Extremely useful for content parsing
- **Challenges**
  - Different scripts are interlaced on a single page
  - Published approaches work at block or page level
  - Different scripts require new features
  - Local features requiring script knowledge not applicable
  - Limited amount of training data

accordingly [əˈkɔːdɪŋlɪ] *adv* 1 照着(办、做等);相应地: We must ascertain the actual conditions and arrange ~. 我们必须了解具体情况,作出相应安排。2 因此;从而: The weather has changed suddenly, and we must alter our plans ~. 天气突然变了,因而我们必须改变计划。

accordion [əˈkɔːdɪən] *n* 手风琴 —*adj* (像手风琴一样)可折叠的: an ~ door 折(叠)门 ‖ **accordionist** *n* 手风琴手 ‖ **accordion pleats** [复] *n* 多道褶裥: a skirt with ~ 百褶裙

Interlaced words with different scripts in an English-Chinese dictionary



अकंपित *a-kampit* [S.], adj. unshaken; firm.

-अक *-ak* [S. & H.], suffix (forms chiefly m. agent nouns from verb stems, e.g. लेखक, m. writer: fem. °इका; बैठक, f. sitting-room, session, &c.; also forms diminutives, e.g. संपुटक, m. small casket).

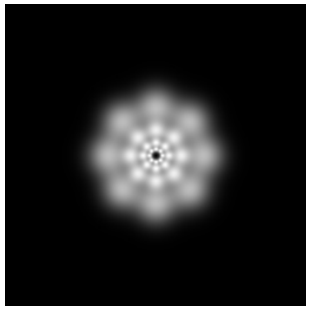
अकज- *akaj-* [cf. H. *akāj*], v.i. Av. 1. to suffer harm, a wrong. 2. to die.

Identified scripts of word is useful for entry segmentation

Our Approach

- Global texture features
- Gabor filter bank
  - Optimal in minimizing the 2D uncertainty in space and frequency
  - Orientation and scale tunable line and edge detector
- Feature vector construction
  - Orientation-invariant

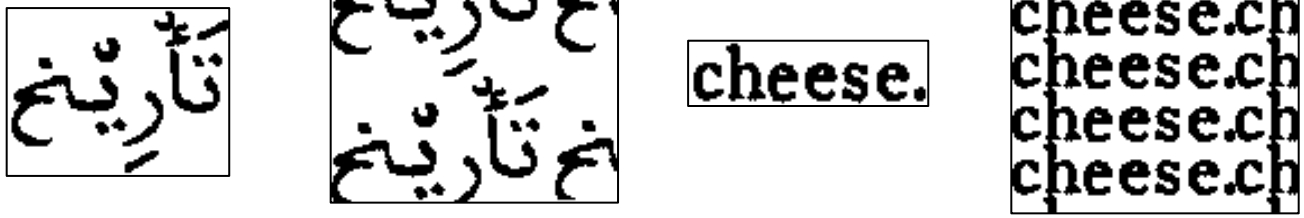
$$\bar{f} = [u_{00}\ \sigma_{00}\ u_{01} \ldots u_{33}\ \sigma_{33}]$$



$$u_{mn} = \iint |G_{mn}(x,y)|dxdy,\ \ \sigma_{mn} = \sqrt{\iint (|G_{mn}(x,y)| - u_{mn})^2 dxdy}$$

Word Image Replication & Scaling

- **Segmented word images are normalized into images with size 64*64 by replication and scaling to guarantee features to be extracted under the same condition.**



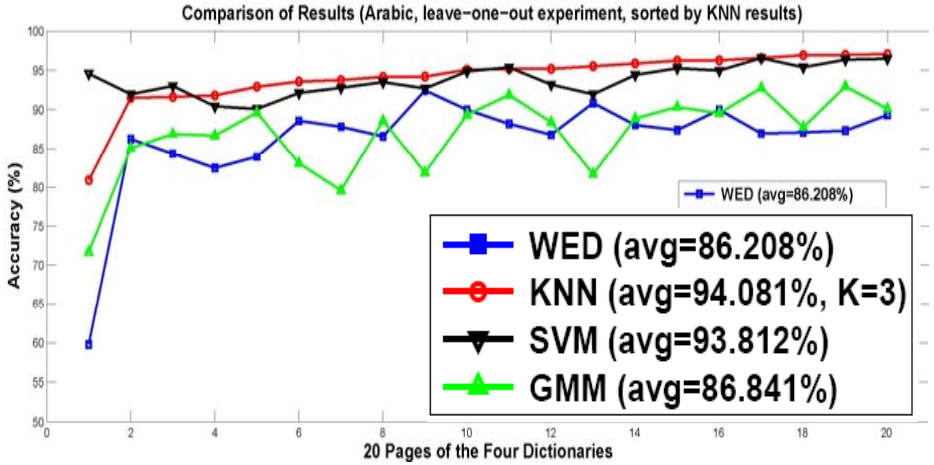
(a) (b) (c) (d)

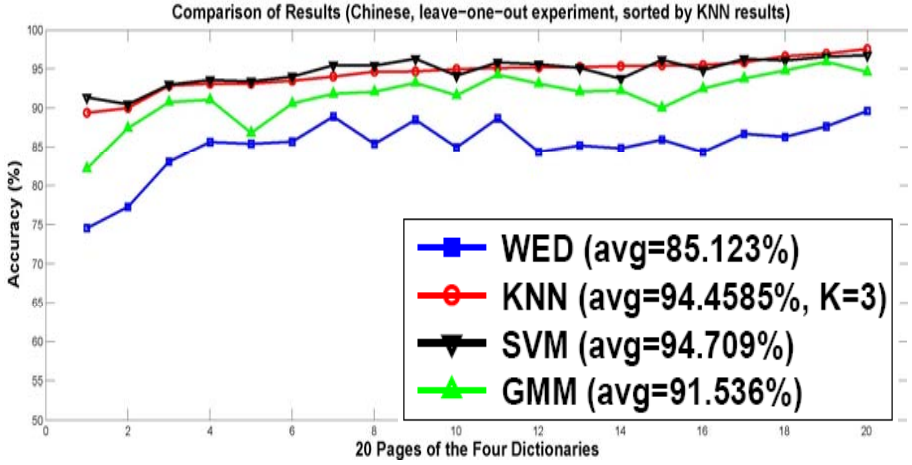
(a,c) Original Image, (b,d) Normalized Image

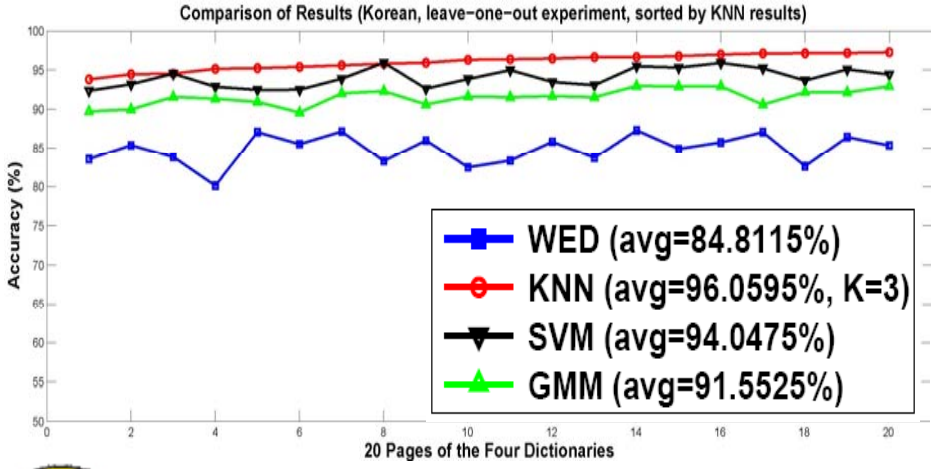
Experiments

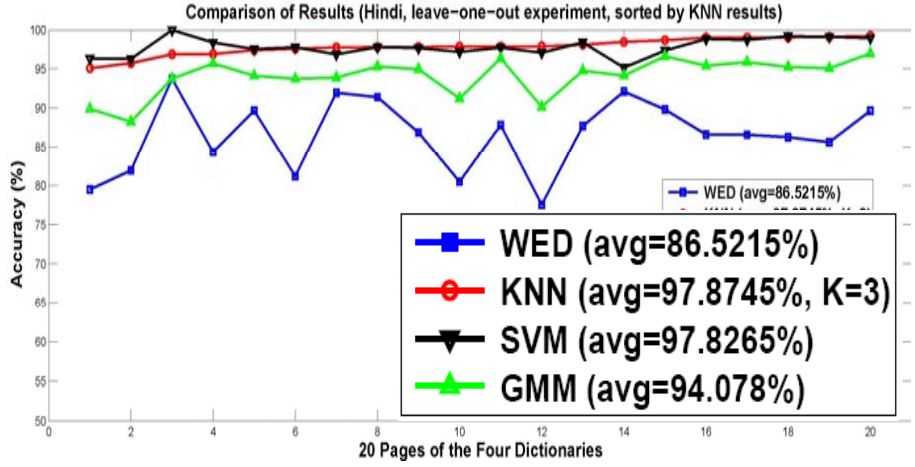
- **Four bilingual dictionaries (each has 20 pages)**
  - Arabic-English
  - Chinese-English
  - Korean-English
  - Hindi-English
- **Four classifiers**
  - Weighted Euclidean distance (WED)
  - k-Nearest-Neighbor (KNN)
  - Gaussian mixture model (GMM)
  - Support vector machines (SVM)
- **Two protocols**
  - Leave one out
  - Single page training

Leave One Out Results

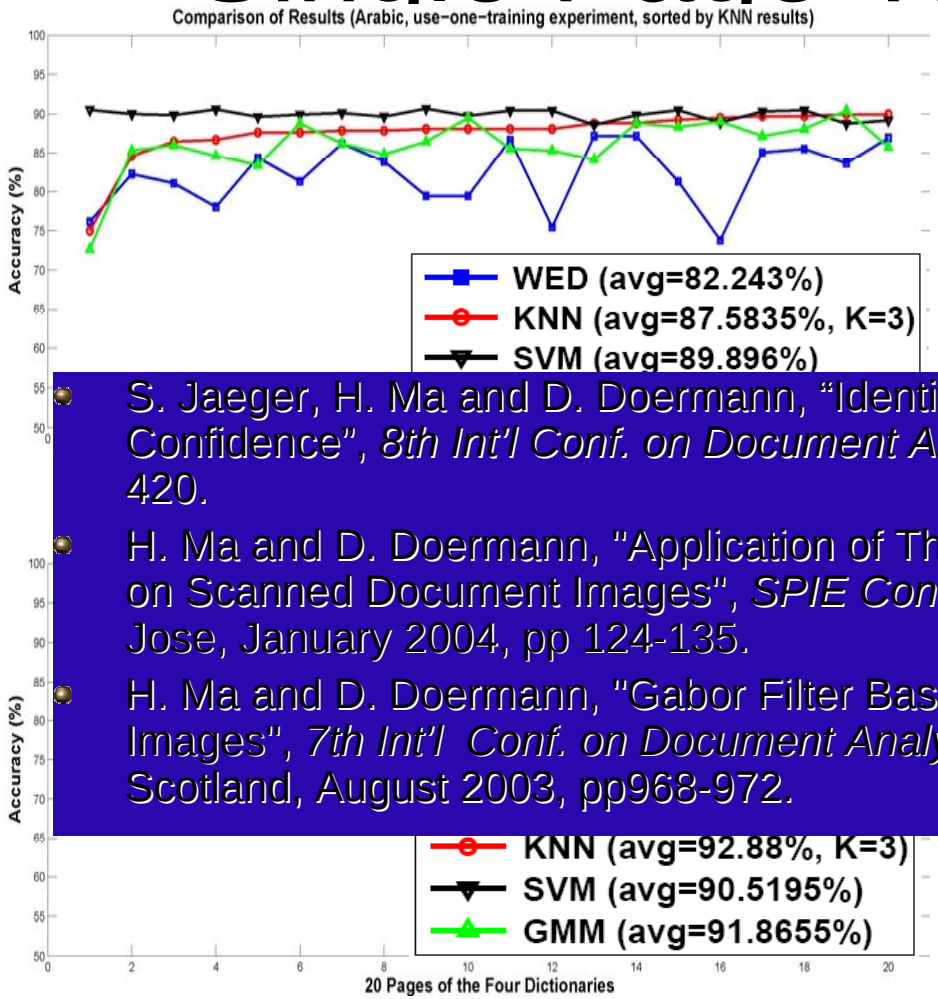


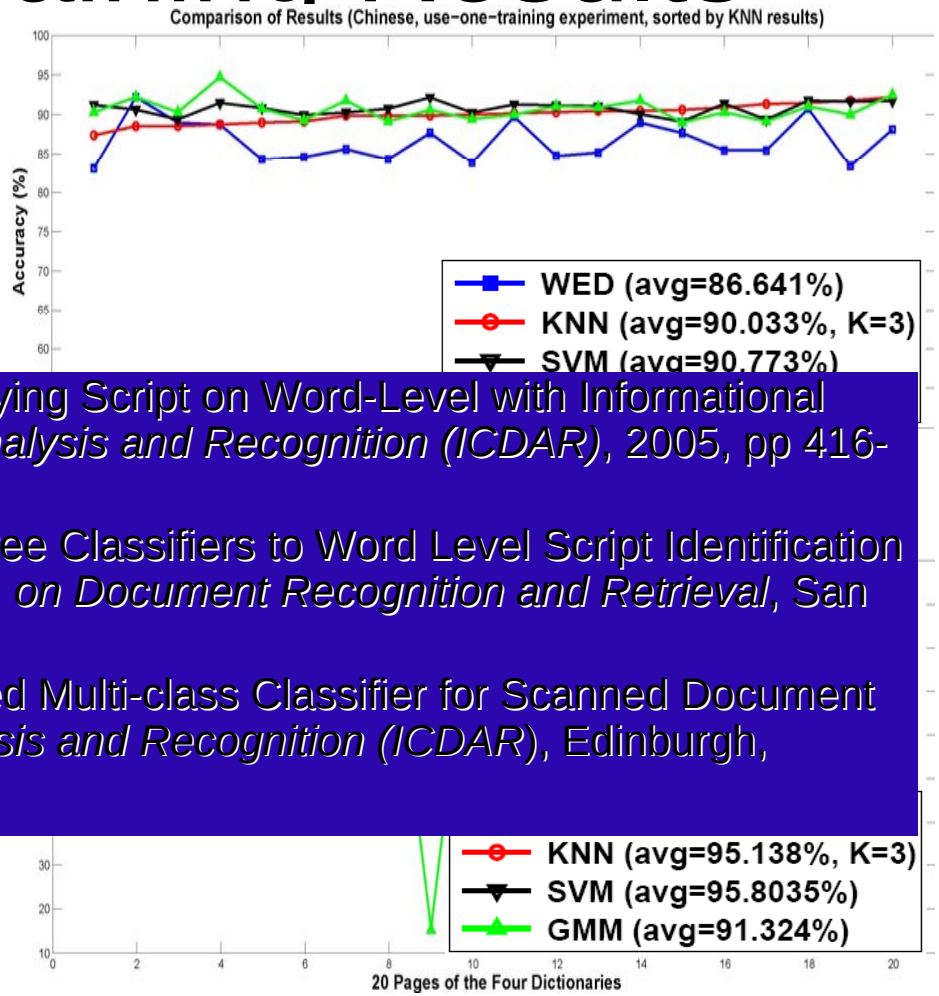






Single Page Training Results





- S. Jaeger, H. Ma and D. Doermann, "Identifying Script on Word-Level with Informational Confidence", *8th Int'l Conf. on Document Analysis and Recognition (ICDAR)*, 2005, pp 416-420.
- H. Ma and D. Doermann, "Application of Three Classifiers to Word Level Script Identification on Scanned Document Images", *SPIE Conf. on Document Recognition and Retrieval*, San Jose, January 2004, pp 124-135.
- H. Ma and D. Doermann, "Gabor Filter Based Multi-class Classifier for Scanned Document Images", *7th Int'l Conf. on Document Analysis and Recognition (ICDAR)*, Edinburgh, Scotland, August 2003, pp968-972.

Agenda

- **Background and Motivation**
- **Challenges**
- **Our approach**
  - **Analysis of text properties**
    - **Script identification**
    - **Font identification**
    - **Style classification**
  - **Recognition of text content**
    - **Adaptive optical character recognition (OCR)**
    - **Automatic training sample creation**
  - **Extraction of page layout**
    - **Bootstrapped logical and semantic page segmentation**
- **Conclusion**

Font Identification

- Locally analyzed typographical features
  - S. Khoubyari, J. Hull (CVIU'96) and H. Shi, T. Pavlidis (ICDAR'97): identify function words such as "the" "of" "and" and perform classification based on the font of these words.
  - A. Zramdini, R. Ingold (PAMI'98): projection and connected component features
- Global texture features
  - Y. Zhu, T. Tan et al. (PAMI'98, PAMI'01): Gabor filter bank to extract texture features.

The identification approach using global texture features is easy to be tuned to work on different scripts and fonts.

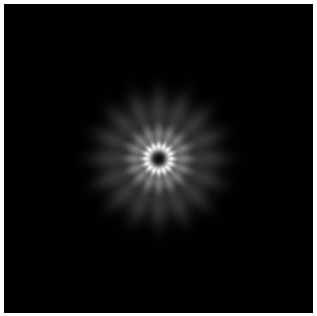
Our Approach

- Also use global texture features
- New texture operator: grating cell operator
- New classifier: back propagation neural network (BPNN) to replace the simple weighted Euclidean distance (WED) classifier
- Result comparison with Gabor filter bank texture operator

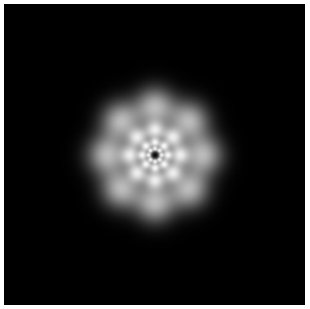
$$g_{\lambda,\theta,\varphi}(x,y) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot exp(-\frac{x'^2 + \gamma^2 y'^2}{2\sigma^2}) \cdot cos(2\pi\frac{x'}{\lambda} + \varphi)$$

$$x' = x \cdot cos\theta - y \cdot sin\theta$$

$$y' = x \cdot sin\theta + y \cdot cos\theta$$



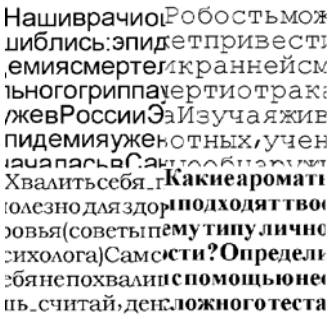
Response of Gabor function



Gabor filter response

Frequency Selection

- Operator is slow
- Speed is crucial
- Minimize number of frequencies
- Maximize the performance



4-font Example

Ground Truth

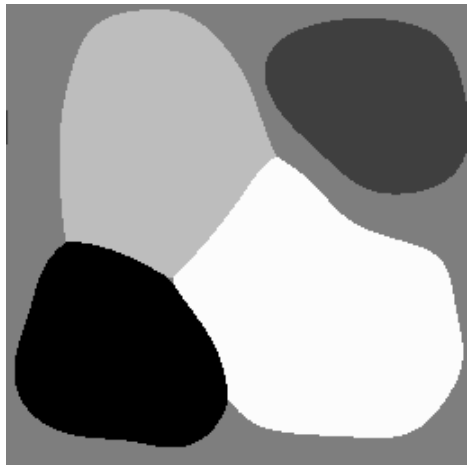
3-frequency Result

4-frequency Result

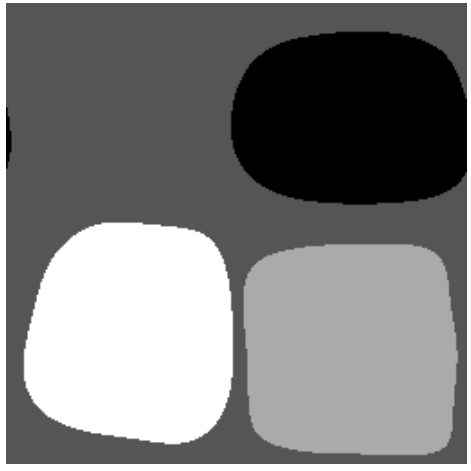
Cyrillic Example

Cyrillic Font Images **Ground Truth** **Result of Grating Cell**

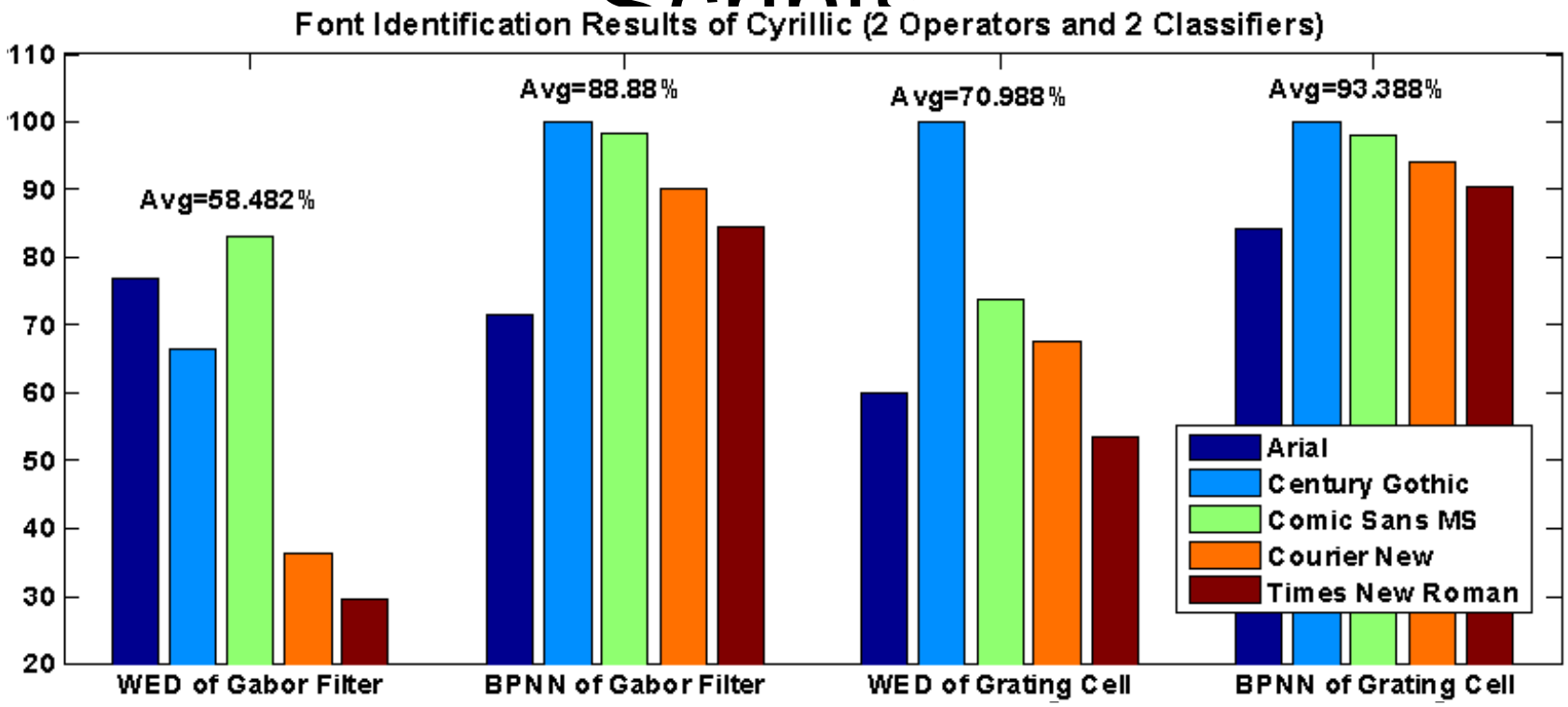
5 Fonts



4 Fonts



Results of Five Fonts, Three Scripts



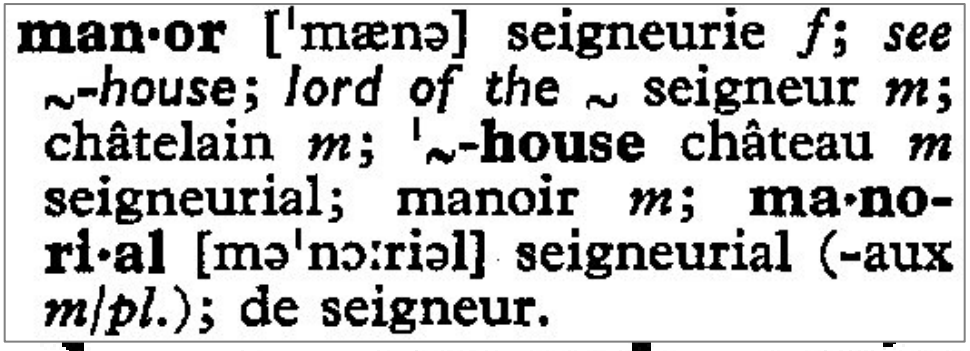
- H. Ma and D. Doermann, "Font Identification of Scanned Documents Based on Texture Features Using a New Texture Operator", *SPIE Conf. on Document Recognition and Retrieval*, San Jose, January 2005.

Agenda

- **Background and Motivation**
- **Challenges**
- **Our approach**
  - **Analysis of text properties**
    - **Script identification**
    - **Font identification**
    - **Style classification**
  - **Recognition of text content**
    - **Adaptive optical character recognition (OCR)**
    - **Automatic training sample creation**
  - **Extraction of page layout**
    - **Bootstrapped logical and semantic page segmentation**
- **Conclusion**

Style Classification

- **Motivation**
  - Provide additional implicit information
  - Represent different functionalities
  - Feature useful for page segmentation and content parsing
  - Significantly depends on image qualities

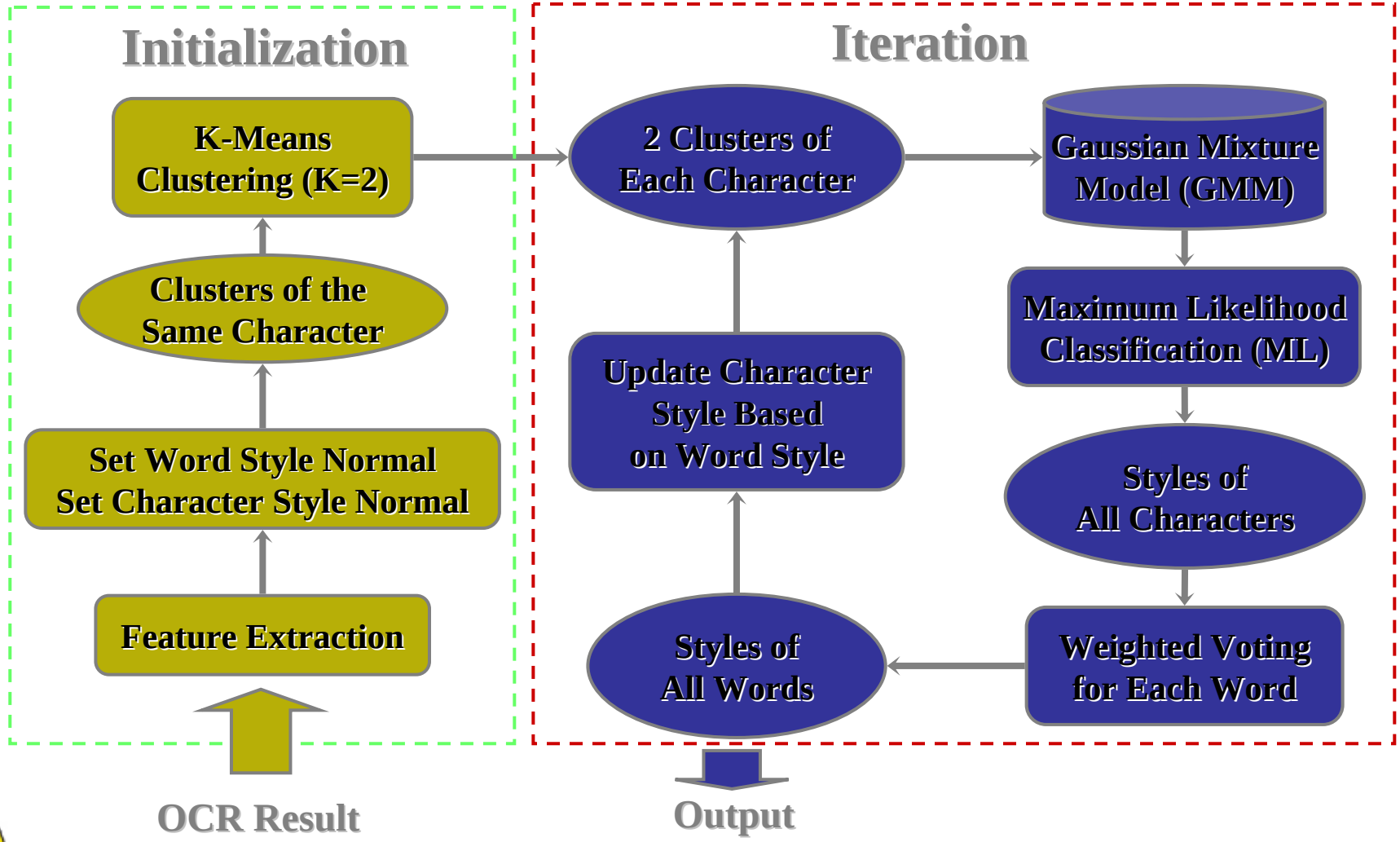


One entry of an English-French bilingual dictionary

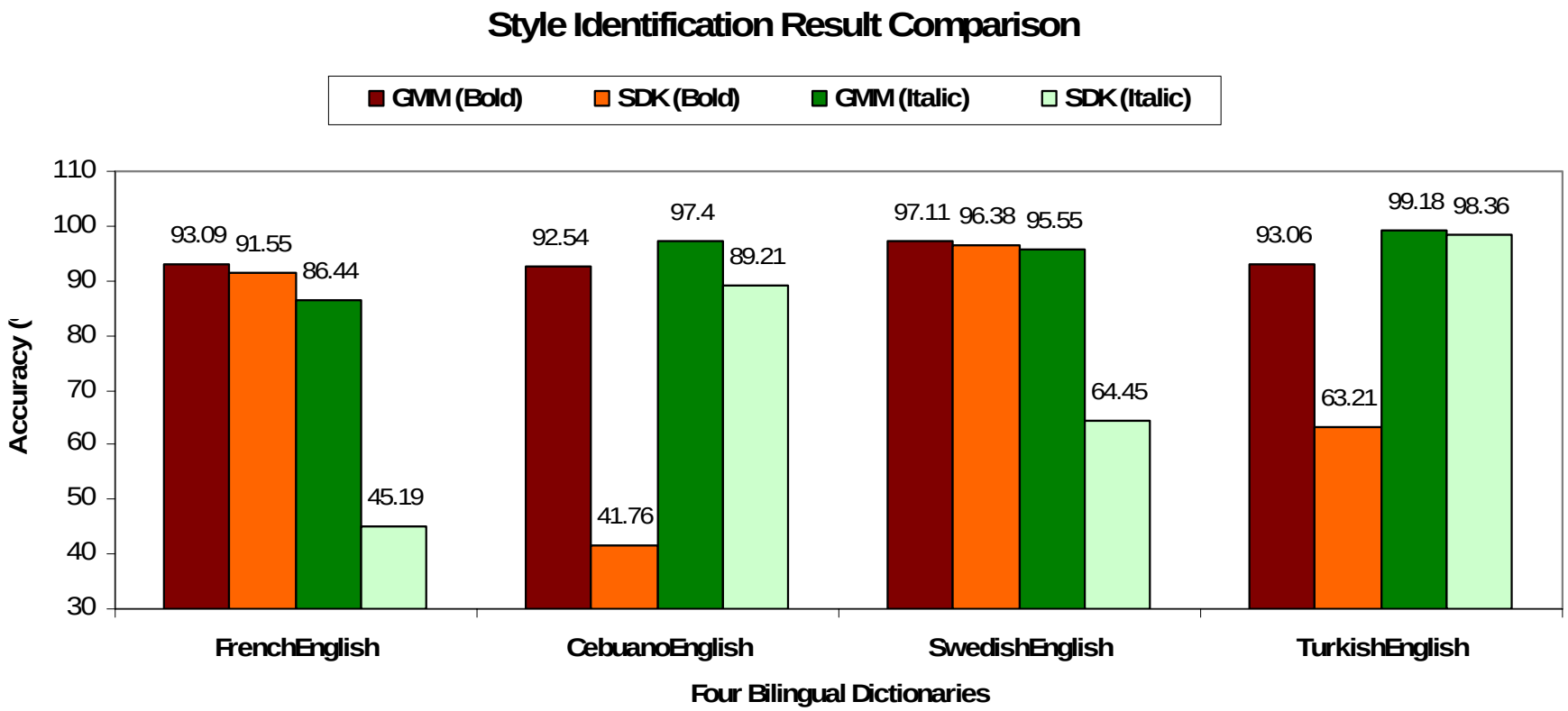
Possible Features for Style Identification

- Stroke width
- Foreground density
- Aspect ratio
- Vertical skeleton pixel ratio
- Possible slant angle
- Nine-zone foreground density

Approach for Style Identification



Experimental Result Comparison



- H. Ma and D. Doermann, "Adaptive Word Style Classification using a Gaussian Mixture Model", *Int'l Conf. on Pattern Recognition (ICPR)*, Cambridge, UK, August 2004, pp 606-609.