MIN Faculty Department of Informatics



# Developments in Computersupported Palaeographic Analysis

Prof. Bernd Neumann, Ph.D.

Department of Informatics University of Hamburg

International Workshop "Digital Support for Manuscript Analysis" Hamburg, 23-24 July, 2010



# Computer Vision, Artificial Intelligence and Palaeography

**Computer Vision** 

**Artificial Intelligence** 

Image Processing Pattern Recognition

Image Understanding Knowledge Representation

**Probabilistic** 

**Models** 

Shape Description

Content-Based Image Retrieval Handwriting Recognition

Forensics

Datamining Le

Learning

**Palaeographic Applications** 





- Image restoration and segmentation
- Writer identification
- Content-based image retrieval
- Computer-based manuscript analysis in Hamburg

**MIN Faculty** Department of Informatics



# **Restoration (1)**

Removing background variations and noise by anisotropic diffusion (Moghaddam & Cheriet 2009)



64

**MIN Faculty** Department of Informatics

(2.9

C31



#### **Restoration (2)**



MIN Faculty Department of Informatics







# **Subpixel Watershed Segmentation**

- A Generate continuous image by spline interpolation between pixels of original image
- B Determine gradient image by differentiating the (analytical) continuous image
- C Trace maxima in gradient image (watersheds")
- D Remove weak edges



**5th order spline interpolation** 





# **Subpixel Segmentation (2)**

Original



**Subpixel Watershed Contours** 







### **Segmentation of Degraded Text**



Uninformed segmentation methods will fail





# **Segmentation Using Shape Priors**

Learnt shape models bias variational segmentation towards realistic results (Bar-Yosef et al. 2009)

original



#### segmentation without shape knowledge



segmentation with shape knowledge





### State-of-the-art in Restoration and **Segmentation**

- Sophisticated mathematical models
- Incorporation of domain-specific knowledge
- Learning and statistics
- Advanced tools





## A Case of Writer Verification (1)

Two pages of a satiricical newspaper of 1846, allegedly handwritten by the New England author Hermann Melville





the Yok Shat 20 S 166C



#### A Case of Writer Verification (2)

Documents known to be written by Melville

lingues (in any 1537 my Can denote - + for with Robert of a speak a been the or advert and I an when my school deald have gove into fronting . and notice a suppose countral of price Witnesser approximations apart your request. Still, however I denside us my prover proving - 5 it is with pleasant the of the I should have taking any here at an earlier bay has not the savity you furtance of and property; that they about a large further of any that more , having because demanchas mayn anded with the unkine of towning - homing established " systim is my minter of instruction - and tring for industrial the the characters & Hafrantins of my wohllers. in wheel, having the new school and, a proper encours 1 sagin \_ a few sultions of since are affected in other Confirm By socialized withing & country the all ages, sides, hands, shereday, a stranding some of the als have a signed the ages of sighter can not be a down in assure, while others have growing things the to the adies. but with the count simplify that they are not working a tight I have as though they have asses paper that tony before. they exhed a situation on a month & replaced hast of the in about for smiles from the sillage, and the house at which an

Mr. prease dicken and What a shapenen want is good The most enough for that in my while fearly under - I don't have then the pleasants this I have along has my product to the production applied of was not more in making the "Herman to be trained or south . Foline gard Capitrating & Material , 40 rades There of That is all of the in wish the in the State & most indere The on Cash they well the I meanaby hidreds to the wave to all an anone of delighter characteristics. her than strong of gentled adde maker were the at all some applied to the other that - don the Calo Sign see are a first and save star distance services by firsting attactions - There the Warter as a gential who are demined hundrene that if no any with repeart of littles, tranks as here of ingene life ?? or desidution they while doubt to forme of the car line town good boots, and as very rate they set are pouring sigly ..... lind ever shutters about sounds for a Saids " & "Handlows in This Apardson's dea" & ale This set of this is go of the Halls who rely



# Writer Verification with CEDAR-FOX (1)

**Center for Excellence in Document Analysis and Recognition** Sargur N. Srihari, University at Buffalo

**Preparing a document for analysis** 

(Ball et al. 09)

- Manual processing: Remove non-text and major noise ۲
- Preprocessing: Binarization, line and word segmentation, computing ۲ global document features, automatic character recognition
- Manual correction of word segmentations ۲
- Transcript mapping to obtain correct ground truth for words •

Word images of

- known documents
- unknown documents

with wi

with with



# Writer Verification with CEDAR-FOX (2)

Comparison of corresponding characters based on GSC features:

**Gradient-based:** 

- A 4x4 grid is placed over the character, dividing it into 16 subfields
- Gradient directions are counted in each subfield for 12 directions
- Direction frequencies above a threshold receive 1 feature bit
- $\Rightarrow$  192 bits

Structure-based:

Presence of corners, diagonal, vertical and horizontal lines

 $\Rightarrow$ 192 bits

**Concavity-based:** 

Major topological and geometrical features, direction of bays, presence of holes, large vertical and horizontal strokes

 $\Rightarrow$ 128 bits

Similar approach extended to comparison of words



# **State-of-the-art in Writer Verification**

- Handwriting recognition technology adapted to requirements in forensics
- 98% reliability for forensic writer verification tasks in USA
- Non-interactive black-box approach
- Not immediately applicable to other writing systems
- Incomplete prototypical systems for palaeographical applications

MIN Faculty Department of Informatics



#### **Content-based Image Retrieval**

# Determine occurrences of example image in large database





العرعليه وماتهم



# **Object Recognition Using SIFT Features (1)**



- SIFT (Scale Invariant Feature Transform) features are gradient-based descriptors evaluated at significant image locations ("interest points")
- For content-based image retrieval
  - database images are annotated with SIFT features
  - SIFT features of input image are compared with SIFT features in database

#### **Object Recognition Using SIFT Features (2)**





# **Word Image Retrieval Using CEDAR-FOX**

1024 bits characterizing the word "Cohen" (Zhang et al. 04)





## **Similarity Measure for Binary Feature Vectors**

**Correlation of binary vectors:** 

$$D^{b}(X,Y) = \frac{1}{2} - \frac{S_{11}S_{00} - S_{10}S_{01}}{2((S_{10} + S_{11})(S_{01} + S_{00})(S_{11} + S_{01})(S_{00} + S_{10}))^{1/2}}$$

S<sub>ii</sub> counts the number of matches with i in the first and j in the second pattern at corresponding positions (i, j = 0, 1)



## State-of-the-art in Content-based Image Retrieval

- Large efforts underway for web applications
- Promising solutions based on precomputed descriptors
- Special-purpose developments for manuscript analysis



# **Computer-based Manuscript Analysis** in Hamburg

Part of the Reserach Group "Manuscript Cultures in Asia and Africa" www.manuscript-cultures.uni-hamburg.de

9 projects for

Indology, Sinology, Iranistics, Japanology, Islamic Sciences / Arabistics, Sanskritistics, Tamilistics, Tibetology, Ethiopistics

**1** project for Informatics



#### **Cognitive Systems Laboratory**

	1971	Image Sequence Analysis			
ł	- 1980	Natural-language Description of Image Sequences	Real Providence		
+	. 1990	Model-based Scene Interpretation	Radiological Image Analysis Image Registration	Aerial Image Analysis	Applications of Knowledge-based Systems
+	. 2000	Learning for Scene	Subpixel Segmentation		Description Logics RACER
	. 2010	Real-time Scene Interpretation	Topology- preserving Sampling Manuscript Analysis	Ocean Current Analysis	Semantic Information Processing



## **Research Goals**

Development of innovative image processing methods for:

- Computing layout and character features for comparative • manuscript analysis
- **Retrieving example patterns from large manuscript databases** •
- Using discriminative features to determine •
  - related cultural origins of manuscripts
  - identities of scribes
- Contributing to a toolbox for palaeographic research ullet

MIN Faculty Department of Informatics



#### Are Hands A and B the same?

Hand A:

Hand B:









Richter 2006: "Tentative Criteria for Discerning Individual Hands in the Guodian Manuscripts"

MIN Faculty Department of Informatics



#### **Comparison of Stroke Angles**

Hand A:





Hand B:







#### Analysis of Column and Line Structure by Anisotropic Filtering

**Column structure** 

Line structure







# **Determining Character Regions**

Automatically determined column and line structure

and the second second		State of the local division of the local div	and the second second	_	_	1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	and the second second	and a start of the start of the		-	Company of the local	
to	好	直	晋	非	佛	杏	充	故	帮	7	生	於
想	者	者	者	者	d-in	- p	佛	吉	故	佛	善	14
忠	讀	請	清	請	欲	能	角	南	A	者	A	E
者	是	是	是	是	求	3	九	无	(年)	,曾.	浙	1
言語	經	錘	妊	征	凌	九	者	佛	充	11	14	守
#	石头	AX.	欷	欲	世	Ŧ	為	南	佛		2	待
紅	得	求	衣	永	3	稱	凝	者	春)	de j	D	家
不	释	黯	佛	主	責	每	怖	歸	者	充	皆	12
谷民	-	あ	道	天	者	充	佛	充	礼	佛	稱	古
作	-	者	者	上	語	佛	吾	者	东	南	阐	自
寒	首	清	請	者	-	即	A	单	者	者	充	救
科	者	星	0		鉅	得	晋	佛	待	(1)	佛	便
部	請	鲑	22	-	補充	*	F	者	伟	开	南	不
者	A	欲	故	41	永	尾		明	者	者	者	兆
語	虹	求	永	海火	这	1	狼	故	唐	÷	易命	集
是	领	錦	和	太	世		-	言	建	佛	充	自
雞	得	政	陕	4	t	-	佛	5	教	者	者	致



# **Shape Context Similarity**

- Determine "Shape context" of a contour by marking each ۲ contour point with the number of contour points for each of 32 relative locations.
- Determine optimal correspondence of contour points of two ulletcharacters by comparing shape contexts
- Remaining difference of shape contexts is similarity measure a •





#### Finding Similar Characters by Shape Context Comparison

Identical characters

# Similar characters sorted by shape context similarity



腹 神 耳 皆 2 戡 堆 出 當 生為 用 五 百 獄 億 下 酒為思 者 遇 者 中 常當愚寢无匹 去 常 受形 便 田 如是 飲消 死 踞 乃竟 者 求 浅 銅 生 魂 消 吏 不 世 銅 包思 玄 當 親 大口 得 1 厚 口堆 鞭榜 太 和 径 31-識 4 有 E 围



#### Weakness of Contour-based Similarity **Analysis**

Contour comparison is sensitive to

- connectivity disturbances, no graceful degradation =>
- contour protrusions (e.g. at junctions) =>





# **Stroke Analysis by Triangulation**

Constrained Delaunay Triangulation (CDT) connects contour points to triangles such that the circumference of a triangle contains no other points.

- CDT generates three types of triangles:
- junction triangles (green) none of the triangle sides coincides with the contour
- sleeve triangles (blue)
- terminal triangles (red)



Junction triangles indicate stroke intersections or sharp stroke corners





A curved line with angle  $\alpha$  and outer contour radius R, drawn with a stylus of radius S, will generate a junction triangle if

 $S > R/2 (1 + \cos \alpha/2)$ 



### **Weak Influence of Contour Point Spacing**



dense spacing

medium spacing

coarse spacing no junction triangles if corners are cut



# **Stroke Segment Merging**



- Segments meeting at a junction may be merged if they are compatible regarding orientation and stroke width
- Segments between two neighbouring junction triangles may be intersections with irregular direction and stroke width
- Global criteria and knowledge of the writing system must be invoked to resolve ambiguities

MIN Faculty Department of Informatics



#### **Results of Stroke Analysis (1)**



MIN Faculty Department of Informatics



### **Results of Stroke Analysis (2)**





#### **Database of Chinese Characters**

#### 339 characters, ca. 60 x 60 pixels each

四	者	致	寤	来	是	腹	中	搭	離	种	面	卧	者	沙	4	見	身	醉	奴
蜚	日	10	无	出	影	堆	當	耳	2	珩	戎	時	醉	門	モ	其	曹	囚	畑
弟	18	县	TT	生	+	铜	挖	the	山	用	得	如	便	道	者	吠	加	語	Ż
トス	-	公	訩	×	倍	T	秋	五	四	酒	酒	死	如	人	醉	廿	被	吉	PF
」開	六	EE	生日	一人	万	调	+	者	者	五	疽	ん	狂	*	便	六	病	衝	輕
四日	*	5	ム	常	出	+	世	醉	醉	四思	症	无	顛	12	不	者	1	D	悔
170	健	T	盐	营	四	如	中海和	便	便	王	任	74	人	者	欲	醉	五	而	++
ド丁	-32	順	T II	里	又形	县	消	死	踞	-	机	識	ト	醉	錘	便	者	出	=
可し	西九	酒	日	恐方	R	求	铜	猫	福	者	病	和	見	便	法	欲	醉	+	者
EL	·赵王 -34	江南	12	彩	音	4	345	魂	長	醉	山	世	2	好	不	前	便	11	醉
入	日山	石石	政	九	7	不	月年日	白田	吏	便	-	-	旺	沃	散	马	吐	者	便
低き	雨	H	世	44	六	得		正	致	朝	者	者	一書	东	明	唐	洋	醉	家
書	大林		伯人	湖	者	志	H	面入	得	庙	醉	醉	H	PF	錘	师	田	便	宝
TE	孔	4	op	A	裕	51-	1	t	御	行ち	便	便	者	里	酹	很天	心雨	臣卜	王
化	祥	5 Esta	飲	田	the	T	难	1	下语	識	天	载	醉	品生	者	THE	命も	賺	大
1	雨田	雨	西九	有	秋	得	~	地	A	E	部	得	便	T	不	里	女子	瞢	循
7	他	西左	T	馬	中	20	腹	私	得	漆	思	雪	臣	n	教	后降	西	時	如
State -	TT	HT	F)	12	A STATE	P. C.K.	1000	Stand and	1.6	1	1.0	0	and the second	and the second	Street 1	-		12	



# **Results for All Characters of Database**

#### Stroke recognition rate decreases with character complexity





### Comparison of Characters Based on Stroke Structure Graphs

Stroke structure graph:

nodes = qualitative descriptions of individual strokes

edges = qualitative spatial relations between pairs of strokes



Comparison of two stroke structure graphs by determining best-matching common subgraph





# **Stroke Decomposition for Tamil Characters**

- Decomposition of a stroke into sections of qualitatively equal curvature
- Qualitative stroke representation in terms of a sequence of equal-curvature segments
- Comparison based on stroke structure graphs with qualitative stroke representations





# **Summary**

- Computer Vision and Artificial Intelligence provide powerful methods which can be harvested for palaeographic applications.
- Handwriting analysis in forensics suggests methods worth looking at. There exist advanced systems for the analysis of Latin handwriting. Blackbox approach may not meet requirements of palaeographers.
- Building an advanced toolbox for palaeographic applications requires close cooperation of Computer Vision and palaeographers (and sustained efforts over many years).