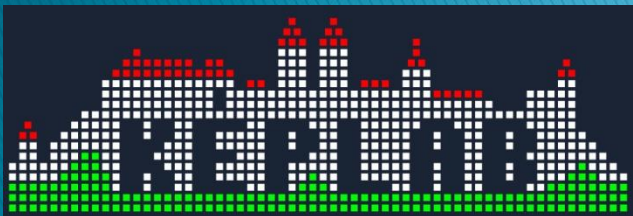


# Processing Historical Documents

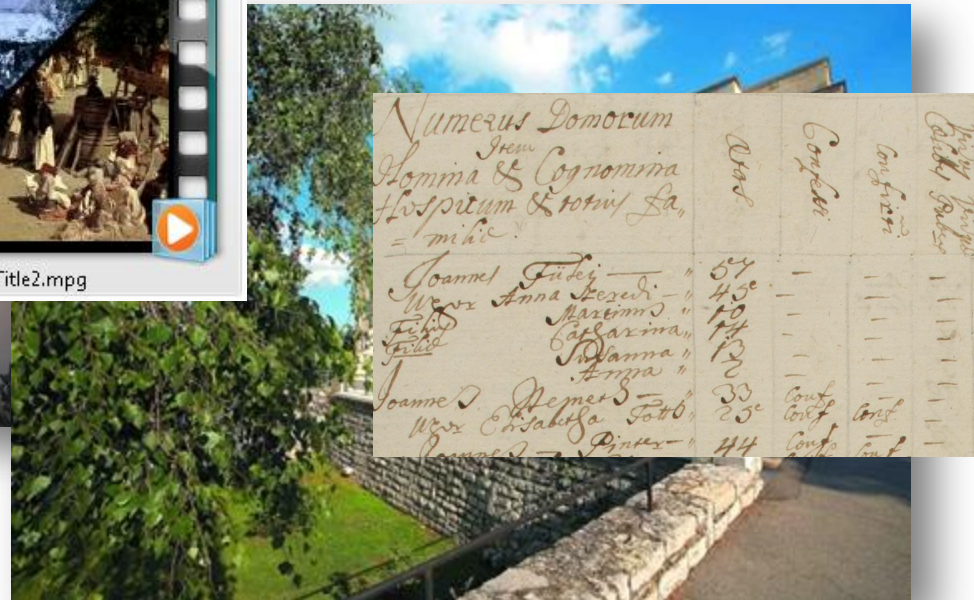
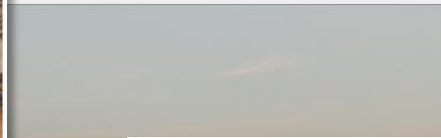
-

## Handwriting Recognition in Historical Documents

Image Processing Laboratory  
Department of Electrical Engineering and Information Systems  
University of Pannonia



# Present and past activities for the preservation of the cultural heritage at the UofP



**Virtual Environments and Imaging Technologies Research Laboratory  
and  
Image Processing Laboratory**



# Optimal Lighting for the Sixtus Chapel

Large surface of paintings of  
Michelangelo, Botticelli, Perugino,  
Domenico Ghirlandaio and others

Reconstruction of the original colours  
under special conditions

3000-3500K colour temp. instead of  
6500K

7000 LEDs

80 % reduced energy consumption





# Results of Optimized Spectral Distribution Lighting in the Sixtus Chapel



Non-optimal

Optimal LED lighting

<http://vision.uni-pannon.hu>

# Restoration of Historical Films

## Typical problems:

- ❖ Vibration→ stabilization
- ❖ Missing dye, dirt, blotches  
→ blotch detection , inpainting
- ❖ Colour fading→ reconstruction
- ❖ Flickering→ correction
- ❖ Scratches→ filtering, inpainting

## Hardware Developments at MTA Sztaki:

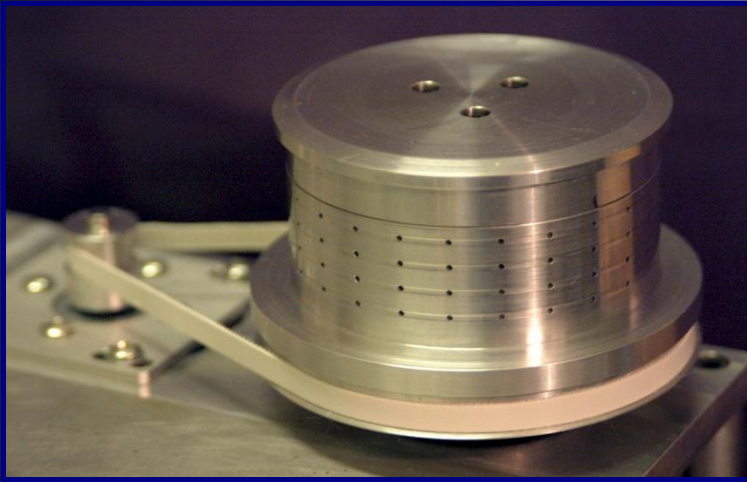
- ❖ High resolution film scanner (2K – 6K)
- ❖ Optical audio reconstruction



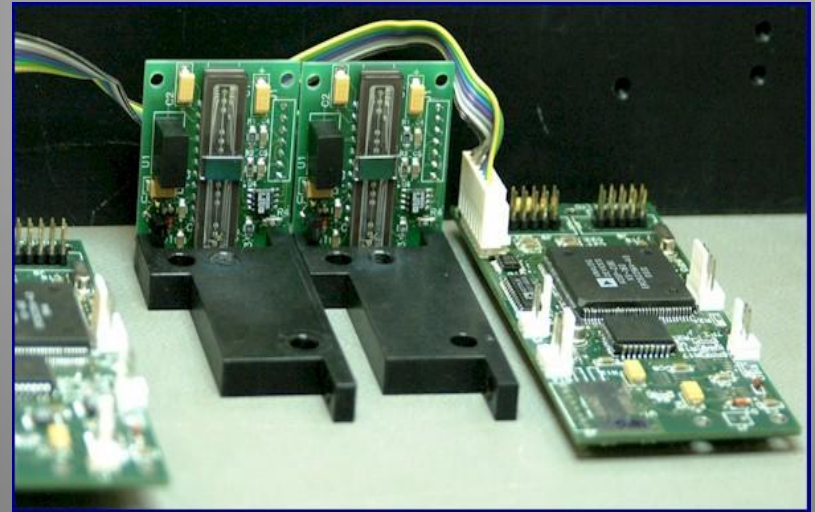


# Special hardware for historical films

Vacuum transfer



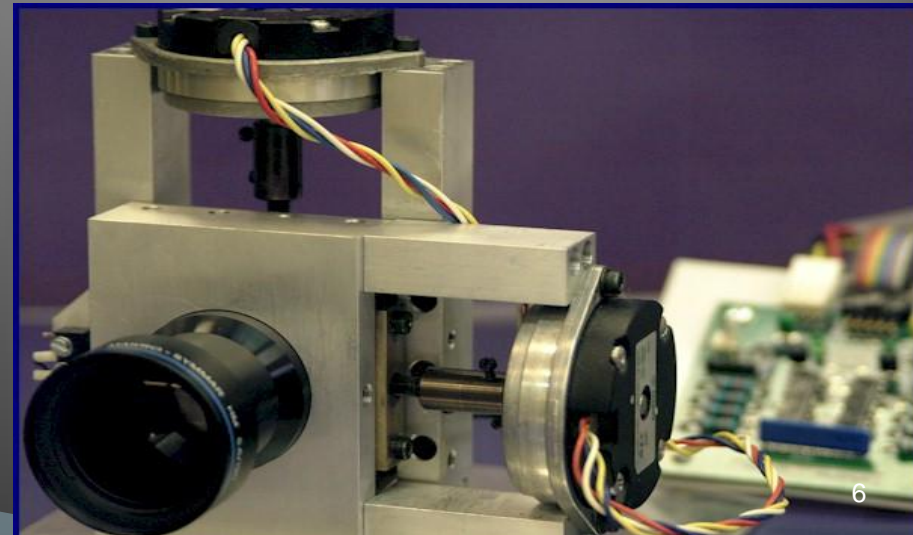
Optical sensors



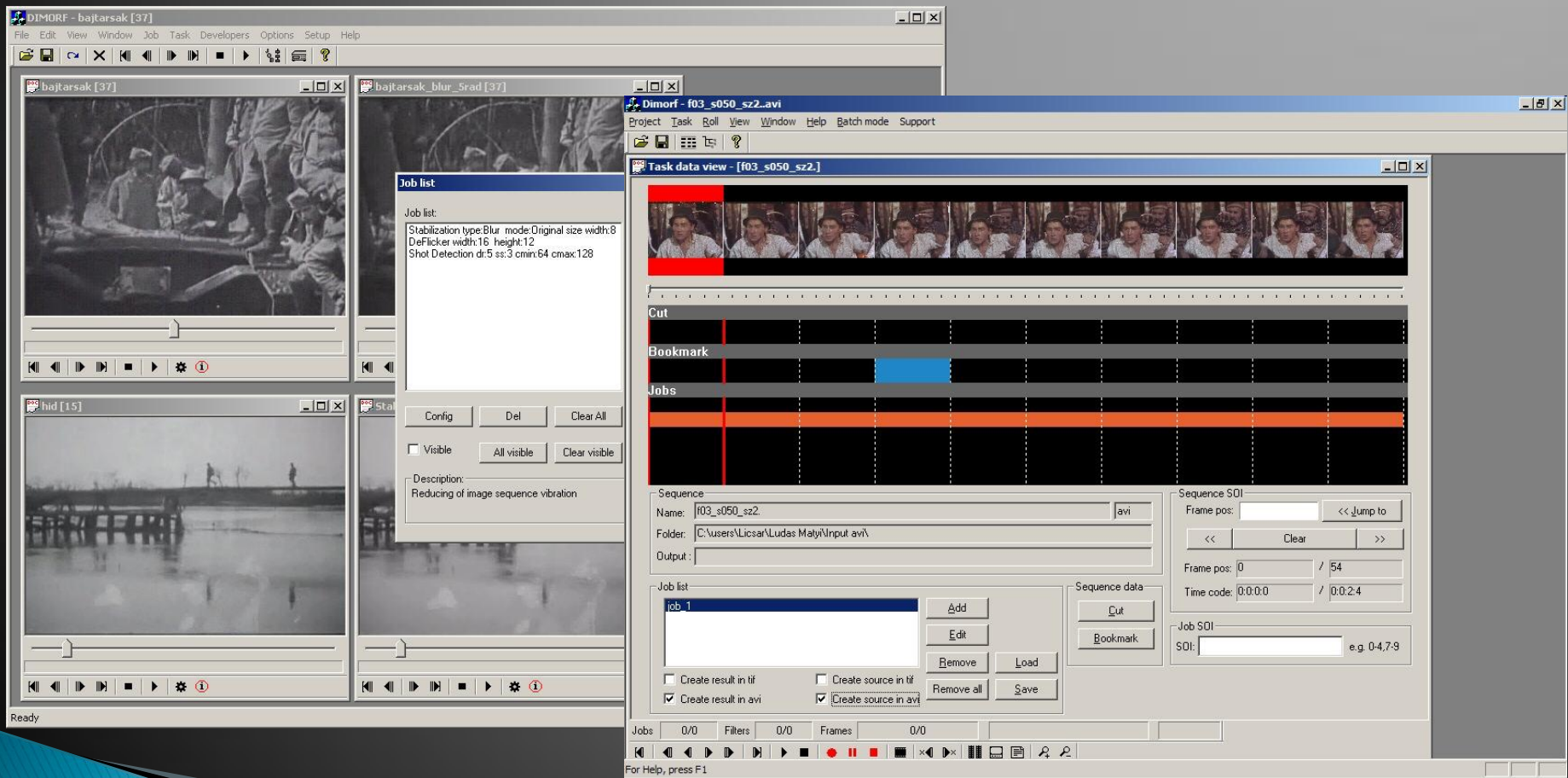
Sound digitization



Positioning of optics



# Film restoration software and algorithms





# Colour reconstruction





# Spatial/Temporal information based processing



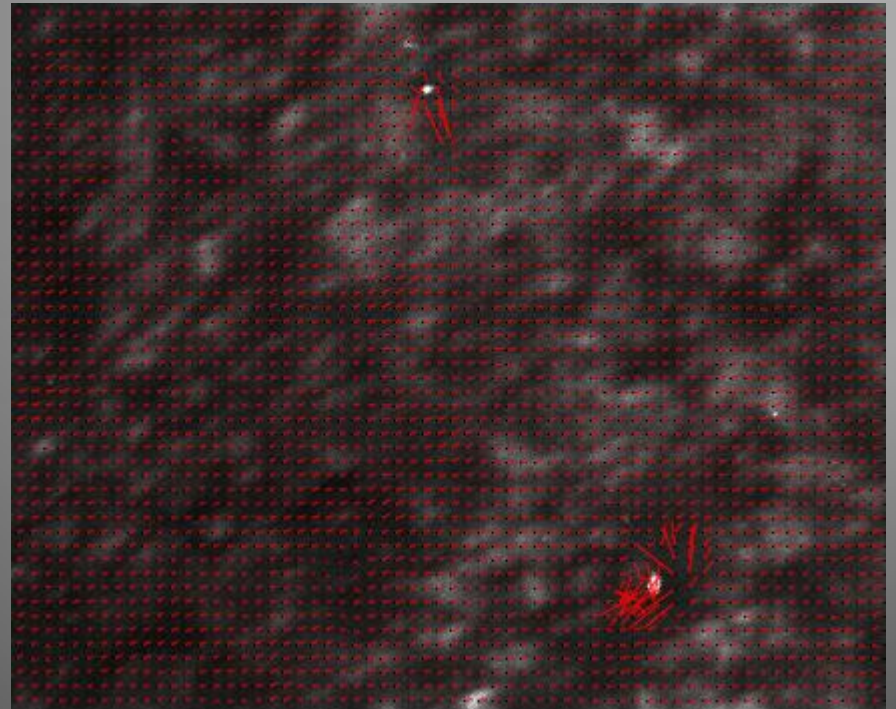
Spatial (still) inpainting



Temporal (film)



# Gradient Based Motion Estimation

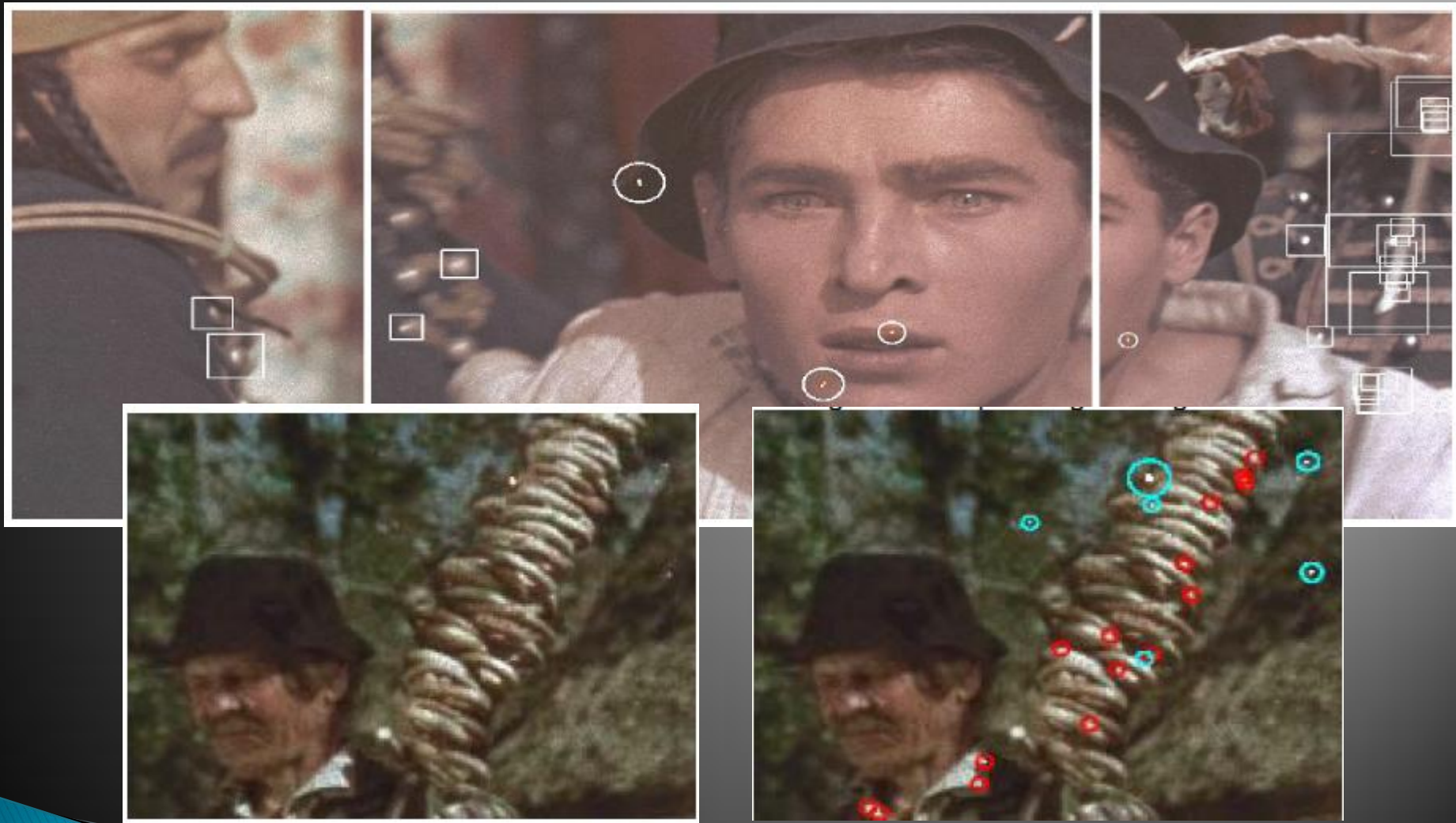


## **Vicious Circle phenomenon:**

- Artifact detection and removal needs reliable motion information
- Gathering reliable motion information is ill posed due to existing artifacts

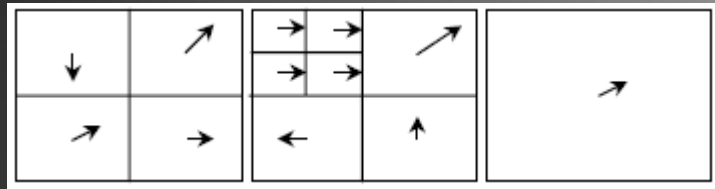


# Blotch decetion and removal



A. Licsár, T. Szirányi, L. Czúni: Trainable blotch detection on high resolution archive films minimizing the human interaction, Machine Vision and Applications, Springer-Verlag, Volume 21, Number 5, 767-777, 2010

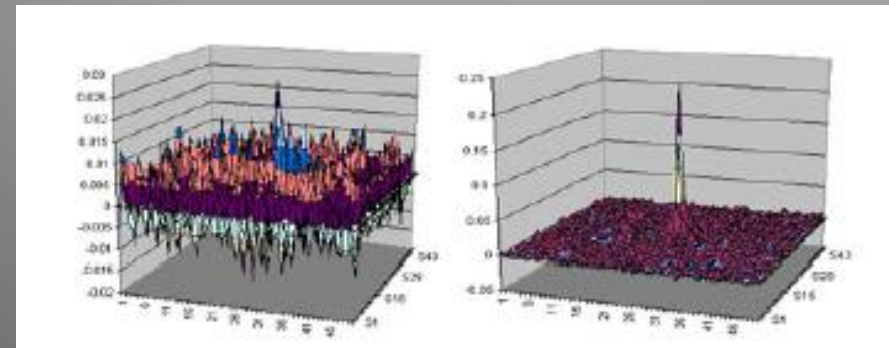
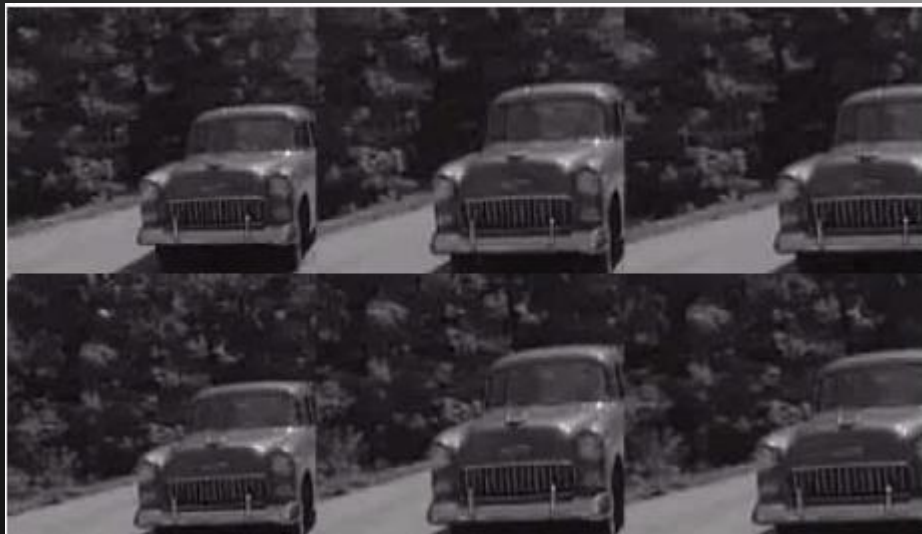
# “Offline” and automatic stabilization



Local motion structure: finding the fixation region.

$$\frac{F_1(\zeta, \eta) * F_2^*(\zeta, \eta)}{|F_1(\zeta, \eta) * F_2^*(\zeta, \eta)|} = e^{j2\pi(\zeta x_0 + \eta y_0)}$$

Cross Power Spectrum (CPS)



Inverse Fourier transform of the CPS  
for non-homogenous and  
homogenous moving areas



# “Offline” and automatic stabilization



Input

Automatic stabilization



Input

Local method

Global method

# Restoration of the first Hungarian colour movie film

- First full colour film produced in Hungary, 1949
- Cultural symbol after WWII
- Over 5 million cinema visitors over the years

## Participants in the restoration project:

- Hungarian National Film Archives
- Hungarian Filmlaboratory
- University of Veszprém (Pannonia)
- MTA Sztaki



# Areas of Historical Document Imaging and Processing

## **Image Acquisition:**

Imaging for fragile materials;  
Multispectral imaging;  
Camera-based/non-invasive acquisition

## **Document Restoration/Improving readability:**

Removing or minimizing damages, defects, ink-bleed;  
Completing and filling in missing pieces based on context, prior knowledge, supporting documents, i.e. inpainting; Machine-learning algorithms for enhancement based on example images

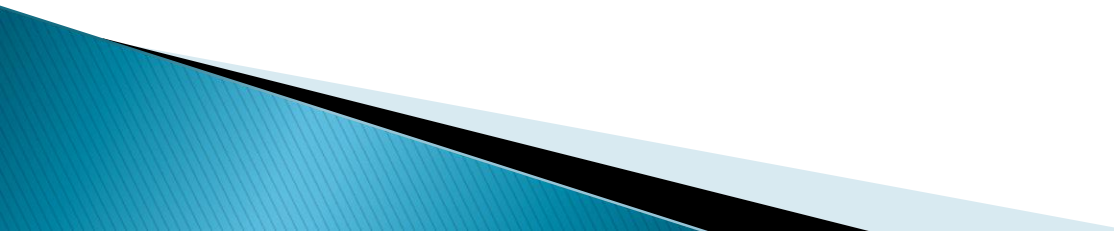
**Digital Archiving:** Compression issues; Measuring essential resolution (color, spatial) and metadata; Modeling of document image degradation;

**Content Extraction:** Content-based retrieval; Automated or semi-automated transcription; Content recognition based on surrounding and supporting context; Ontologies for modeling historical document content; Extracting and linking names, dates, places, personal and family histories and narratives; Discovering historical social networks

## **Automated Classification, Grouping and Hyperlinking**

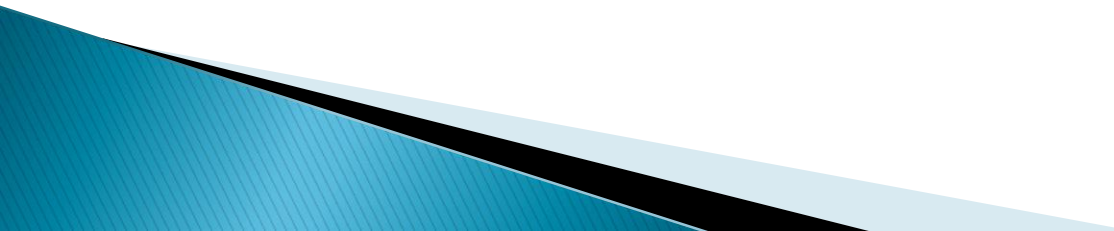
Style identification (typography of printed text, handwriting style recognition for manuscript authentication or author identification...); Searching for Documents over the Internet; Searching/querying, retrieval, summarizing/condensing of document images; Parallel tagging of images, transcripts, and other document layers

# Special Areas of Historical Documents

- ▶ Military records, personal journals, church records, medieval manuscripts
  - ▶ Historical Collections
  - ▶ Scientific, technical and educational documents
  - ▶ Government archives, documents from the world cultural heritage, multi-language
  - ▶ Multimodal information (motion picture, audio)
  - ▶ Family history documents and genealogies
- 



# Outline of handwriting recognition

- ▶ OCR (Optical Character Recognition)
  - ▶ Handwriting recognition
  - ▶ Document segmentation
  - ▶ Signature recognition
  - ▶ Handwriting recognition in archive documents
    - Introduction of the problem
    - Recognition by SIFT points
    - Pivot based search for faster recognition
- 

# Types of Text

## Character coded texts

Word  
editor files  
...

Web  
pages

## Embedded texts

### Physical media

Books

Films

Letters

...

**Digitization**

### Electronic documents

Still  
images  
...

Video

**OCR**



# Character (word) recognition

- ▶ OCR (Optical Character Recognition)
  - Widespread applications (books, journal papers, etc.)
  - Problems only in noisy/distorted/undersampled environments
- ▶ Handwritten text recognition
  - **Online** recognition (mobile devices, touchpads, bank signature verification systems), dynamic: uses pen's speed, position, pressure, acceleration, etc.
  - **Offline** recognition: uses only static images
- ▶ Signature recognition: learn personal characteristics of handwriting (signature verification or writer identification) -> also for historical documents

# History of Handwriting Recognition

- ▶ 1914 Hyman Eli Goldberg , U.S. Patent 1,117,184, On-line recognition of hand-written **numerals** to control a machine in real-time. Controller: conversion of handwritten numbers to electronic data by inductive ink to controll equipments.



- ▶ 1938 George Hansel, U.S. Patent 2,143,875, machine recognition of handwriting
- ▶ 1957 Tom L. Dimond: Stylator the first on-line handwriting recognizer prototype

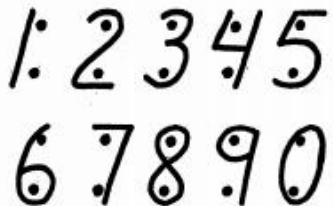


Fig. 4—Numerals with dot constraint.

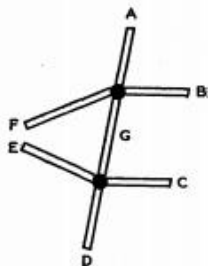


Fig. 5—Set of bipolar coordinates for character recognition.

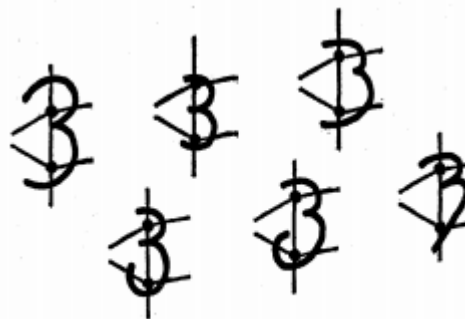


Fig. 6—Range of variation permissible.

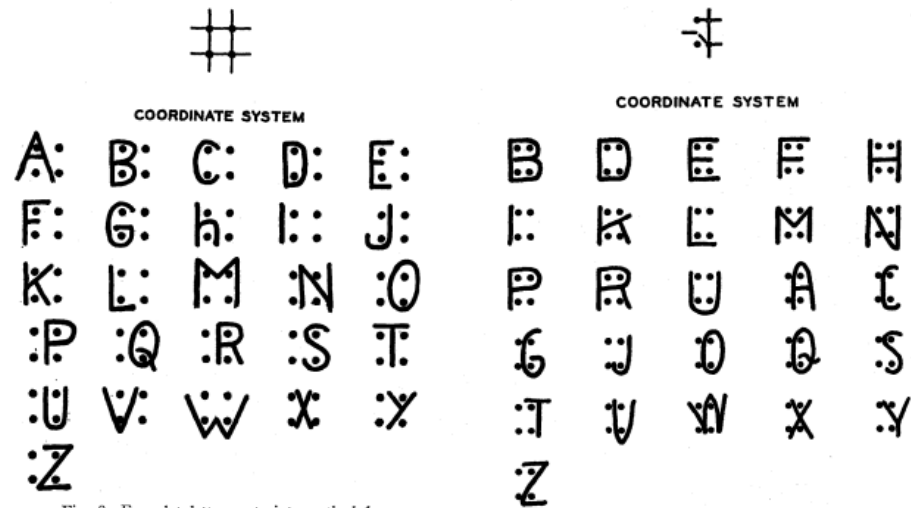
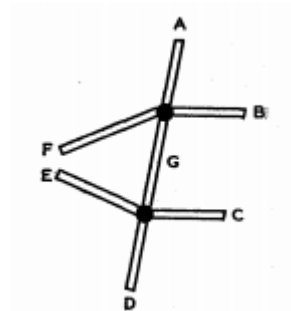


Fig. 8—Four-dot letter restraint—method 1.

Fig. 9—Four-dot letter constraint—method 2.



► T. L. DIMOND: Devices for Reading Handwritten Characters



ALLOWED CONFIGURATIONS	CRITERIAL AREA						
	A	B	C	D	E	F	G
1	0	0	0/1	0	0/1	1	0
2	0	1	0/1	0	0/1	0	0
3	1	1	0	1	0/1	0	1
4	1	1	1	1	0/1	0	0/1
5	0	0/1	0/1	0	0/1	1	1
6	1	0	0/1	0/1	0	1	0/1
7	0/1	0	0/1	1	1	1	0/1
8	1	1	0/1	0	0/1	0	0/1
9	0/1	1	1	1	0/1	1	1
0	1	1	0/1	0	0/1	1	0/1
0	0/1	1	1	1	0/1	1	0

Fig. 7—Truth table for numerals.

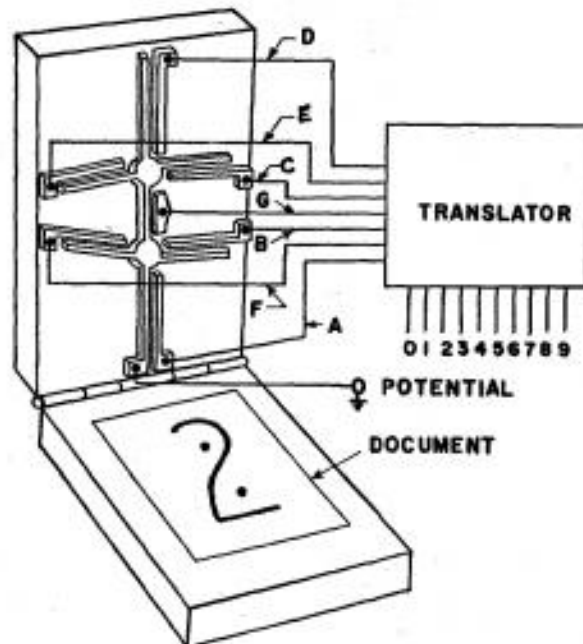


Fig. 10—Reader.

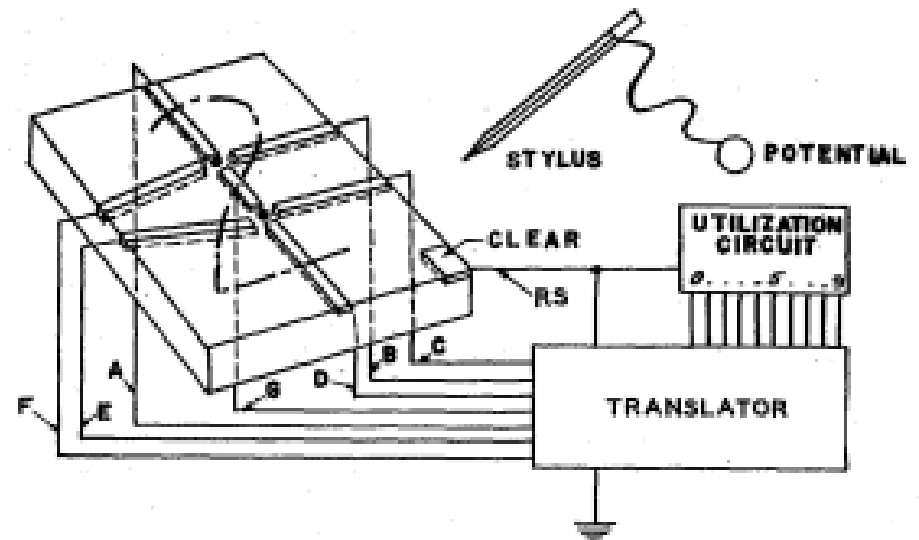
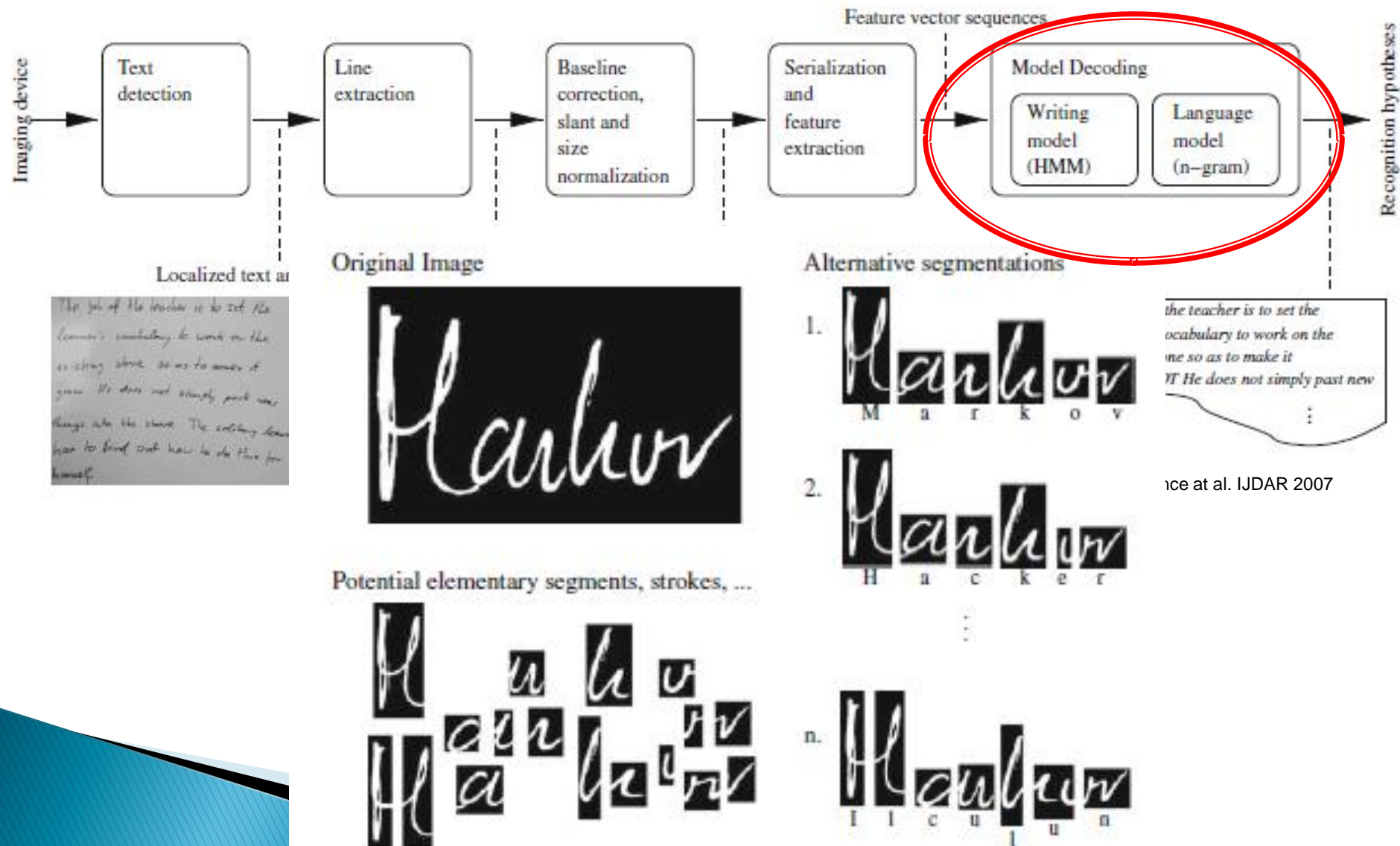


Fig. 12—Stylator.

# Overview of offline HWR

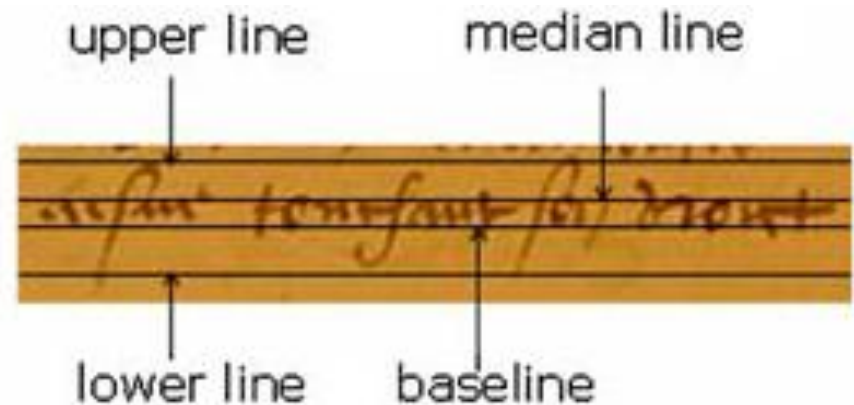




# Document Sementation for Text Recognition

Aim: to find the correspondence between document image and its content by text/image alignment techniques

- ▶ *Baseline: fictitious line which follows and joins the lower part of the character bodies in a text line (Fig. 1).*
- ▶ *Median line: fictitious line which follows and joins the upper part of the character bodies in a text line.*
- ▶ *Upper line: fictitious line which joins the top of ascenders.*
- ▶ *Lower line: fictitious line which joins the bottom of descenders.*
- ▶ *Overlapping components: overlapping components are descenders and ascender located in the region of an adjacent line*



# Problems in Document Segmentation for Text Recognition

## Line level:

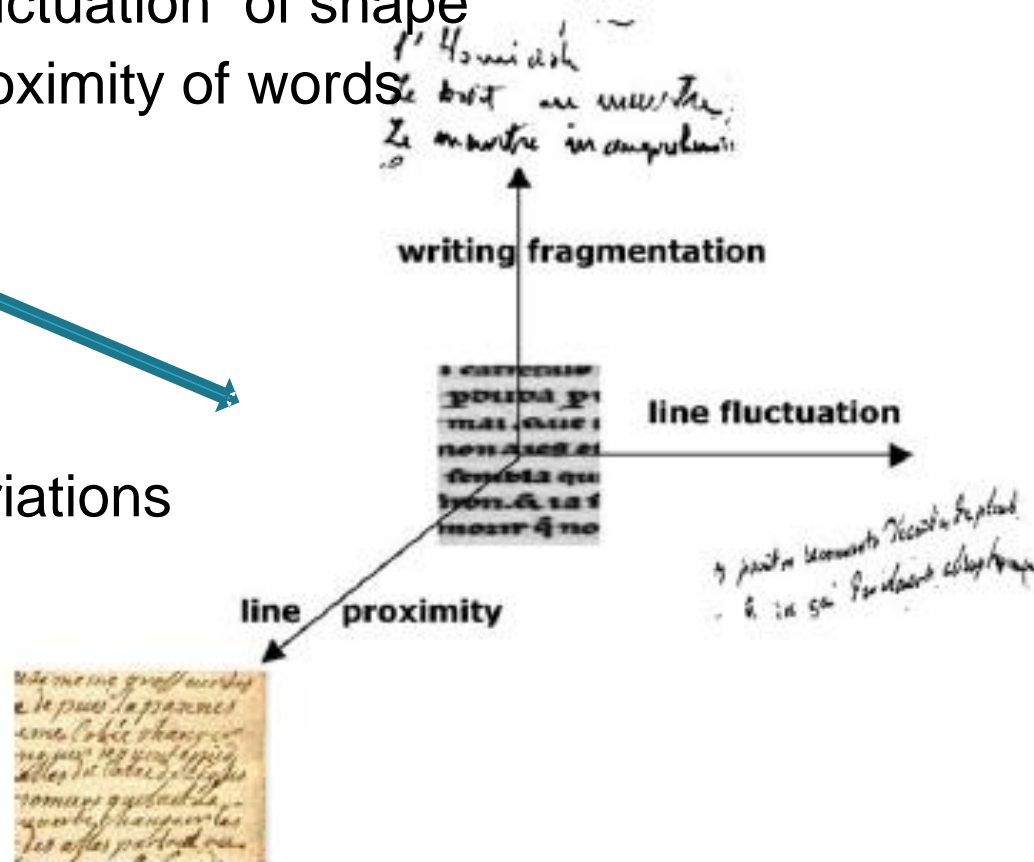
- ▶ Fragmentation
- ▶ Fluctuation
- ▶ Proximity

## Word level:

- ▶ Fragmentation of letters and words
- ▶ Fluctuation of shape
- ▶ Proximity of words

## Sources of noise:

- ▶ Blotches/dirt
- ▶ Background intensity variations
- ▶ Transparency of paper
- ▶ Tears
- ▶ Scanning problems

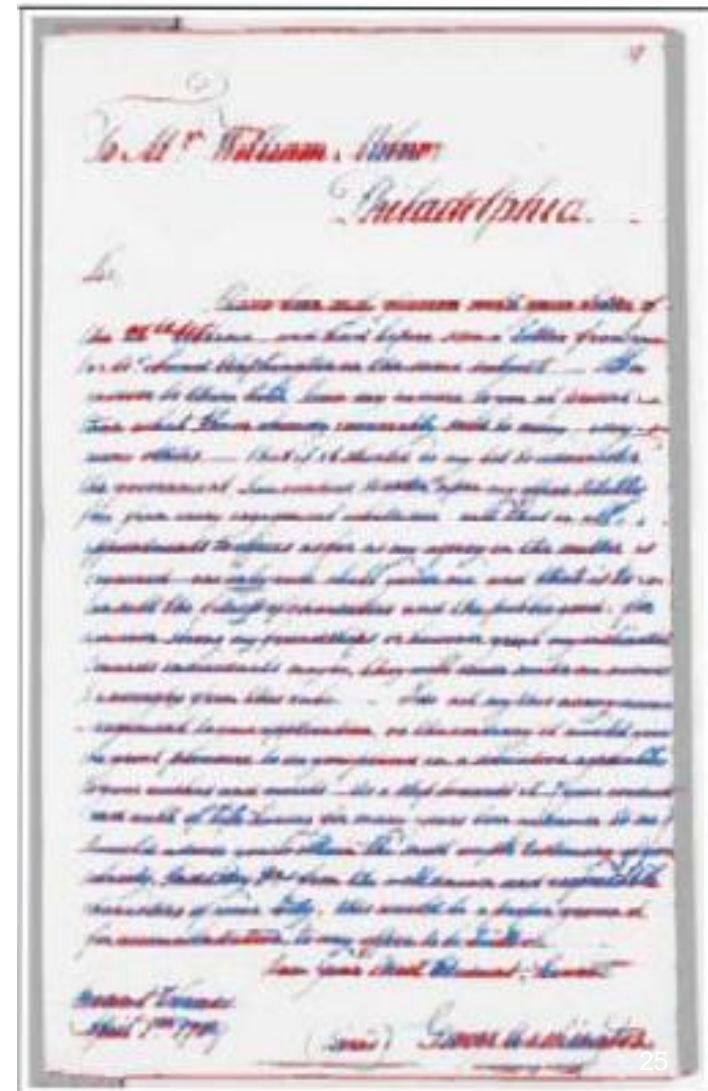




# Document Sementation for Text Recognition

- ▶ Projection-based methods
- ▶ Grouping methods: aggregating units in a bottom up strategy
- ▶ Smearing methods (horizontal smearing then bounding box detection)
- ▶ Hough transform based methods
- ▶ ...etc.

laissons cela. Ramie disait: "Hé! hé! anté, d'apuyez, he! - Ce  
profane ou sa folie, c'est quelle à'a feoté", du beau jeu.  
- les séductions de "Pélie", : jamais je ne meurs au P. lea.  
- propriétaire d'un "talant": la seule affaire était de me ven-  
- dre dans le mains, à la dans les poche - par le travail et les-  
- tance. Donc, la pure option de s'élever endessus de person-  
- ne, acquiescent, sans ostentation pour sur eux tout entier à l'œuvre  
- me saur tout entier. Si possible, le possible Salut au sage  
- des accessoires, garest-t-il? Tout un homme, fait de tous les  
- et qui les fait tous et qui fait à l'aise qu'il se



# Super Resolution (SR) Based Character (NP) Recognition

**Problem:** low resolution number plates in security videos



Optical zoom

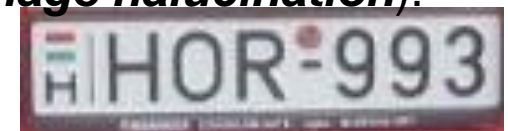


Optical zoom

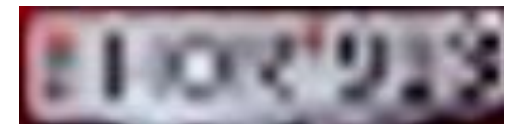


**Solution:** apply statistical image processing with the knowledge of what we expect to see (“**example based**” **super resolution**, **image hallucination**).

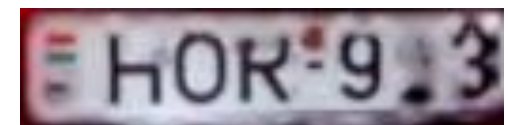
1. Learn low resolution – high resolution patch pairs by image examples
2. Retrieve high resolution patches from low resolution observation applying local constraints
3. Recognition: use reconstruction code statistics



Original known NP



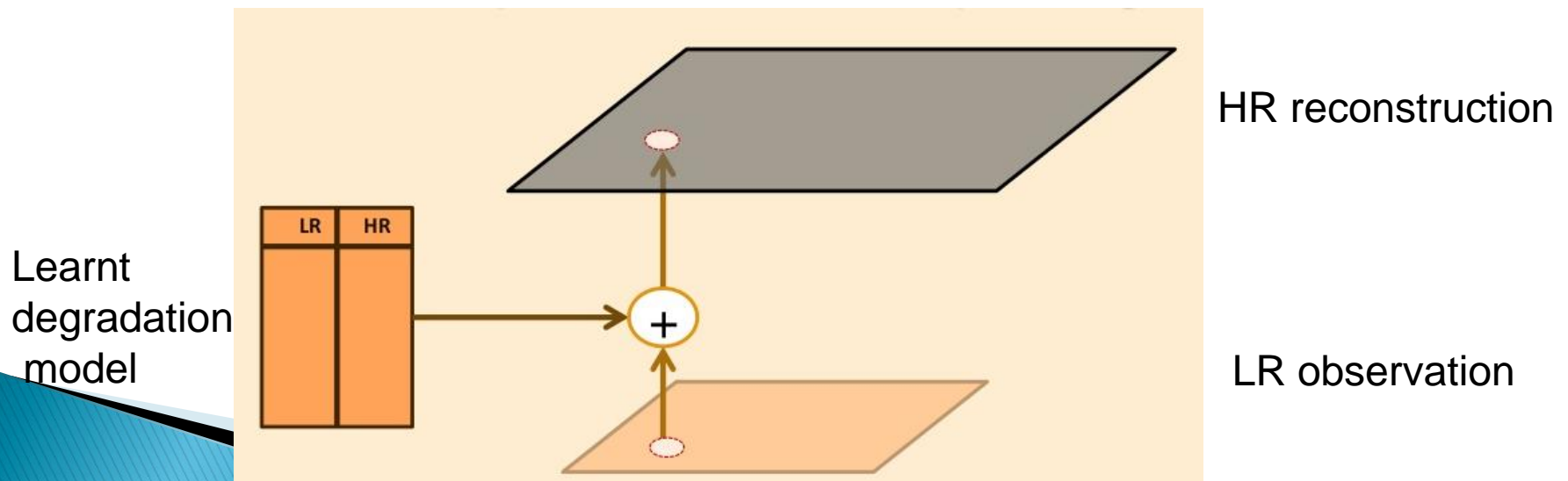
Low resolution observation



Reconstructed NP

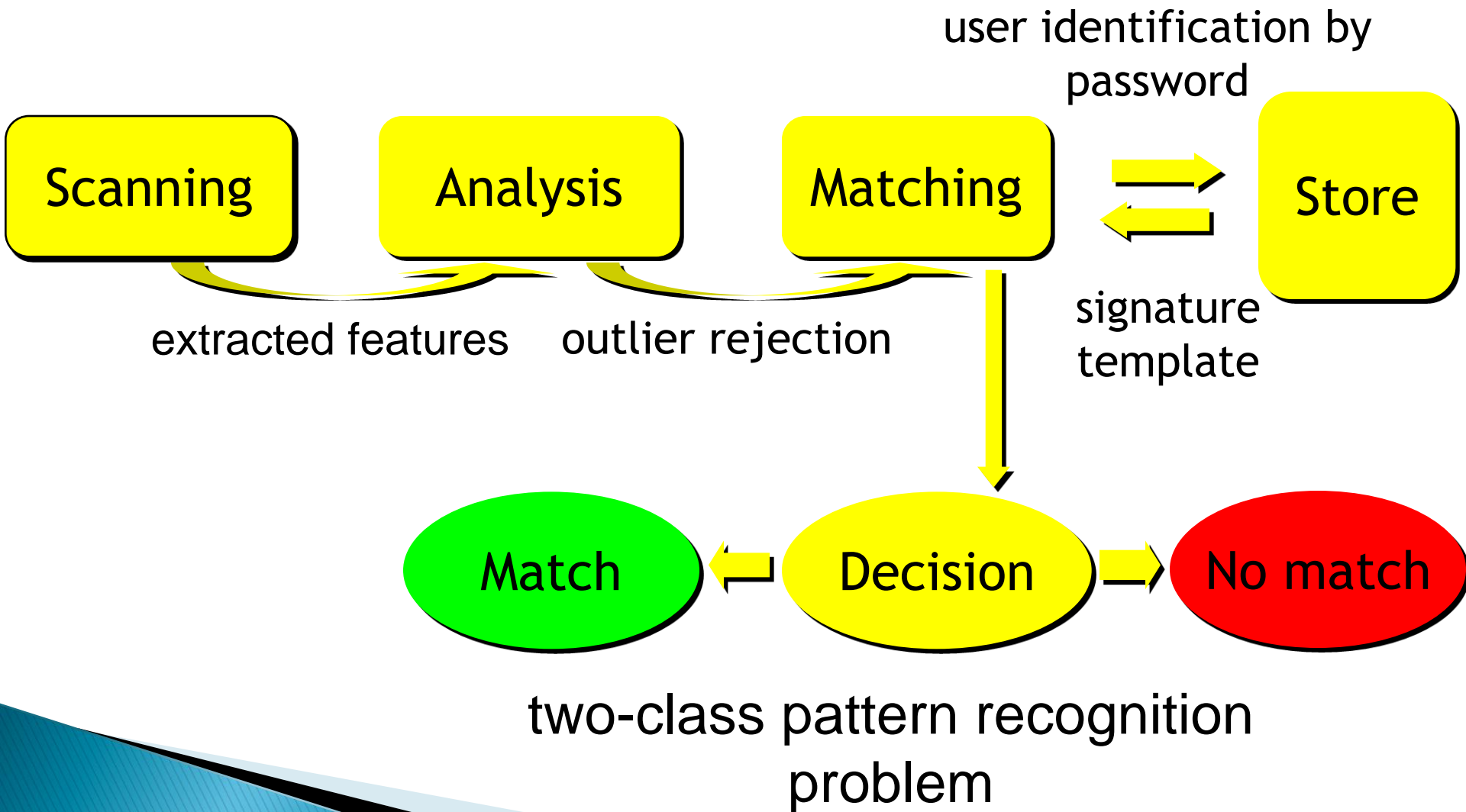
# Example-based SR

- ▶ Learn LR-HR image patch pairs by example images
- ▶ Build up a database from LR-HR pairs
- ▶ Replace LR patches with corresponding HR patterns also considering ***neighborhood fitting***





# Signature-based biometrics

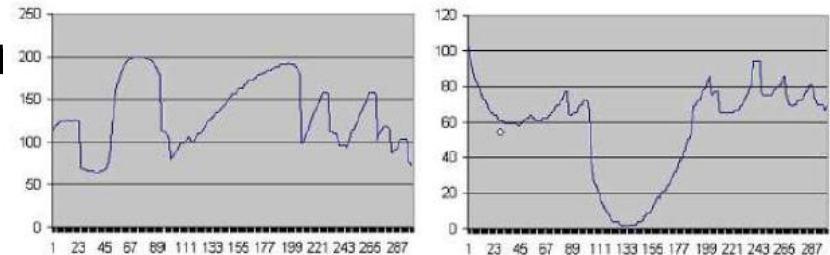


# Signature recognition

- ▶ Alignement
- ▶ Feature extraction
  - Baseline Slant Angle
  - Aspect Ratio
  - Normalized area of the signature
  - Center of Gravity
  - Slope
  - Upper profile/lower profile
  - Etc.

A sample off-line signature in black ink, appearing to read 'Seraphin'.

Figure 3.3: Sample off-line signature.

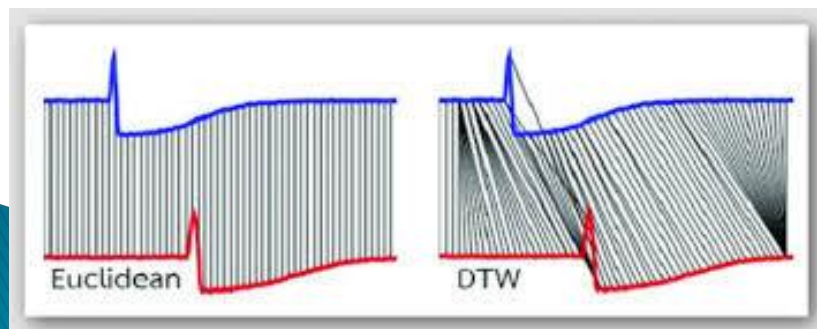
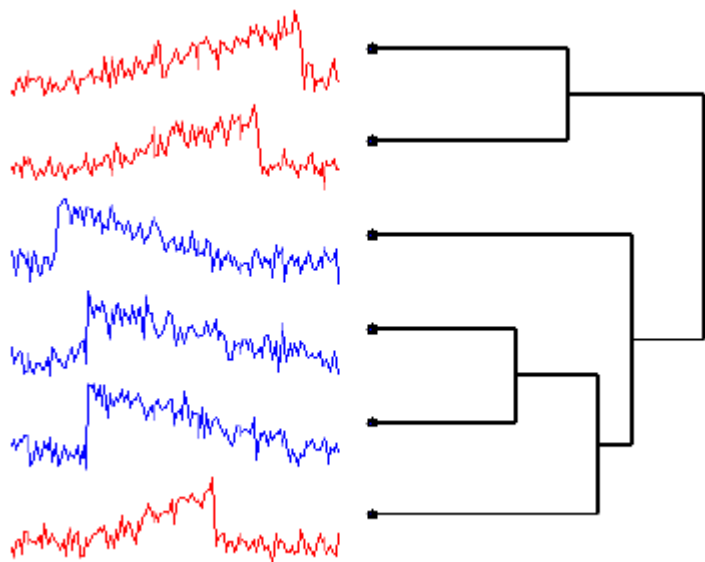


Upper profile/lower profile

- ▶ Comparison
  - Several types of metrics... (do not work alone) but
  - **Dynamic Time Warping**
  - Hidden Markov Models can help...

# Dynamic Time Warping...

- ▶ To find local correspondence...



- Horizontal non-linear stretching of objects to find the best matching
- Local gradient algorithms work well



# The amount of information in archive documents...

Consumed by an average person on an average day

- ▶ corresponds to 100,500 words
- ▶ and 34 gigabytes
- ▶ newspapers, books, portable computer games, satellite radio, and Internet video,
- ▶ (information at work is not included!)

How Much Information? 2009 Report on American Consumers, University of California, San Diego)

Estimated number of books:

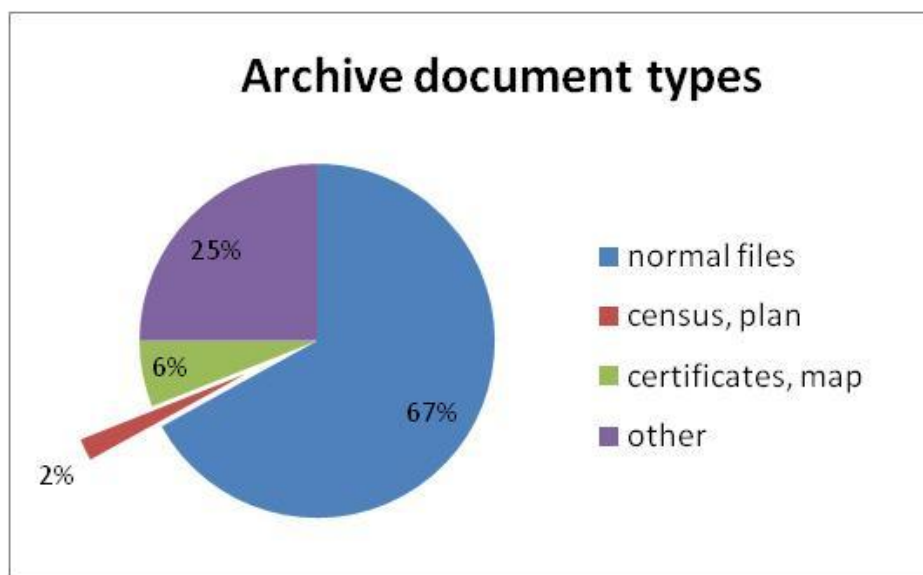
- ▶ 129,864,880.
- ▶ „at least until Sunday”

(Google Books research, 2010)

- ▶ What about old documents...?

# What about archive paper documents?

- ▶ The number of archive pages (only in Hungary): 3 500 000 000 – over 3 billion!
- ▶ The number of archive pages recommended for digitization: 200 000 000 (5.7%)



Orosz Katalin, Rácz György, Reisz T. Csaba, Vajk Ádám, Véber János,  
Középkori oklevelek tömeges digitalizálása, Magyar Országos Levéltár,  
(2008)

# Aims of Digitization

- ▶ To preserve information for future generations
  - ▶ To make them analyzable for researchers
  - ▶ To make them searchable for the public
- 
- ▶ Central European Virtual Archives Network of Medieval Charters Project: ... Digitization of medieval charters within the stocks of the participating archives...





# Handwriting styles

**J**ohann Neudörffer the Elder's  
~ 1538 writing manual ~  
fascinated the German designer  
Hellmut Gomm for years.

Fraktur handwriting

*Business Writing was developed from  
Spencerian Script as a simpler, monoline  
version intended for everyday use. I am a  
recent convert, and find this style of  
writing very attractive in its own right,  
and a real joy to write. In my opinion,  
this very beautiful script takes its place  
alongside italic as an ideal basis for a per-  
sonal style of handwriting and is perfect  
for those who prefer to write monoline.*

Cursive handwriting

*If doctors are so smart  
why is their handwriting  
so messy?*

„Normal” cursive handwriting



# Traditional OCR software products?

- ▶ FreeOCR,
- ▶ TOCR viewer,
- ▶ SimpleOCR,
- ▶ Abby FineReader,
- ▶ TOPOCR,
- ▶ ...

simply do not work...archivists process information manually...



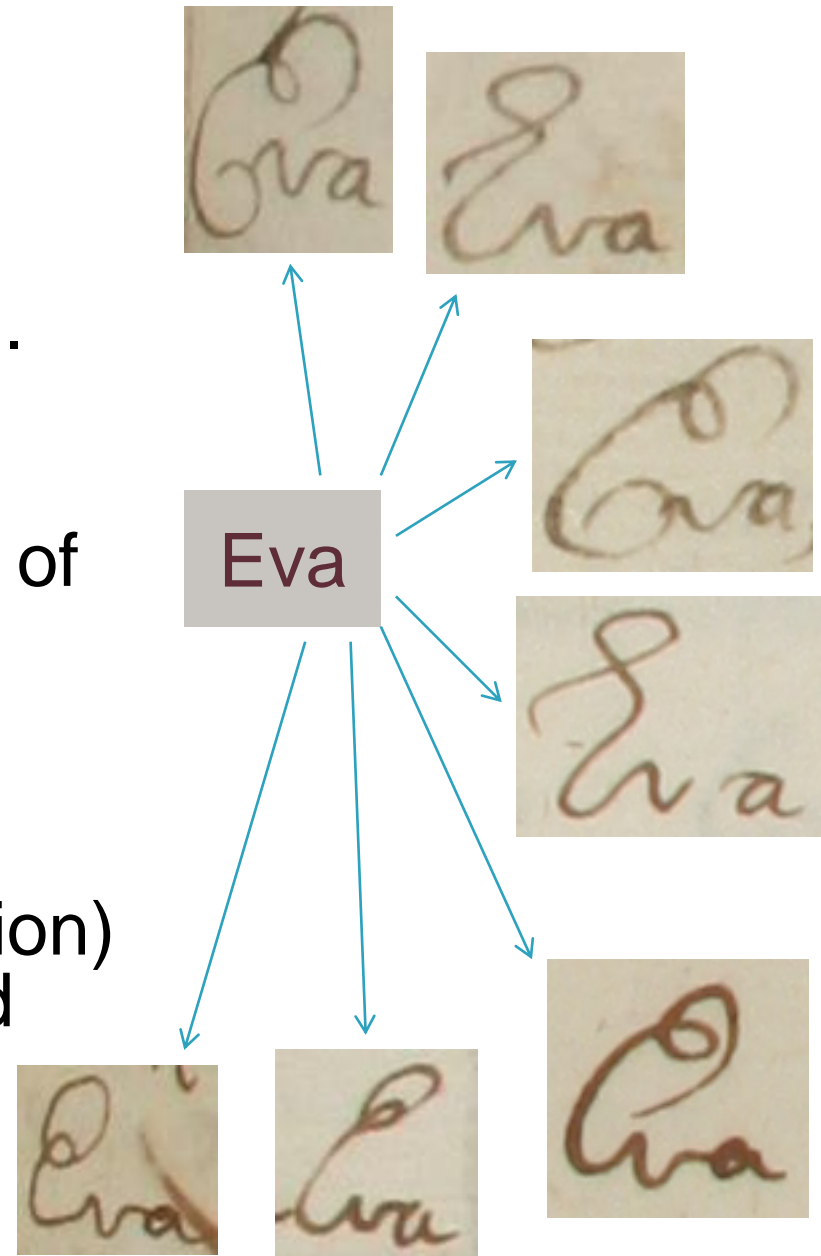


# Typical Problems of Archive Cursive Handwriting

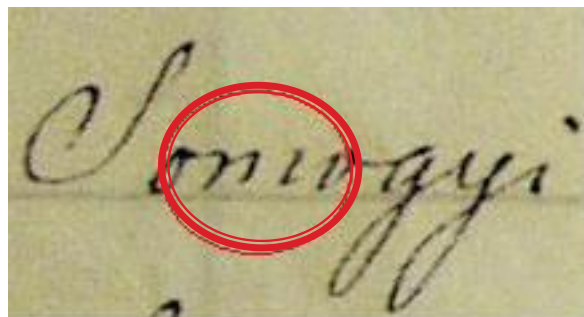
- ▶ The same letters have different appearances (e.g. „E” in Eva)
- ▶ The beginning and ending of letters can not be easily recognized



separation (segmentation)  
of letters is a (too) hard  
problem

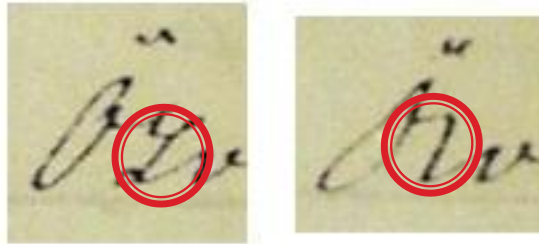


# Typical Written Problems



Broken line transforms "m" into "n"  
and "r"

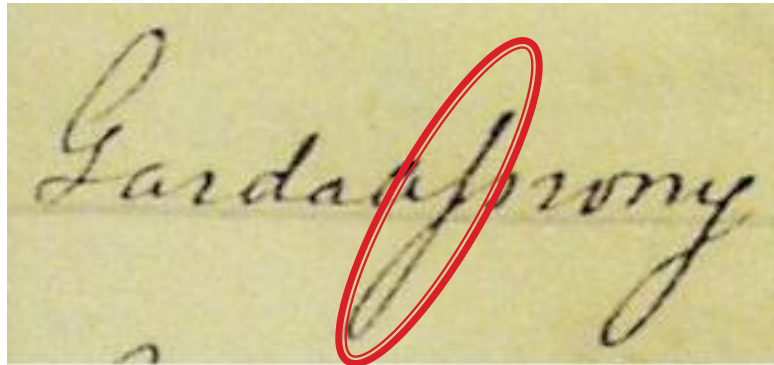
# Typical Written Problems



Different appearances of the same letter "z" in the same hand-writing (beginning of "Özvegy")



# Typical Written Problems



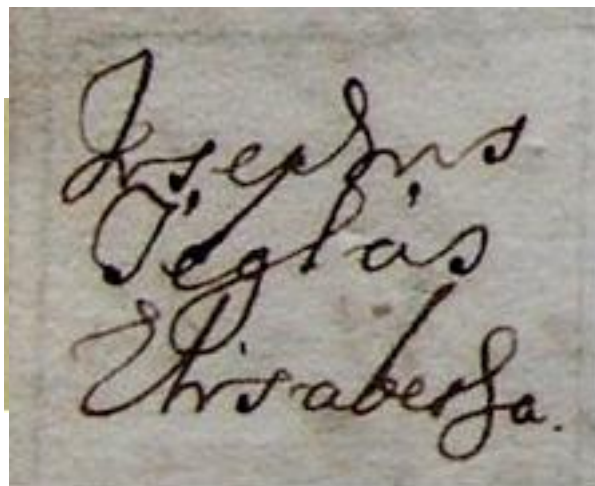
Misspelling of the 7th letter which should look like the 8th letter.

# Typical Written Problems



Similarities of different letters in the same hand-writing (beginning of "István", "János", "Sámuel")

# Typical Written Problems



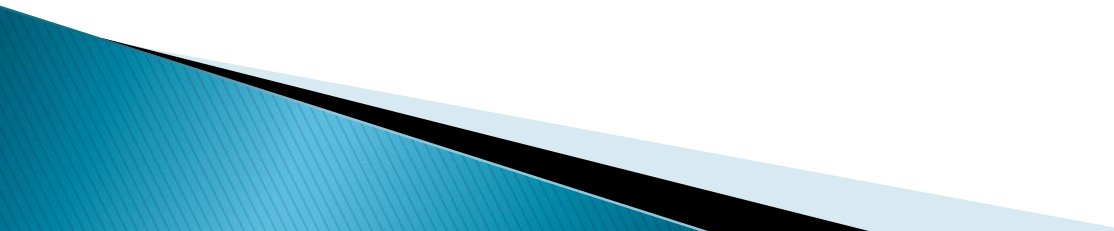
Word overlapping.



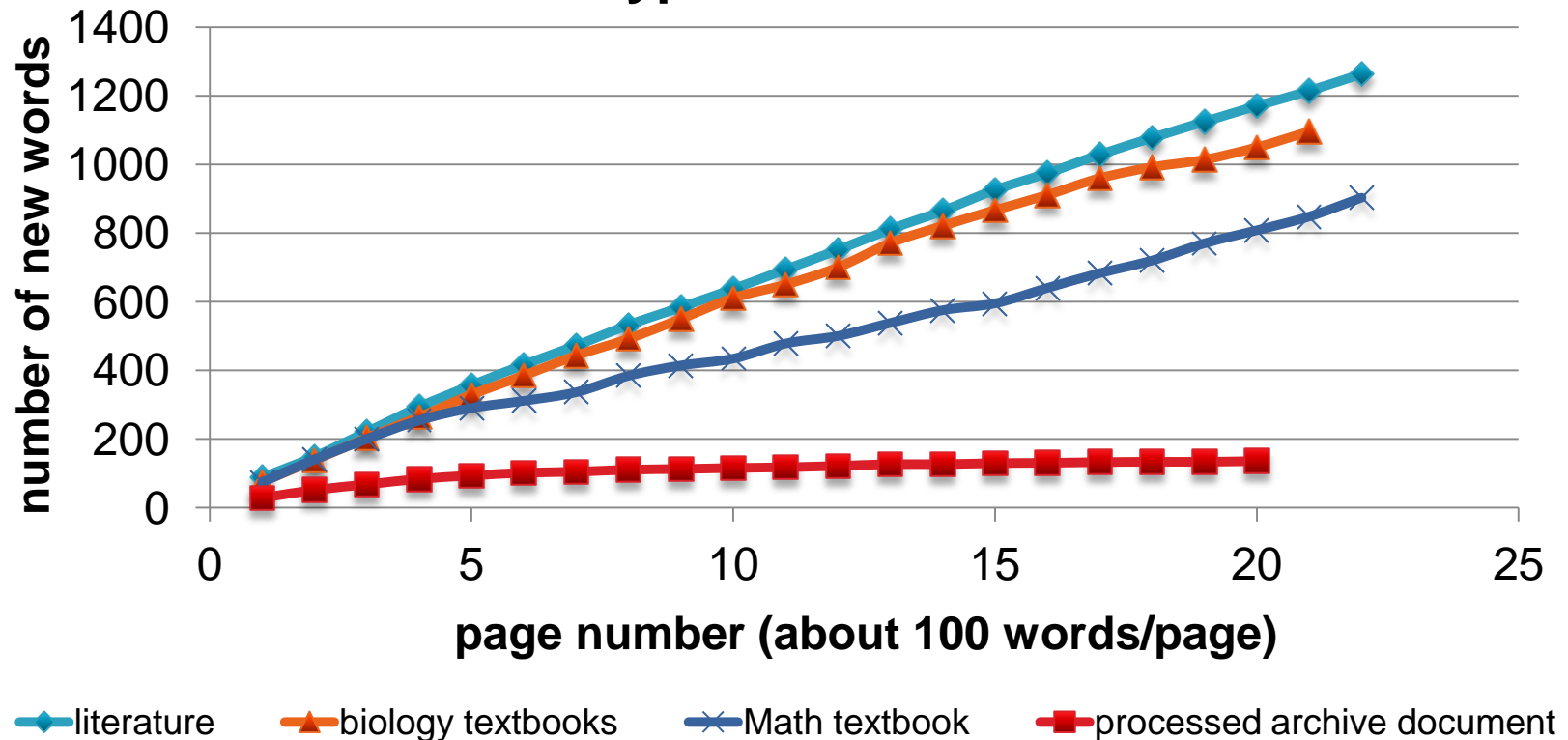
Imerus Domorum  
 Item  
 ma & Cognomina  
 ditum & totius Sa,  
 mihic.

		Das	Confessi	Confirma
annet	Fülei — "	54	—	—
vor	Anna Pereri — "	45 <sup>e</sup>	—	—
	Maxim — "	10	—	—
	Cassarina — "	14	—	—
	Isanna — "	13	—	—
	Anna — "	2	—	—
e	Werner — "	33	Conf	—
vor	Elisabetta Fottb — "	25 <sup>e</sup>	Conf	Conf
anne	Pinter — "	44	Conf	Conf

# Consequences

- ▶ Character-based recognition in several cases does not work.
  - ▶ Is it worth trying word-based recognition – word spotting?
  - ▶ What is the amount of word classes?
- 

## Cumulative Distribution of New Words in Different Types of Documents



➡ A continuously „learning system” seems to be reasonable, the amount of necessary annotation decreases exponentially from page to page in the archive document to be processed.



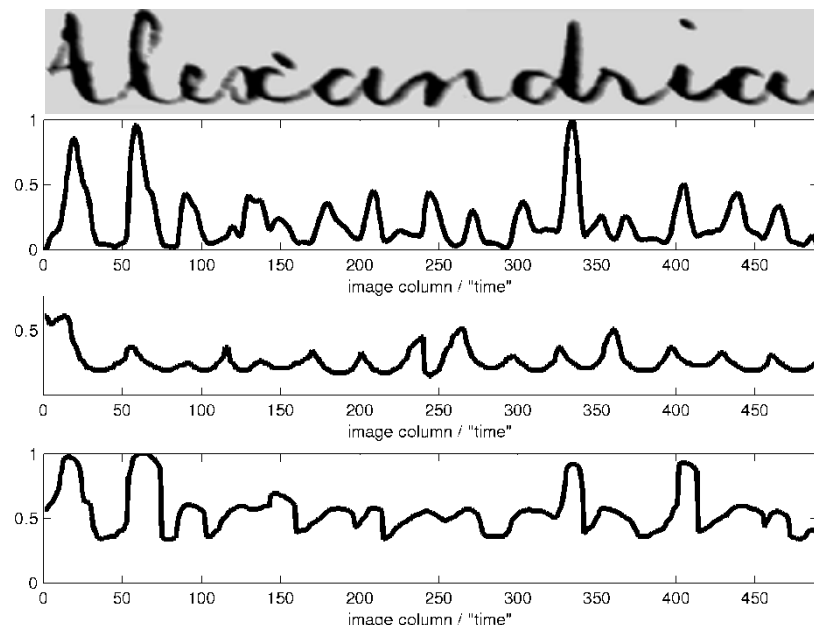
# Global word shape based classification

## ▶ Tested descriptors of length 329

- horizontal and vertical size and their ratio;
- minimum, maximum, and average intensity;
- average intensity derivatives;
- upper profile; lower profile;
- right profile; left profile;
- center of gravity;
- black-white transitions; black-white ratio;
- black count;
- black density;
- image moments

## ▶ Tested classifiers: k-NN, Random Tree, Random Forest, Naive Bayes ect.

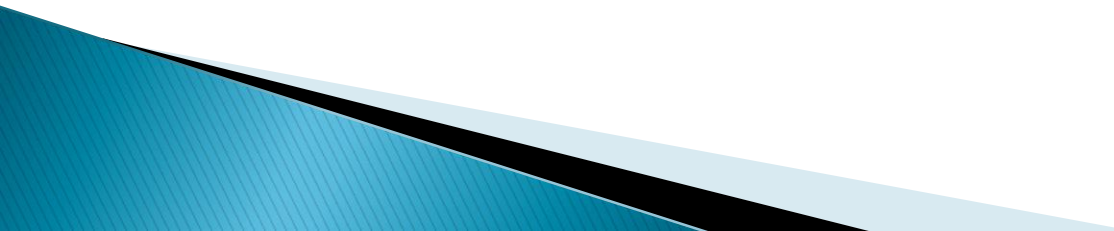
## ▶ Average performance is around (only) 50% recognition rate



# Global word shape based classification

- ▶ **Global** word feature descriptors are
  - Sensitive to the individual (inter class) variations of word shape
  - Sensitive to extreme decorations
  - Sensitive to dirt and noise
  - are „ad-hoc”
- ▶ What about **local** feature descriptors in word spotting?
  - SIFT, SURF, FAST, ... successfully applied to complex images
  - Invariant to transformations (rotation, scaling)

# Local features for word spotting

- ▶ *Has it been already applied?*
  - ▶ *Is scale invariance of descriptors important to be considered?*
  - ▶ *Is rotation invariance of descriptors important to be considered?*
  - ▶ *Is word structure (f.e. skeleton) itself proper to extract local features?*
- 



# Existing solutions

- Lawrence Spitz: Using Character Shape Code for Word Spotting in Document Images (1995)
  - „SIFT-like” descriptor
  - Applied to Chinese symbols
  - *Not scale and rotation invariant*
- J. A. Rodríguez, F. Perronnin: Local Gradient Histogram Features for Word Spotting in Unconstrained Handwritten Documents. *Frontiers in Handwriting Recognition* (2008)
  - Gradient histogram descriptor in a moving window
  - DTW or HMM for classification
  - 80% hit rate for a low number of classes
  - *No information selection*
- Uchida, S.; Liwicki, M., Part-Based Recognition of Handwritten Characters, *Frontiers in Handwriting Recognition (ICFHR)*, 2010 International Conference on, 545–550 (2010)
  - Tested and applied only for the 10 digits
  - SURF points without positions (*not real localization*)
  - Feature point votes for character class

More comprehensive overview is available in Czúni et al., CBMI2013

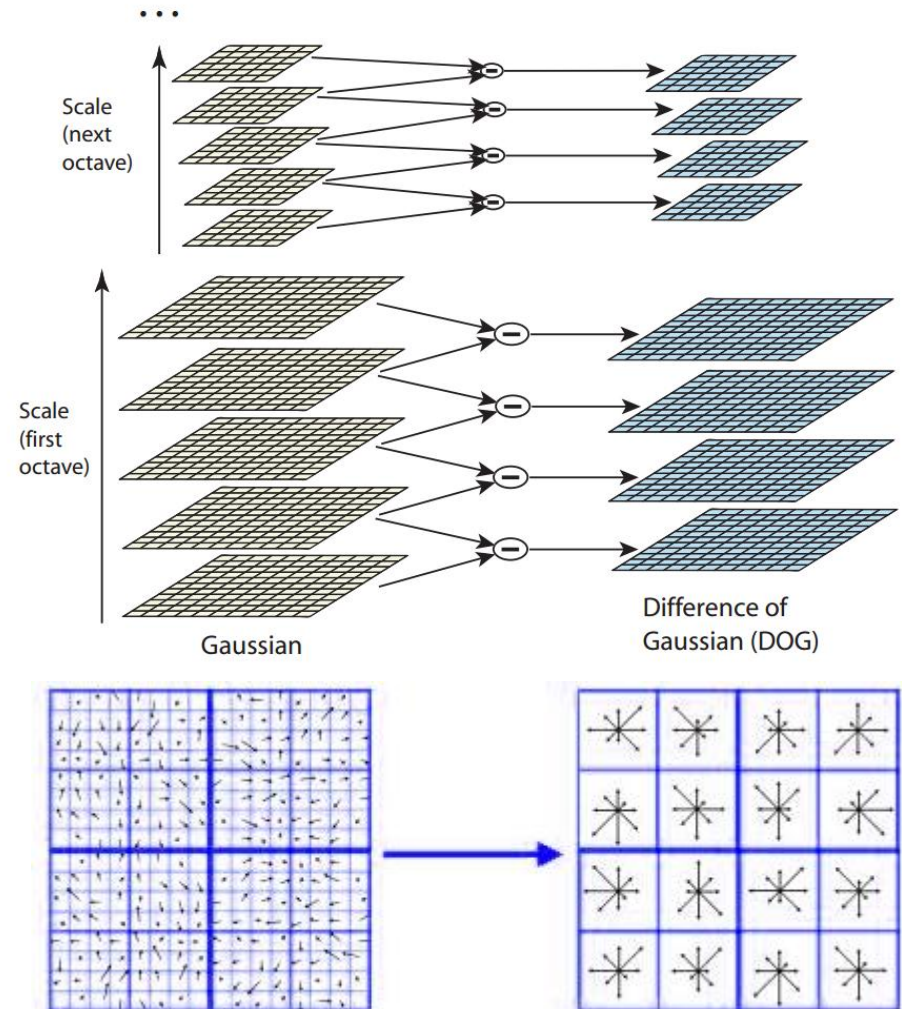
# SIFT local descriptor

## ► Scale Invariant Feature Transform

- Difference of Gaussian pyramid
- Finding local extreme points (position, scale)
- Leaving out low contrast and edge points
- Finding the maximal gradient (for orientation invariance)
- Setting the local coordinate system
- Generating the descriptor vector

## ► Properties

- Invariant to affine transformations (scaling, rotation, etc.)
- Computationally expensive



1. Localize SIFT points and generate SIFT descriptors both in the query (q) and in the candidate words (c).
2. Normalize SIFT point positions by the physical size of the words.
3. Define a disk shape area around each feature point of the query (q): only candidate points (c) within this area will be compared.
4. Find the best two matching points

$$D(q_i, c_j) = \sqrt{\sum_{k=1}^{128} (q_i(k) - c_j(k))^2}$$

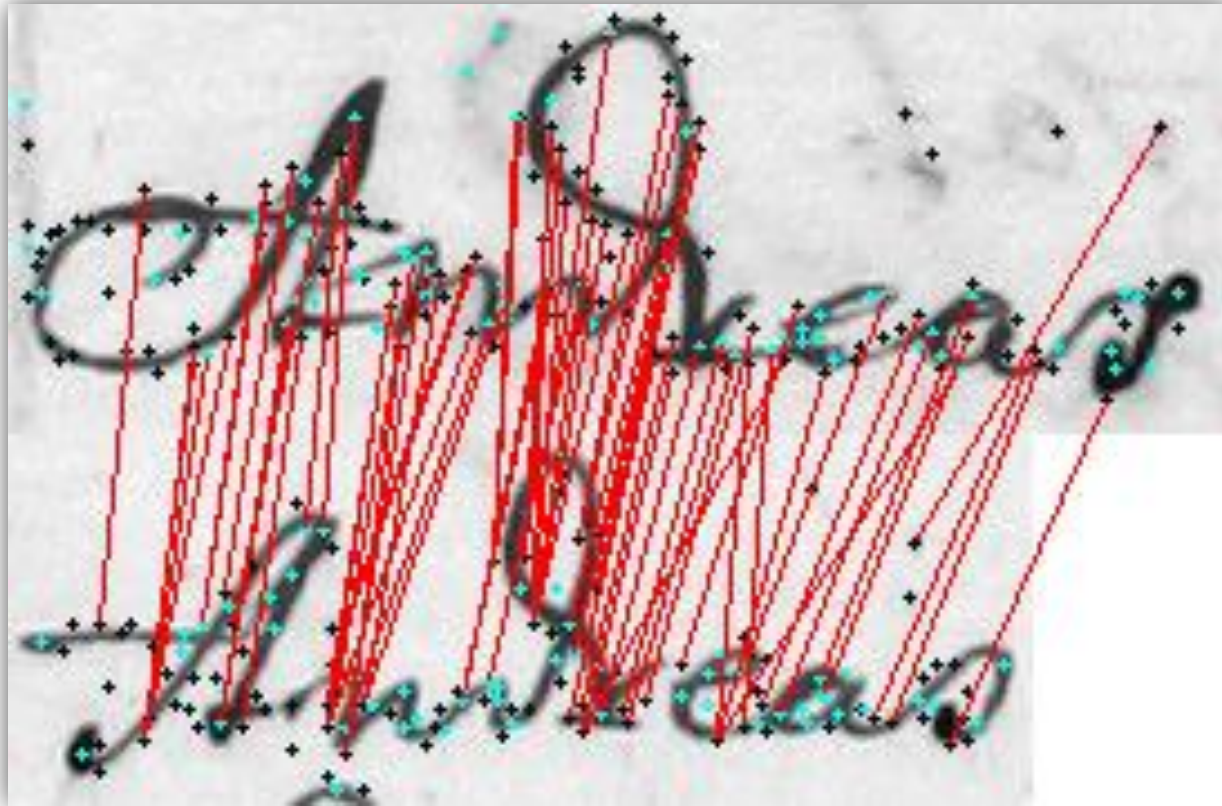
$$c_{i,min} = \min_{c_j} D(q_i, c_j)$$

$$c_{i,min2} = \min_{c_j} D(q_i, c_j) \text{ s.t. } c_j \neq c_{i,min}$$

5. Apply a threshold to orientation difference
6. Constrain the uniqueness of the best matching point  $\frac{D(q_i, c_{i,min})}{D(q_i, c_{i,min2})} < T_D$
7. Calculate the similarity value for the query and candidate words with the use of the matching points, rank candidates according to this similarity value:

$$S(Q, C) = \sum_{j=1}^N (\sqrt{255^2 \cdot 128} - D(q_j, c_j))_{(q_j, c_j) \in M_{Q,C}}$$

# Example for matching points

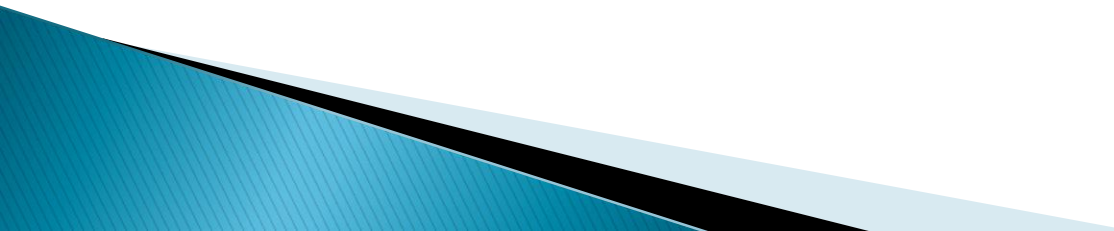




# Advantages

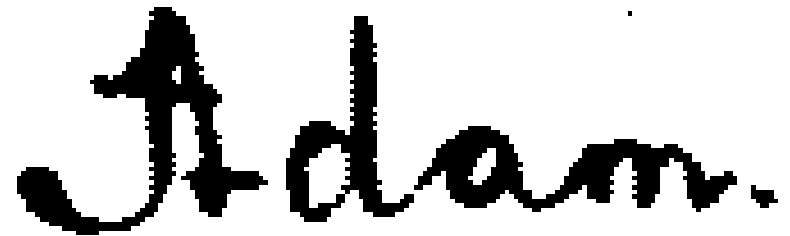
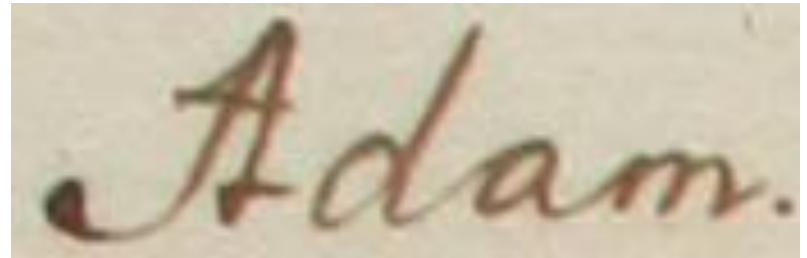
- ▶ Scale and rotation invariance (in some degree)
- ▶ No need for preprocessing (e.g. binarization, slant correction, noise removal, morphology, etc.)
- ▶ No need for precise segmentation of words.
- ▶ The searching area is symmetrical around query points, contrary to most methods using DTW, where matching cannot go backwards.
- ▶ Stable in noisy environments: the algorithm can neglect most noisy points.
- ▶ Only extrema points in scale–space are considered: there is no need to correlate points with small information content.

# Experimental setup

- ▶ 22 manually annotated pages of the 177 with 1638 word images.
  - ▶ 103 random query image compared to the remaining 1637 images
  - ▶ 111 word classes
  - ▶ most frequent word: 116 occurrence
  - ▶ 68 words with only 1 occurrence
  - ▶ SIFT (OpenSIFT, Lowe), SURF
- 

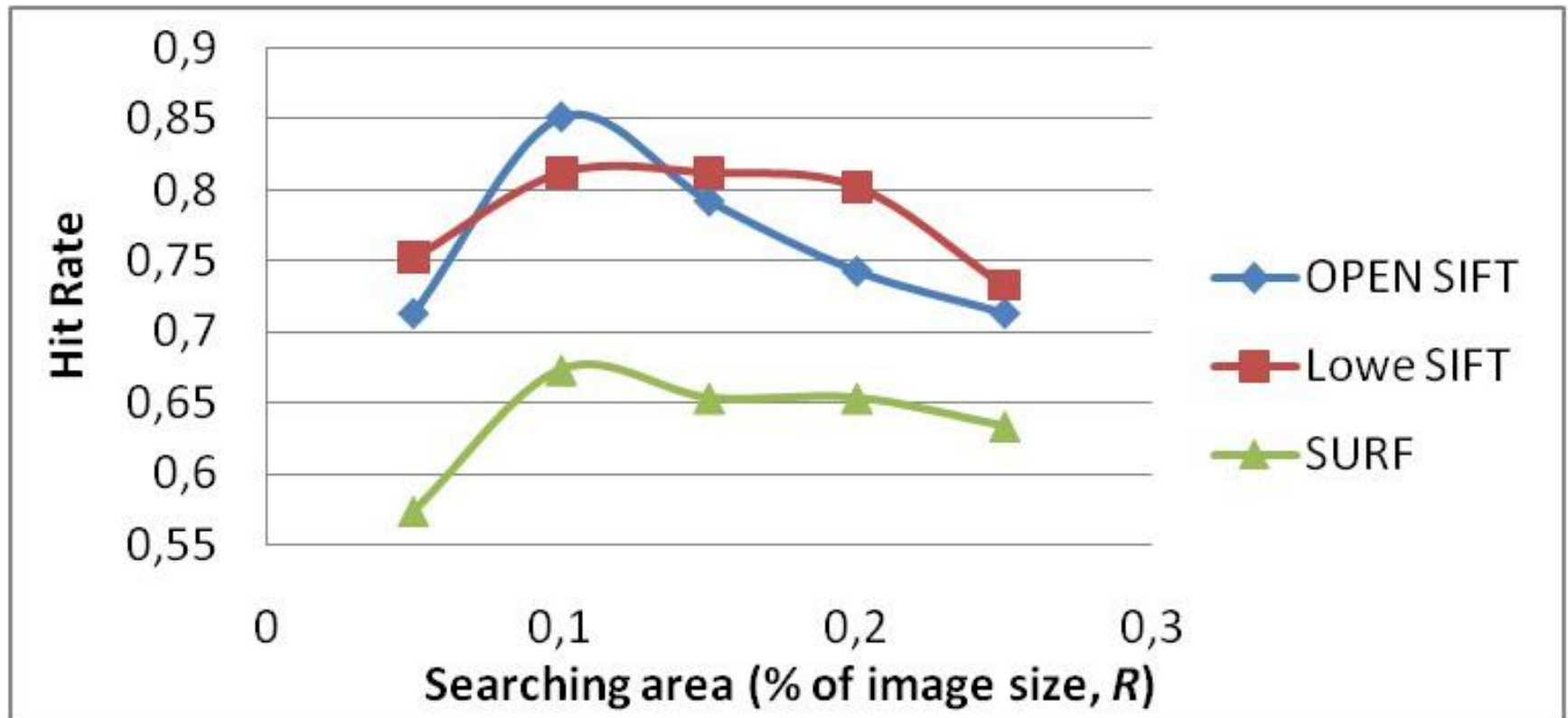
# Preprocessing

- ▶ Segmentation - manually ✓
- ▶ Noise-filtering ✗
- ▶ Slant correction ✗
- ▶ Word image resizing ✓
- ▶ Binarization ✗
- ▶ Skeletonization ✗



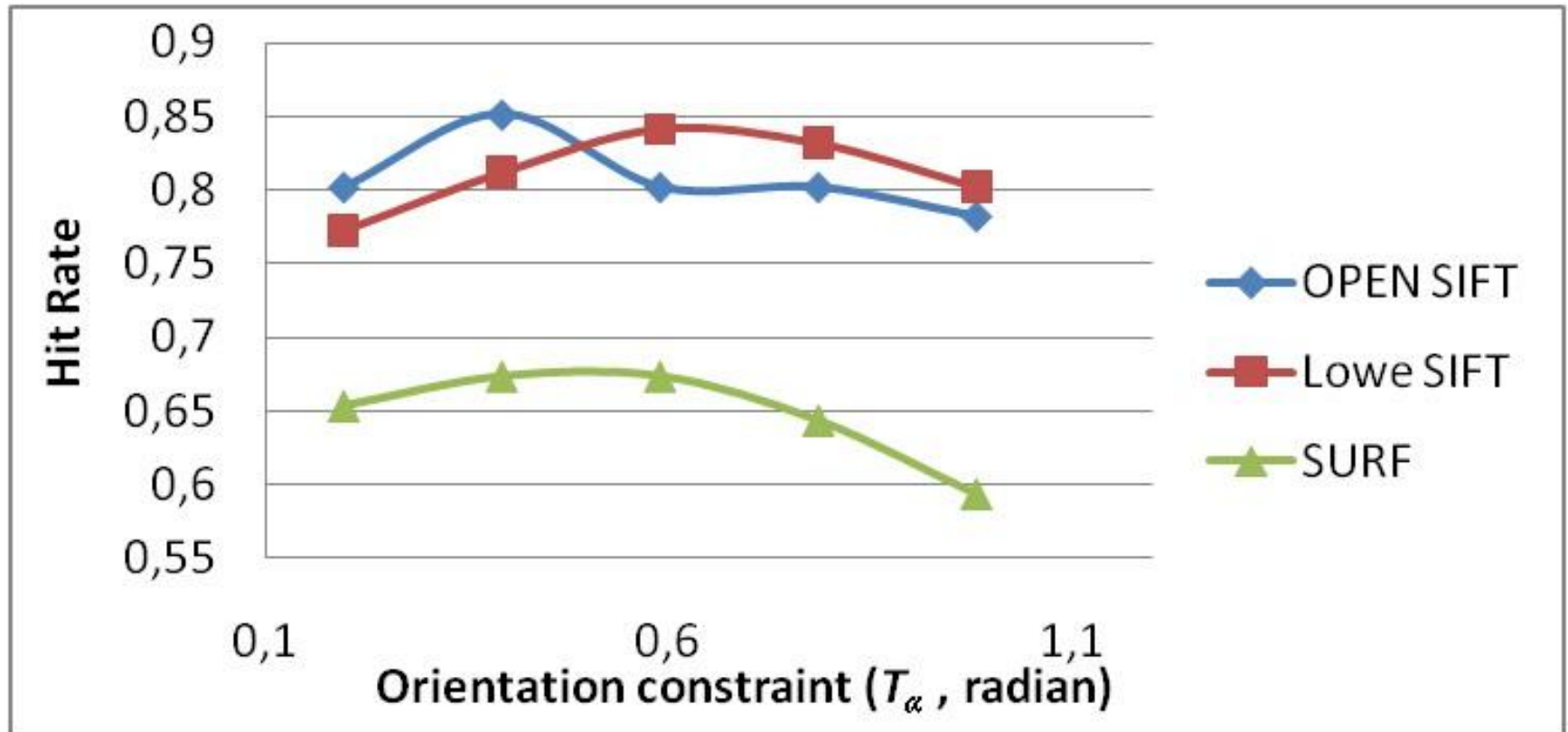
methods caused new problems... gave no real improvement

# Effect of searching distance



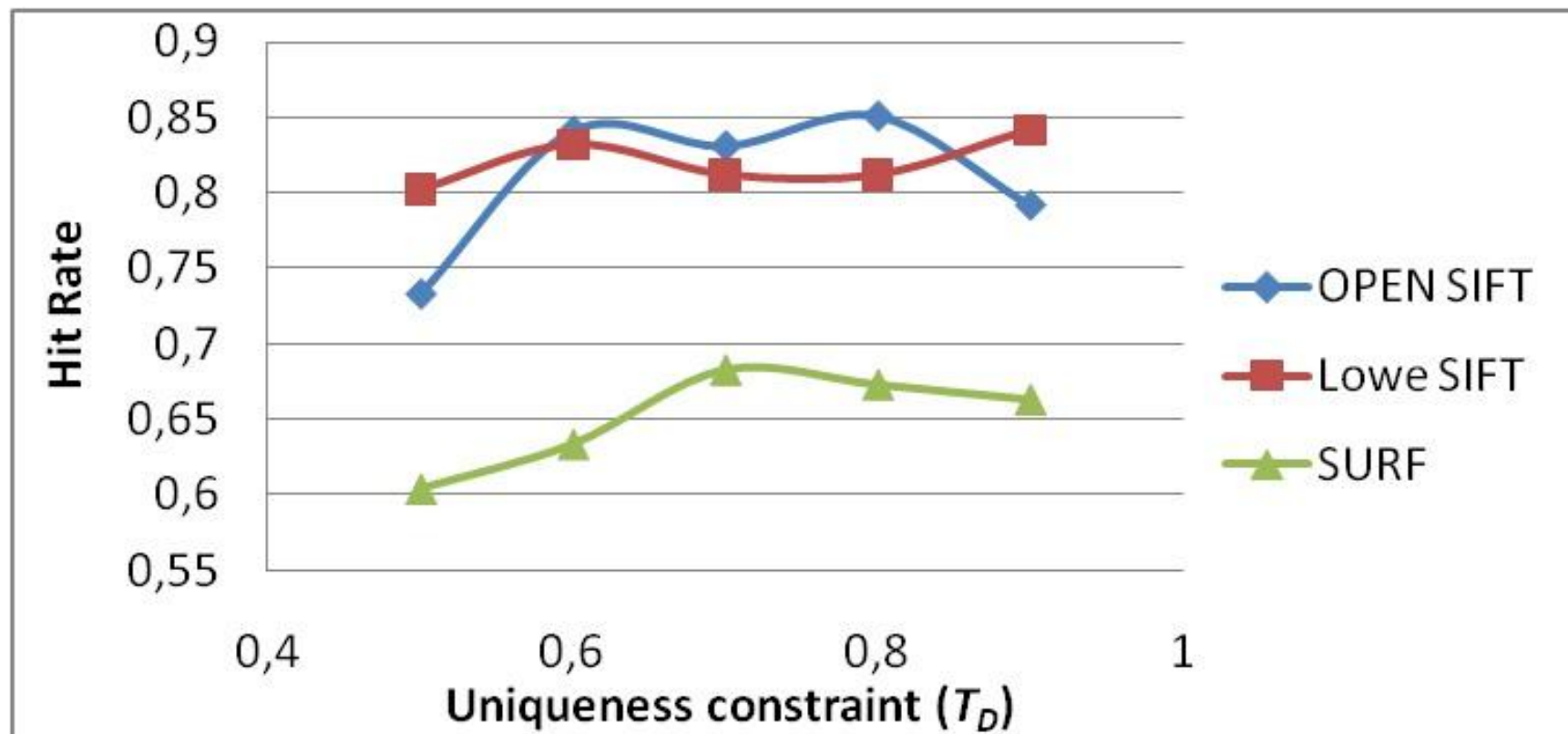


# Effect of orientation constraint



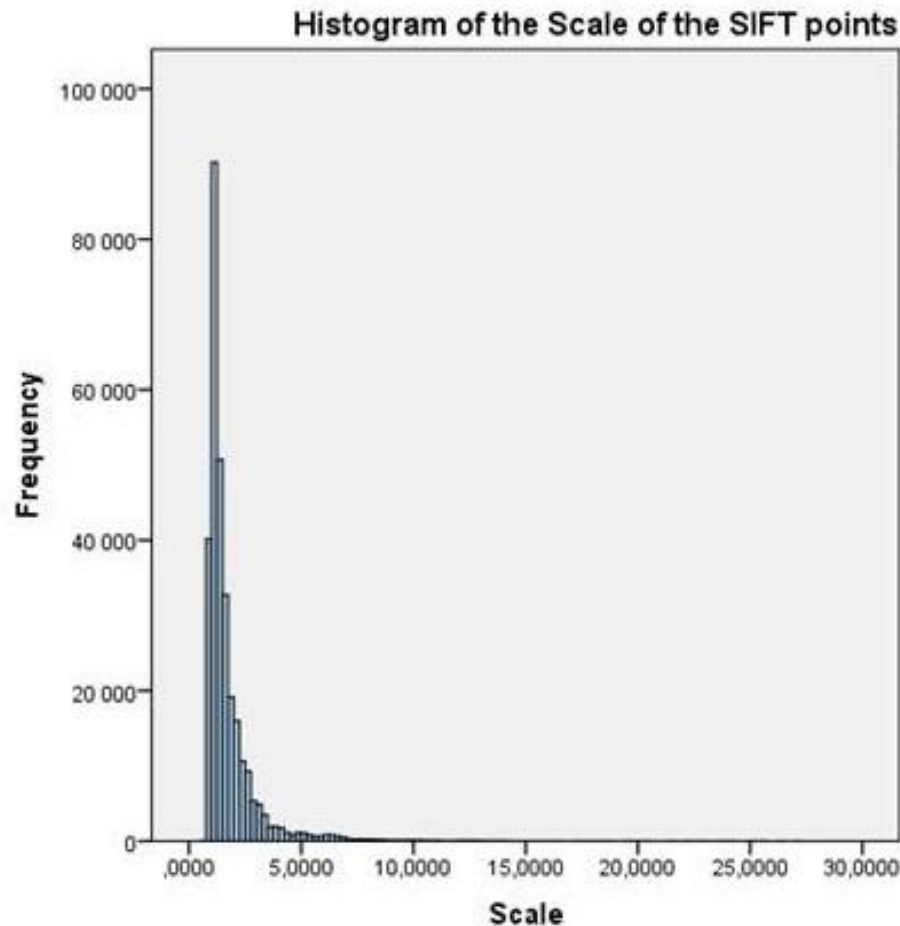
- Relatively stable performance of SIFT
- SURF is much behind

# Effect of uniqueness constraint

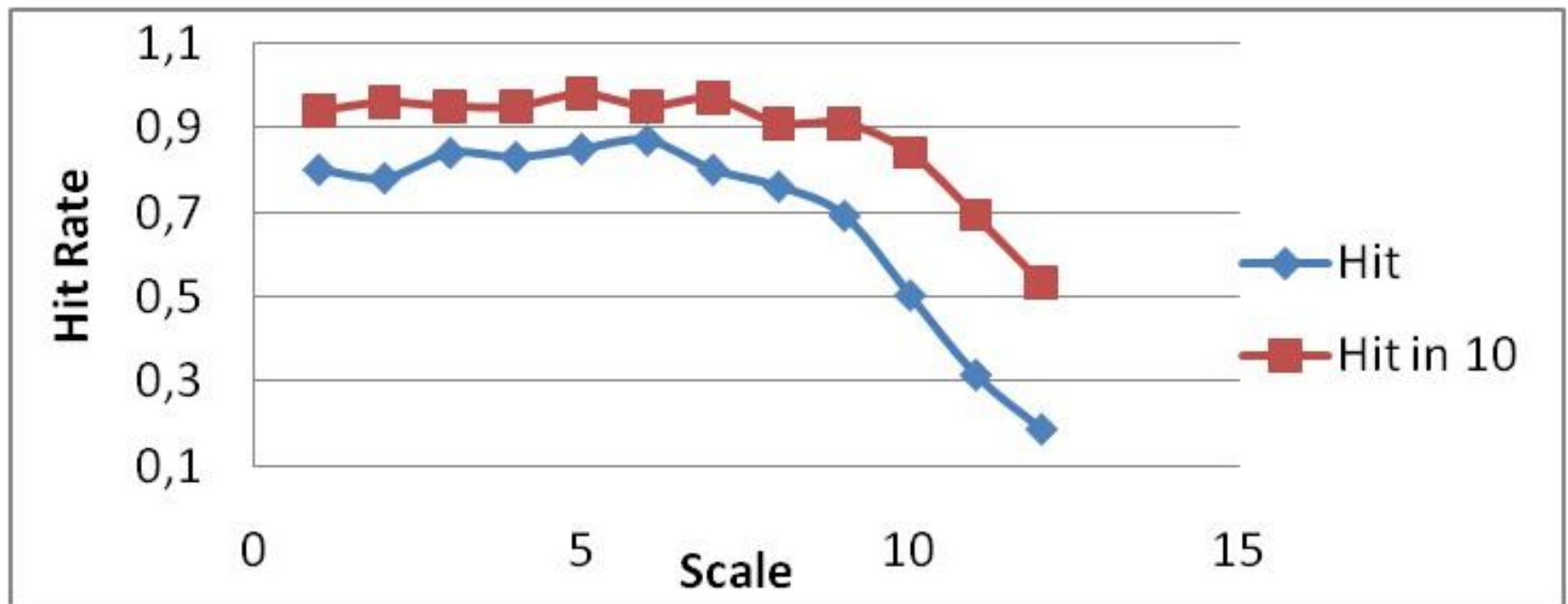


- Relatively stable performance of SIFT
- SURF is much behind

# Distribution of Scale of over 300000 SIFT points from archive text



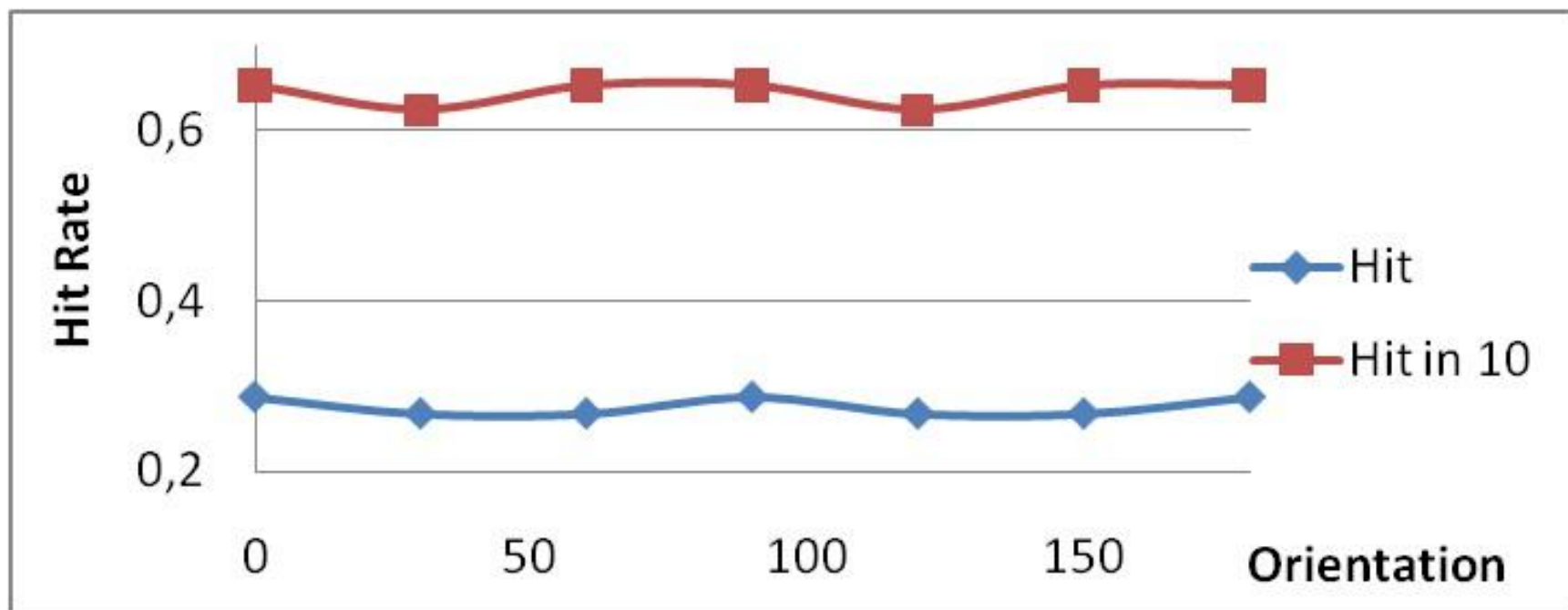
# Effect of fixing the Scale



**Scale information has relatively low importance in recognition.**

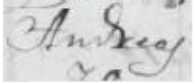
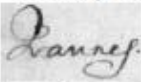

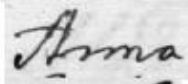
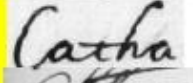
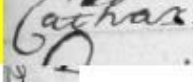
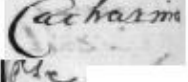

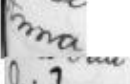
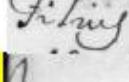
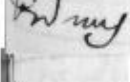
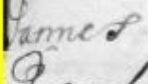
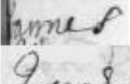
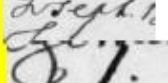
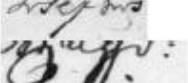
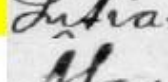
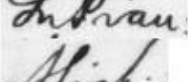
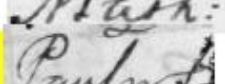
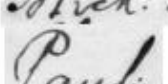
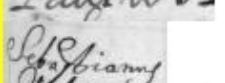
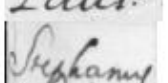
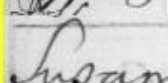
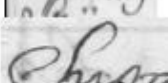
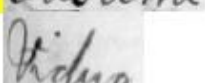
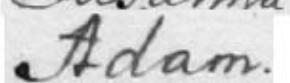
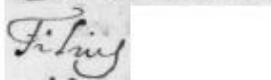


# Effect of fixing the orientation



**Constantly low performance of rotation sensitive descriptors tells us that, while written text is basically horizontal, but rotation invariance can't be neglected!**

# Analysis of results

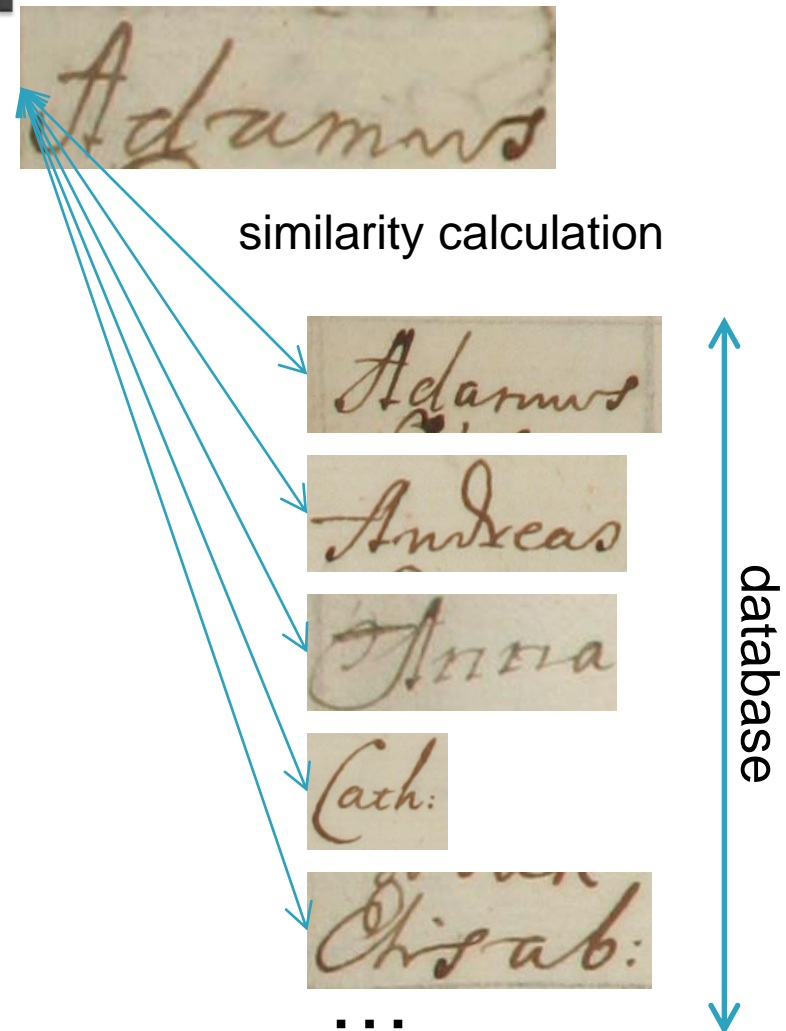
Ground truth	Wrong recognition	Ground truth	Wrong recognition
Andreas	Ivannes		
Anto	Anna		
Catha	Cath		
Cathar	Catharina		
Eva	Anna		
Filius	Viduus		
Jannes	Ivannes		
Joseph	Josephus		
Julia	Julian		
Math	Mich		
Paulus	Paul		
Sebastianus	Szephanus		
Susana	Susanna		
Vidua	Adam		
Viduus	Filius		

The list and images of mistaken recognitions from 101 random queries. Yellow words indicate classes of almost the same names.

Recognition error (in the test database ) could be halved by grapheme processing.

# Sequential search

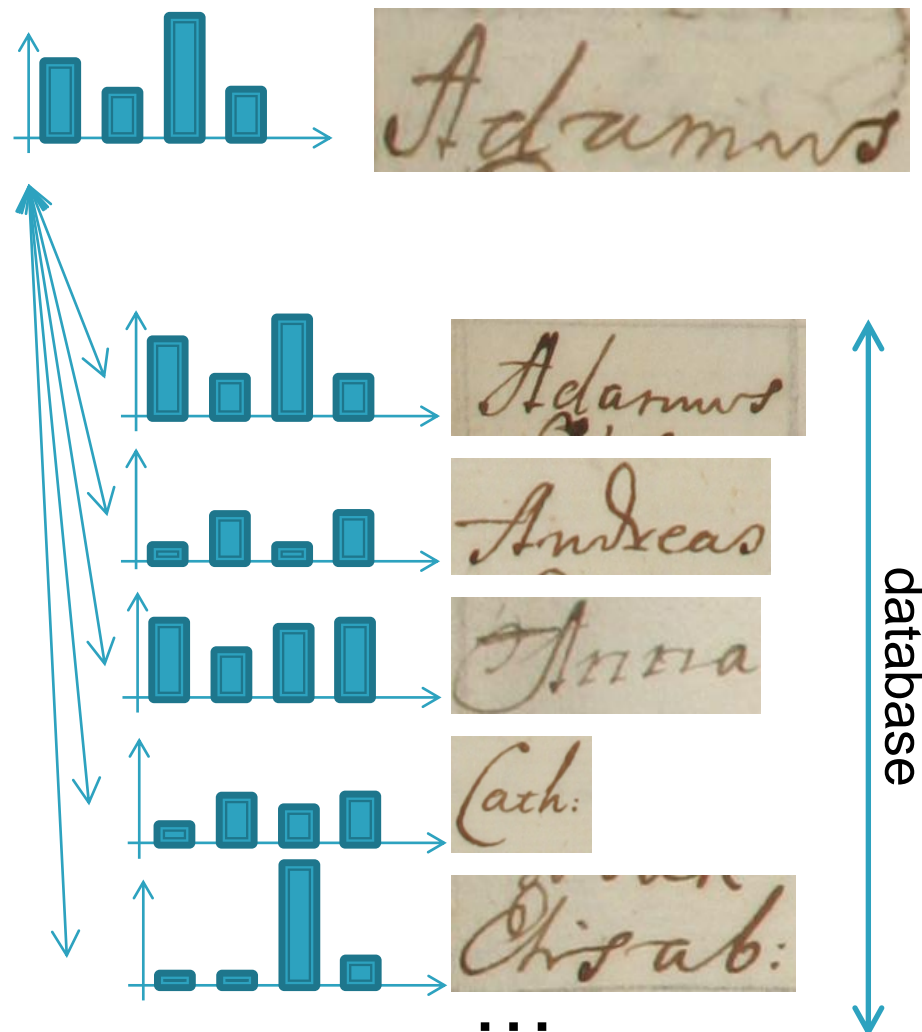
1. local feature extraction (SIFT)
  2. calculating similarity value with the images of the database
  3. searching the word with maximal similarity value
- ▶ the similarity calculation is slow
  - ▶ long running time



# Bag of Words (typical for SIFT)

1. local feature extraction (SIFT)
2. create feature cluster histograms
3. calculating similarity values between histograms (eg. correlation)

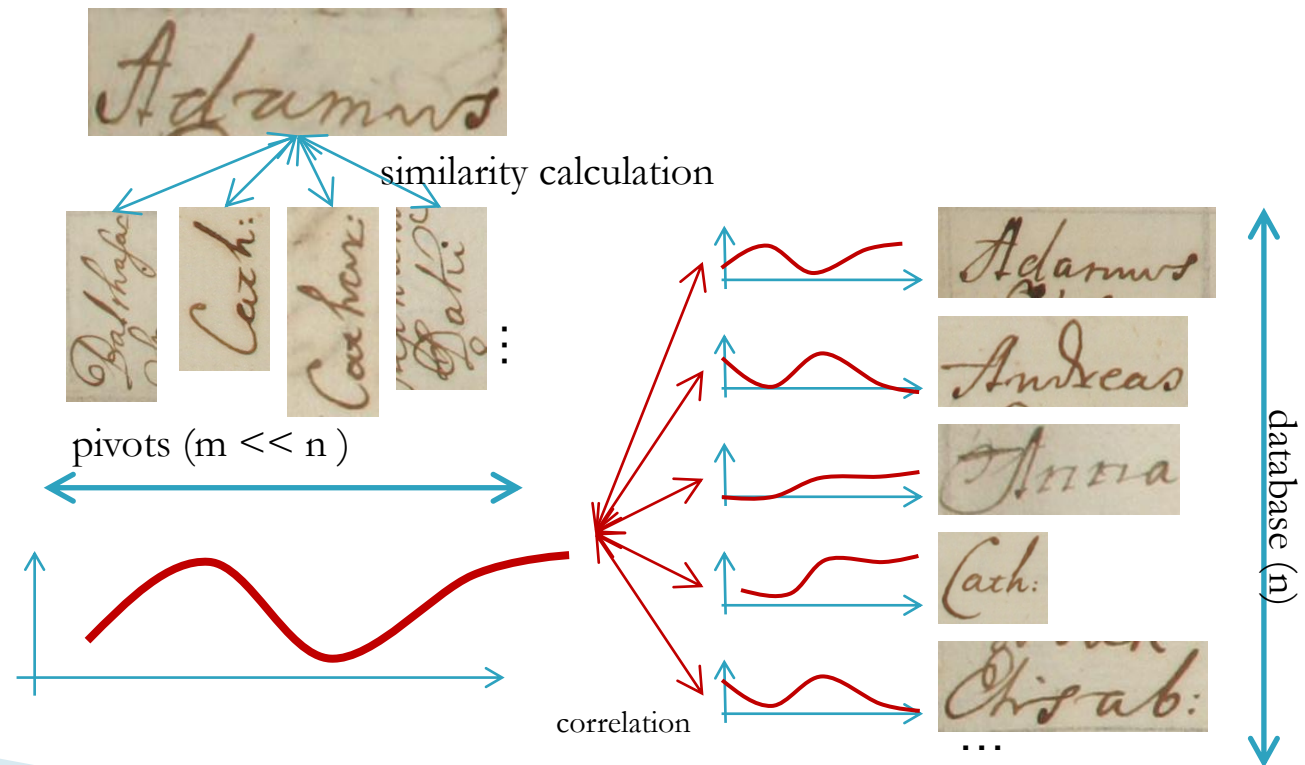
- features are too sparse/similar
- poor recognition rate





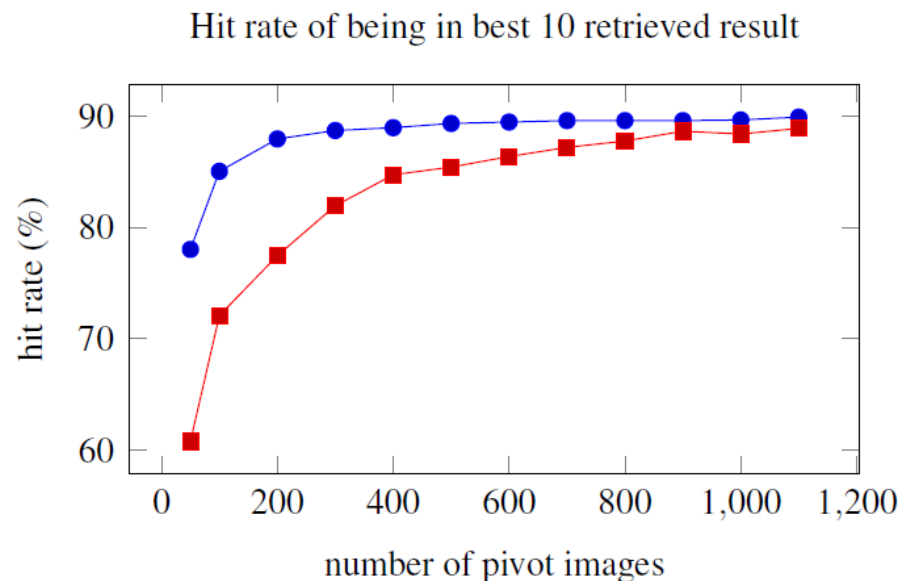
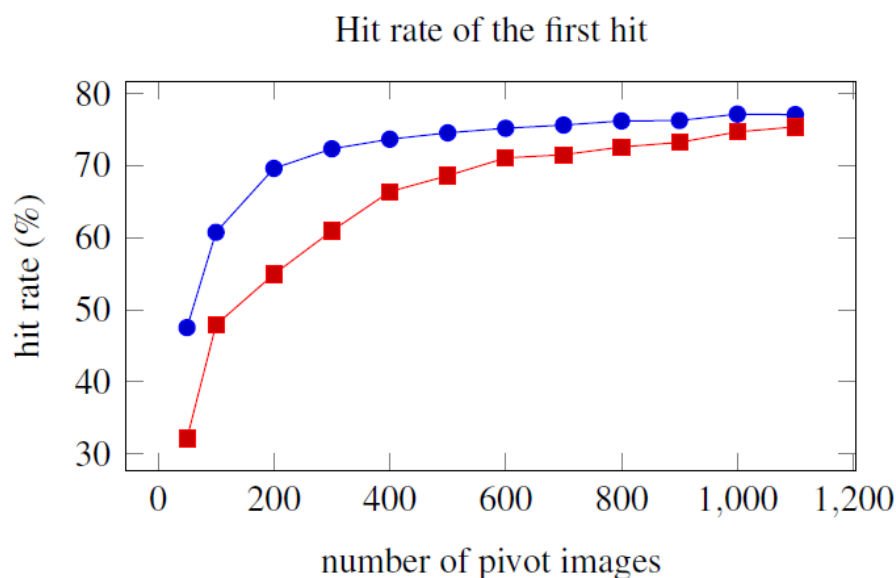
# Pivot based searching

1. local feature extraction (SIFT)
2. create similarity values with the ***pivot images***
3. correlation calculation between the pivot similarity values („function”) and the images of the database



# Pivot based searching results

- ▶ hit rate is about 70-75 %
- ▶ 2-3 times faster searching depending on the size of the database
- ▶ pivot selection problem (SVM, FSTAT)



—●— using pivots by SVM; —■— using pivots by FSTAT

# Conclusion

- ▶ **Not localized feature** descriptors are not proper for noisy archive handwriting recognition
- ▶ **SIFT based retrieval** can reach around **85% hit-rate** in case of cursive handwritten text with limited vocabulary
- ▶ Around 100% in the first 10 of the **result list** (manual correction is possible)
- ▶ **No need** for:
  - Preprocessing (e.g. binarization, slant correction, morphology)
  - Noise filtering
  - Precise segmentation
- ▶ **Rotation invariance** is more important than scale invariance
- ▶ Pivot based search can increase speed 2-3 times with small loss in retrieval rate
- ▶ Not all popular descriptors are adequate (SURF is faster but has significantly lower performance)

# Thank you for your attention!

**ACKNOWLEDGEMENTS :** This research was supported by the Hungarian Government and the European Union and co-financed by the European Social Fund under project TÁMOP-4.2.2.C-11/1/KONV-2012-0004. László Czúni was supported by the Bolyai scholarship of the Hungarian Academy of Sciences.