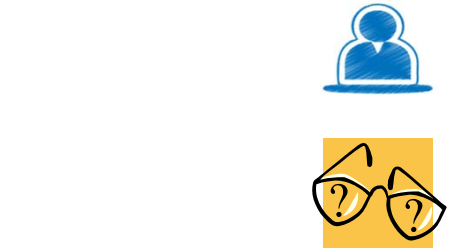
Historical Document Analysis

Marcus Liwicki
University of Fribourg
University of Kaiserslautern
Insiders Technologies GmbH
marcus.liwicki@unifr.ch

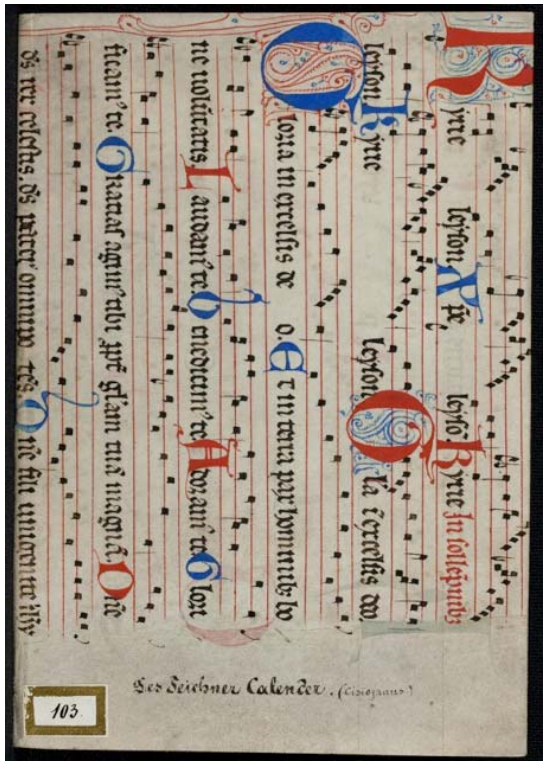
Typical Tasks of Scholars in the Humanities

- Cataloging
- Transcribing
- Searching
- Comparing texts
- …



- It’s hard to find interesting and relevant doucments

State of the Art Tool in the Humanities: Catalogs



Donaueschingen 103

Cisioianus Ps.-Heinrichs des Teichner · Cisioianus des Steyrer · Mariengebet

Papier · 4 Bll. · 20,5 × 14,5 · Österreich · um 1370/1380

Zustand: Griffspuren und Verschmutzung (v. a. Bl. 1v/2r, 3v); Wasserflecken (Bl. 2v/3r); Wurmfraß (nur Bl. 4).

Moderne Bleistiftfoliierung *1-4*. Wz. (nur Bl. 4): Einhorn (fragmentarisch, wohl halbe Figur), zur Motivgruppe PICCARD Fabeltiere III 1447-1455 (Perugia, Treolis, Volterra, Siena, Arezzo, Sta. Fiora, Cortona 1380-1382). Das Papier weist eine grobe Rippung auf (Abstand zwischen den Siebdrähten ca. 2,5-3 mm), die im Verlauf der 1380er Jahre üblicherweise außer Gebrauch kam. Lage: I + 2^4. Bei Bl. 2 handelt es sich um ein Einzelblatt mit einem Falz, an den Bl. 3 angeklebt ist. Schriftraum, wenig sorgfältig mit Tinte begrenzt: 15 × 11-12. 24-26 Zeilen, häufig kein gerader Zeilenverlauf. Textualis auf einfachem kalligraphischen Niveau von einer Haupthand der 2. Hälfte des 14. Jh.: doppelstöckiges, meist nicht kastenförmiges a, kleiner, hochgezogener g-Bogen, rundes und gerades r regellos verwendet (vgl. SCHNEIDER, Paläographie, S. 51-53). 3v: Kursive von einer Nachtraghand der 1. Hälfte des 15. Jh. Verse nicht abgesetzt. Rubrizierung (nur 1v-3r): Majuskelstrichelung, rote Überschriften, ein- bis zweizeilige Lombarden. 3v Manicula des 15. Jh.; nachträgliche Federzeichnungen des 16. Jh.: drei Köpfe, links von späterer Hand ungelenk ein Körper ergänzt.

Einband um 1860-1865: Pergament auf Pappe unter Verwendung eines Blattes wohl aus einem Kyriale (2. H. 13. Jh.), vgl. die Angaben bei Don. 69. Auf dem Vorderdeckel unten Titel von der Hand Baracks, der wohl die Einbindung veranlasste: *Des Teichner Calender. (Cisiojanus.)*. Zu weiteren Donaueschinger Hss. mit gleicher Einbandart s. Don. 69.

Geschichte: Entstanden im Süden Österreichs um 1370/1380. Auf eine Entstehung im Erzbistum Salzburg oder in dessen Suffraganbistum Graz spricht neben der Schreibsprache auch die Auswahl der Heiligen, vgl. PICKEL (s. u., Lit.), S. 60 (für den Cisioianus Ps.-Heinrichs des Teichner) und M. J. SCHUBERT, Der Cisiojanus des Steyrer in Krakau, in: ZfdPh 116 (1997), S. 32-45, hier S. 38 (für den Cisioianus des Steyrer). Die Entstehungszeit lässt sich durch die Beschaffenheit des Papiers und den paläographischen Befund eingrenzen. Als weiteres Indiz kann die frühe Parallelüberlieferung der beiden Cisioiani herangezogen werden: Wien, ÖNB, Cod. 2817, Schwaben um 1370/1390 (71vb: Cisioianus Ps.-Heinrichs des Teichner; vgl. MENHARDT I, S. 327-340); Stockholm, Kgl. Bibl., Cod. A 175, Österreich 1375-1382 (346v: Cisioianus des Steyrers; vgl. KURRAS 2001, S. 39f. u. Abb. 12). Bereits im 15. Jh. scheint sich die Hs. im ostschwäbischen Sprachgebiet befunden zu haben (Schreibsprache des Nachtrags).
Auf 1r Bleistiftnummer des 19. Jh.: *588e*. Auf dem Vorderspiegel Bleistiftsignatur: *Don. 103*, auf dem Vorderdeckel Donaueschinger Signaturschild *103*. 1r, 4v Donaueschinger Bibliotheksstempel 2.

Schreibsprache Haupttext: Südbairisch. Merkmale: 1) nhd. Diphthongierung; 2) ai-Schreibung für mhd. ei; 3) Dehnung von mhd. i (*wíer, dier*); 4) b erscheint als p im Anlaut (außer bei der Vorsilbe be-); 5) k erscheint in allen Positionen als ch (*chind, chrau'tz, iunchfrawn, danch*) oder kch (*iarmarkcht*).
Schreibsprache Nachtrag: wohl Ostschwäbisch. Merkmale: 1) teilweise durchgeführte nhd. Diphthongierung (*beÿ, auß, raw* neben *min, duzzent, riuwe*); 2) Anlautend sch vor m (*schmerczen*); 3) immer b- und k-Anlaute, nie bairisch p-/ch-; 4) die für das Schwäbische typische Vorsilbe her- (*herwelte*), vgl. MOSER, Frnhd. Gramm. I.3., § 128, 4b, S. 8; 5) teilweise d-Schreibung für mhd. t im Anlaut (*duzzent, dun*), vgl. MOSER, Frnhd. Gramm. I.3., § 143, 1a, S. 158. Auffällig ist die Schreibung au, aw für mhd. iu in *kauschе, raw*. Diese Form ist in HSS, Karte 49 nur für das Rheinfränkische belegt.

Literatur: BARACK, S. 99f.; K. PICKEL, Das Heilige Namenbuch von Konrad Dangkrotzheim, Elsässische Litteraturdenkmäler aus dem XIV.-XVII. Jahrhundert I, Straßburg 1878, S. 47; SCHUBERT (s. o., Geschichte), S. 33; M. MUELLER, Beherrschte Zeit. Lebensorientierung und Zukunftsgestaltung durch Kalenderprognostik zwischen Antike und Neuzeit (Schriften der Universitätsbibliothek Kassel - Landesbibliothek und Murhardsche Bibliothek der Stadt Kassel 8), Kassel 2009, S. 38.

1v-2v Cisioianus Ps.-Heinrichs des Teichner

>Daz ist dez Teichner kalender.< >B<esniten ist daz chind, / drei chünig sagtt Erhartz gesind, / der stern weist si. / Wenn chom Marcellus, Antoni? … – … Herr, gib vnserm leben frist. / Thomas chündet vns geporn Christ. / Stephan, Hansen, chindel, Thomas freunt ist.

Edition: PICKEL (s. o., Lit.), S. 61-65. Literatur: A. HOLTORF, Cisioianus, in: ^{2}VL 1 (1978), Sp. 1285-1289, hier Sp. 1288 (Nr. B.10); PICKEL, S. 59-65; H. A. HILGERS, Versuch über dt. Cisiojani, in: V. HONEMANN u. a. (Hg.), Poesie und Gebrauchsliteratur im dt. Mittelalter, Würzburger Colloquium 1978, Tübingen 1979, S 127-161, hier S. 155f.
Dieser Text auch in BLB, St. Georgen 60, 158r-160v.

2v-3r Cisioianus des Steyrer

>Daz ist dez Steyrer kalender.< >N<eu ist daz iar in perichten langt, / Erhart, nach dier ist dem Felix gar ant. / Brisca, Fab., nes, Vinczent wart. / Paulus, der hat sich bechart … – … Hilf suezzen Barbar, Nyclos fraw, / daz vns der teufel lützel schw. / Hie lau^{t}tt mit fuchs swantz To^{o}mel mi^{n}r spat, / lau^{t}t Christ, Stephan, Hanns, chind, Toemel dort drat.

Editionen: PICKEL (s. o., Lit.), S. 49-51 (nach dieser Hs.); R. M. KULLY, Cisiojanus, Studien zur unemnonischen Literatur anhand des spätmittelalterlichen Kalendergedichts, Schweizerisches Archiv für Volkskunde 70 (1974), S. 99-123, hier S. 106 (Teilabdruck); HILGERS (s. o., zu 1v-2v), S. 157-159; SCHUBERT (s. o., Geschichte), hier S. 43-45 (synoptischer Abdruck der verschiedenen Fassungen). Literatur: PICKEL, S. 46-51; HILGERS, S. 154, 157-161; SCHUBERT (s. o.).

3v Mariengebet (Nachtrag)

(Überschrift:) *Wer daz gebet teglichen spricht mit raw vnd andacht, den wil Maria, gotes muter, behuten vor sunden vnd vor schanden.*

(Gebet:) *Maria, du bist gnaden vol, / al diß welt dich loben sol. / Du bist ein heil des todes schmerczen. / Hilff, Maria, allen betruten ellenden herczen … – … der ward gedach jn der hohen gotheit / vnd floß durch die heilgen triveltikeit / vnd ward gewirket da von dem heilgen geist. Amen.* – 4r/v leer.

- But automatic methods can help!


Vision: DIVADesk
A scientific workbench for scholars

Outline

- **Challenge: Why historical Documents?**
- State-of-the-Art
- Recent Trends
- DIVAServices: Approach Towards Interoperability

What is the main Challenge?

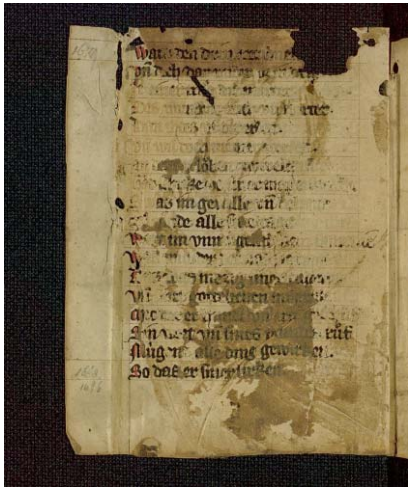
- Data variation?

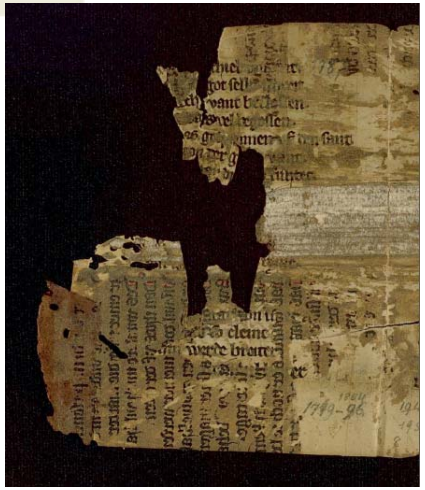
Data Variation

- Different languages and alphabets
- Writing style differs
- Quality of the images/data
- Changing writing instruments
- Abbreviations and misspellings
- Graphics & handwriting
- Language and writing evolves
- Annotations
- Change of support

What is the main Challenge?

- Data variation?
- Degradation?



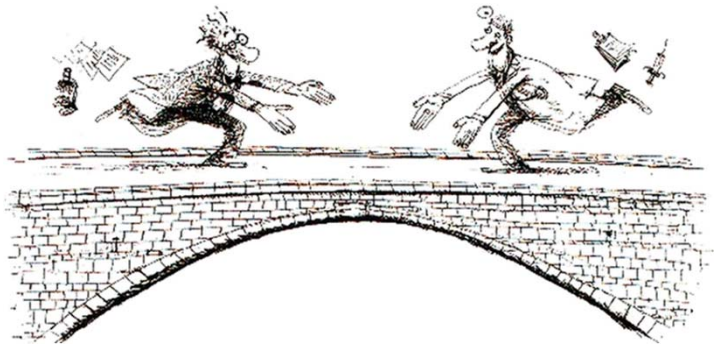


What is the main Challenge?

- Data variation?
- Degradation?
- Communication between humanist scholars and computer science experts!

Communication between Humanist Scholars and DIA Experts

- Different expectations
  - Clearly defined challenging datasets vs. useful systems
- Bridging the gap is the biggest challenge



Success in Computer Science ?!?

- HIP 2011 (27 papers accepted)
  - Information retrieval (text / graphic)
  - Projects
  - Text/Character recogniti
  - Visualization
  - Digitization
- HIP 2013 (18 papers accepted)
  - Information Extraction and Retrieval
  - Reconstruction and Degradation
  - Text and Image Recognition
  - Segmentation, Layout Analysis and Databases
- HIP 2015 (18 papers accepted)
  - Text Transcription
  - Segmentation and Layout Analysis
  - Templates, Date Estimation, and Script Specific Approaches

But: ask a random scholar attending the Digital Humanities conference: Do you know about HIP?

Thanks to Mickael Coustaty, IDAKS 2016

Overview of Projects on Hist-OCR

* If you ask scholars who want to use the systems

- EU IM[illegible]ct (2008-2012)
- EU [illegible]IBUS (2012-2016)
- EU READ (2016-now)
- CIS, LMU München, Post-OCR Correction
- OCR-D Projekt DFG (since 2015, 1.5 Mio books)
- Early Modern OCR Project, Texas A&M(2012-2015)
- Kallimachos (Uni Würzburg, 2014-2017)
- Ocular, University of California, Berkeley (2013-now)
- …

Communication Problems and Approaches for Solution

For Computer Science Experts:

- Not a unique representation of knowledge
- Same content has a lot of interpretation
- A description is not shared by all scientists
- Focus on different aspects

For Scholars in the Humanities

- Methods are not understandable
- Not clear what 95% means
- Systems not accessible
- Too specific solutions

- We need more interdisciplinary discussions
- Reduce black box effects (describe methods, give examples)
- Approximate results are not enough
  - Interfaces needed
  - Alternatives to be reported

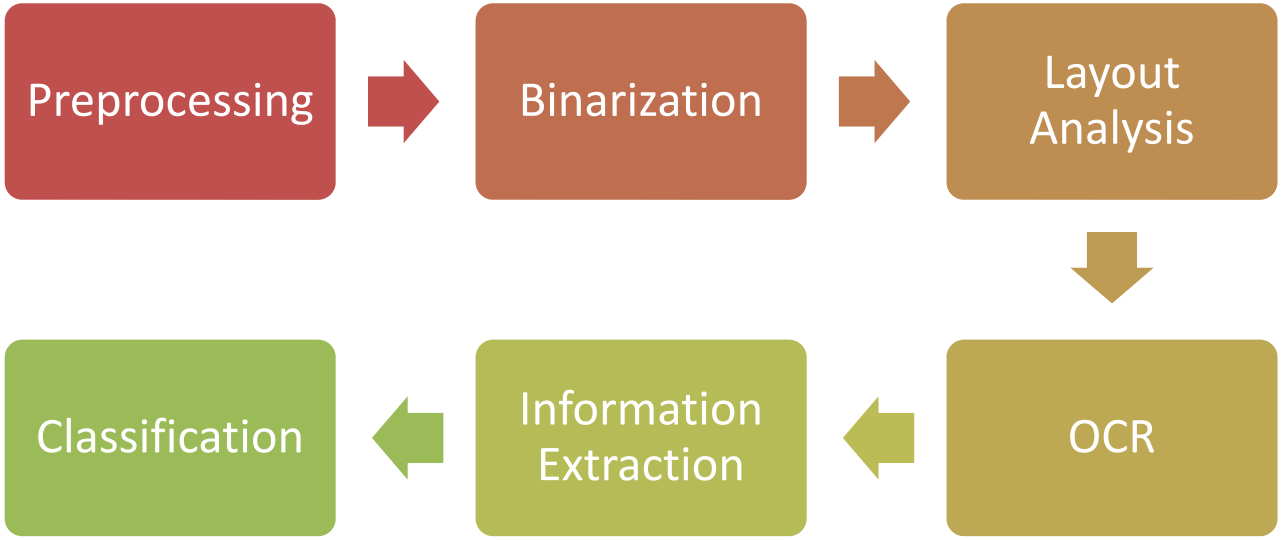
Outline

- Challenge: Why historical Documents?
- **State-of-the-Art**
- Recent Trends
- DIVAServices: Approach Towards Interoperability

Processing Steps of Automatic DIA

Threshold (local, global)
- Sauvola
Classification

Top-down vs bottom-up



Layout Analysis Methods

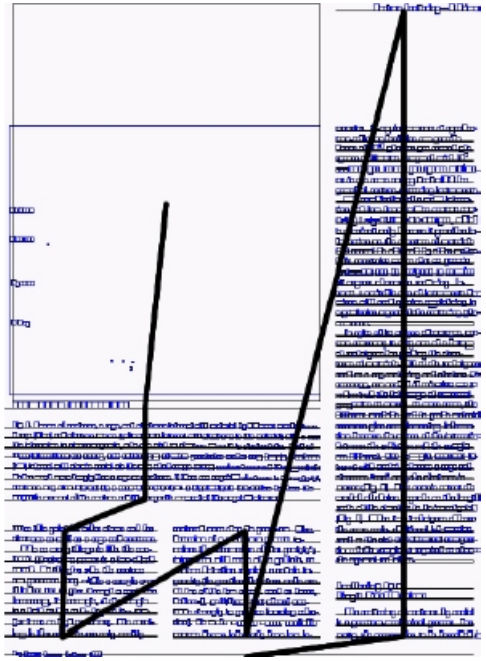


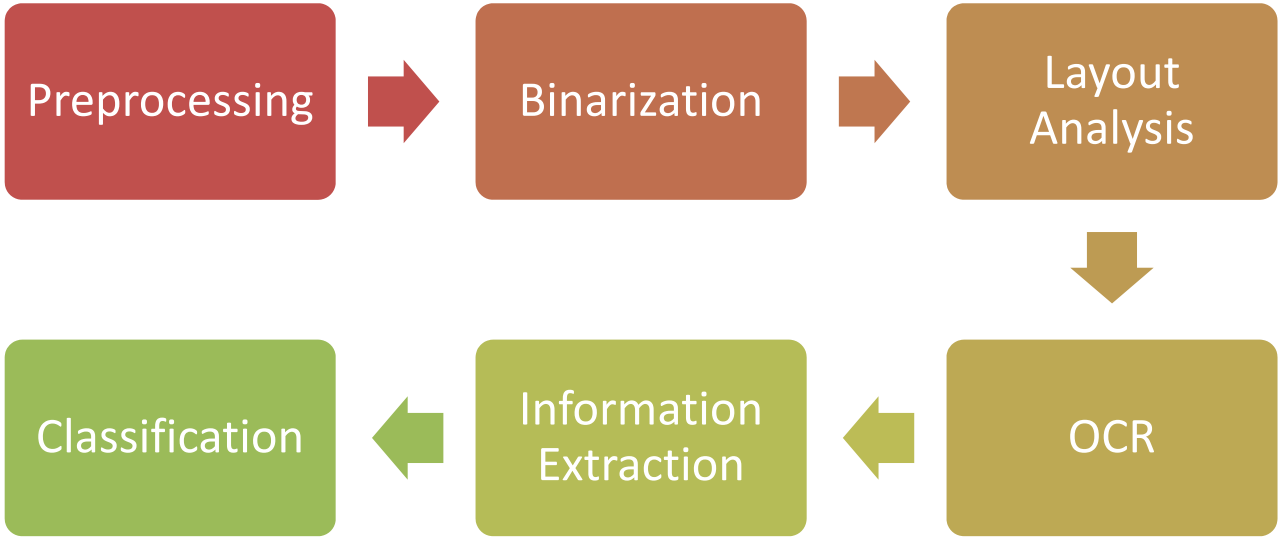
Figure 3 Projection profiles of a simple text binary image.

- Based on connected components
- XY-cut
- Other histogram-based approaches

Processing Steps of Automatic DIA

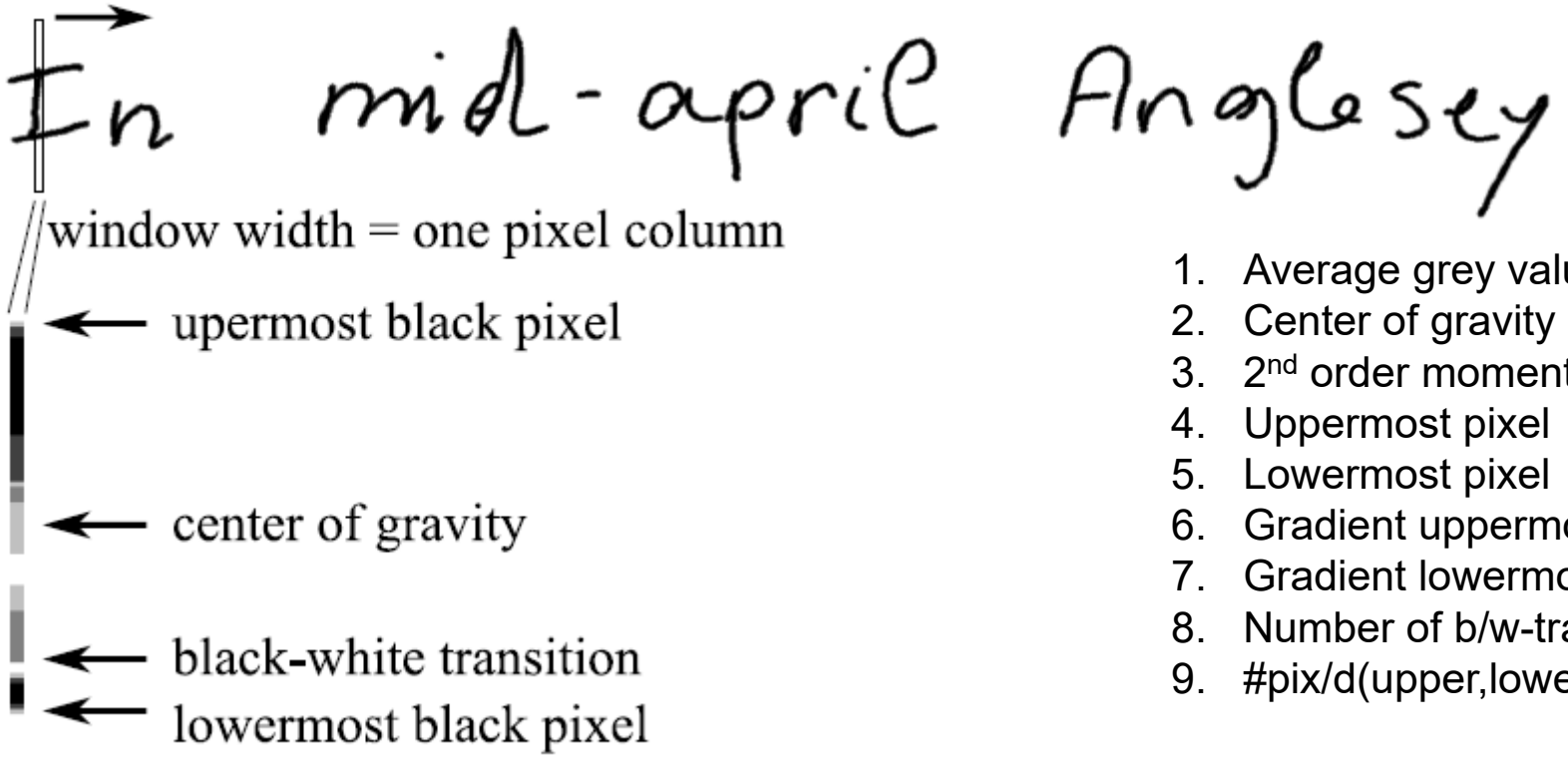
Threshold (local, global)
- Sauvola
Classification

Top-down
- XY-cut, histograms
Bottom-up
- Connected components



Feature Extraction

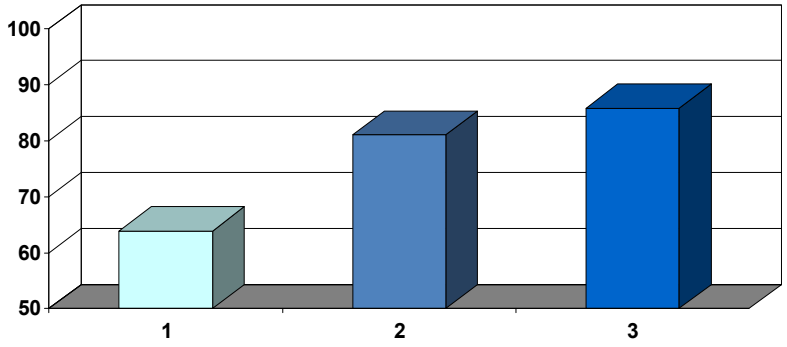
- Marti, Bunke (2001)



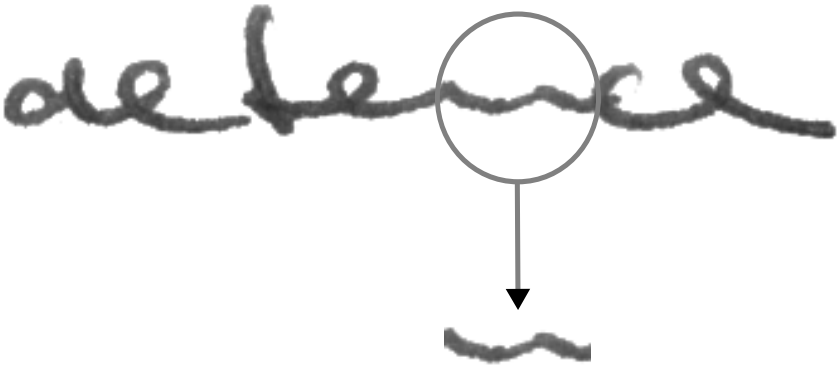
1. Average grey value
2. Center of gravity
3. 2nd order moment vert.
4. Uppermost pixel
5. Lowermost pixel
6. Gradient uppermost
7. Gradient lowermost
8. Number of b/w-transitions
9. #pix/d(upper,lower)

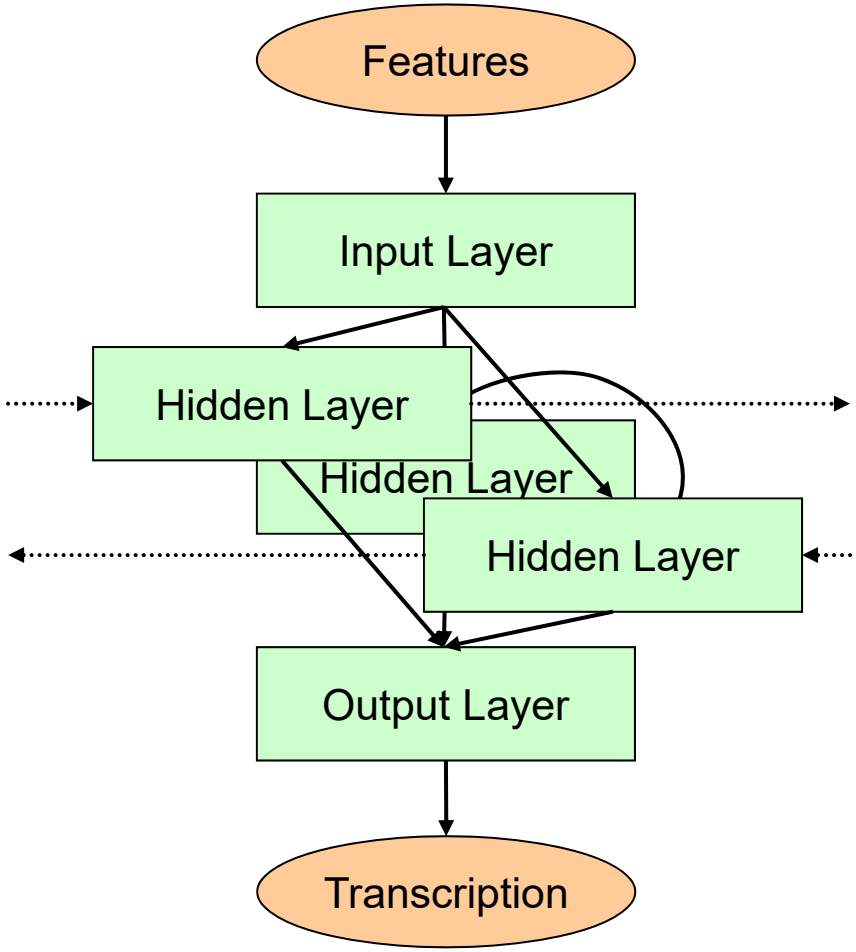
Classification

- Machine learning methods for sequences
  - HMMs
  - Recurrent NNs



Bidirectional Long Short-Term Memory Network



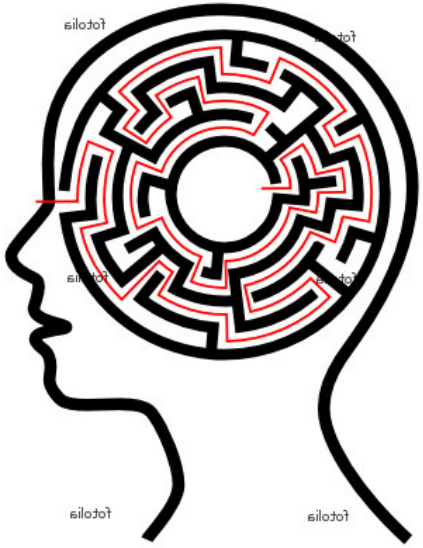


- Multilayer perceptron network
- Recurrent connections
- Bidirectional
- Memory instead of perceptron

Limits of MLP

- Limit: static input/output operation

$$x_1, \ldots, x_n \rightarrow y$$

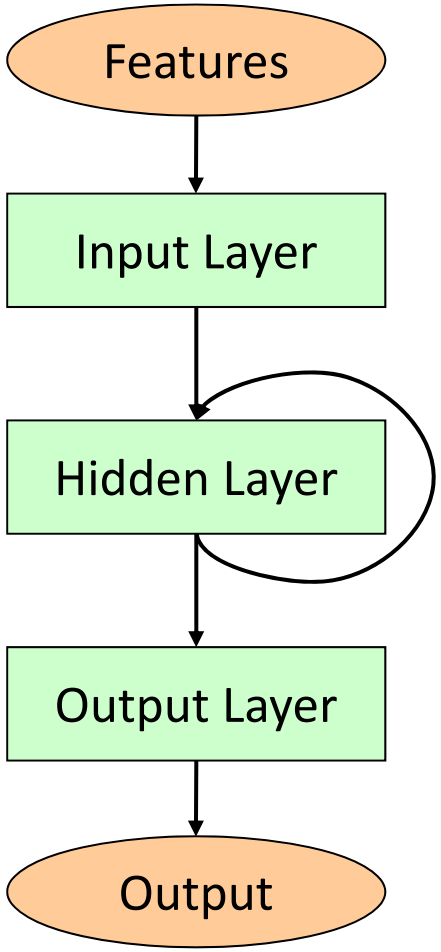


- Human brain is capable of memorizi
- Needed for solving many problems
  - Sequence recognition
  - Navigation through a labyrinth
  - Video analysis

$$\left((x_1^1, \ldots, x_n^1), \ldots, (x_1^T, \ldots, x_n^T)\right) \rightarrow (y^1, \ldots, y^U) \,|\, U \leq T$$

- Idea: add backward-connections to maintain state

Recurrent Neural Networks (RNNs)

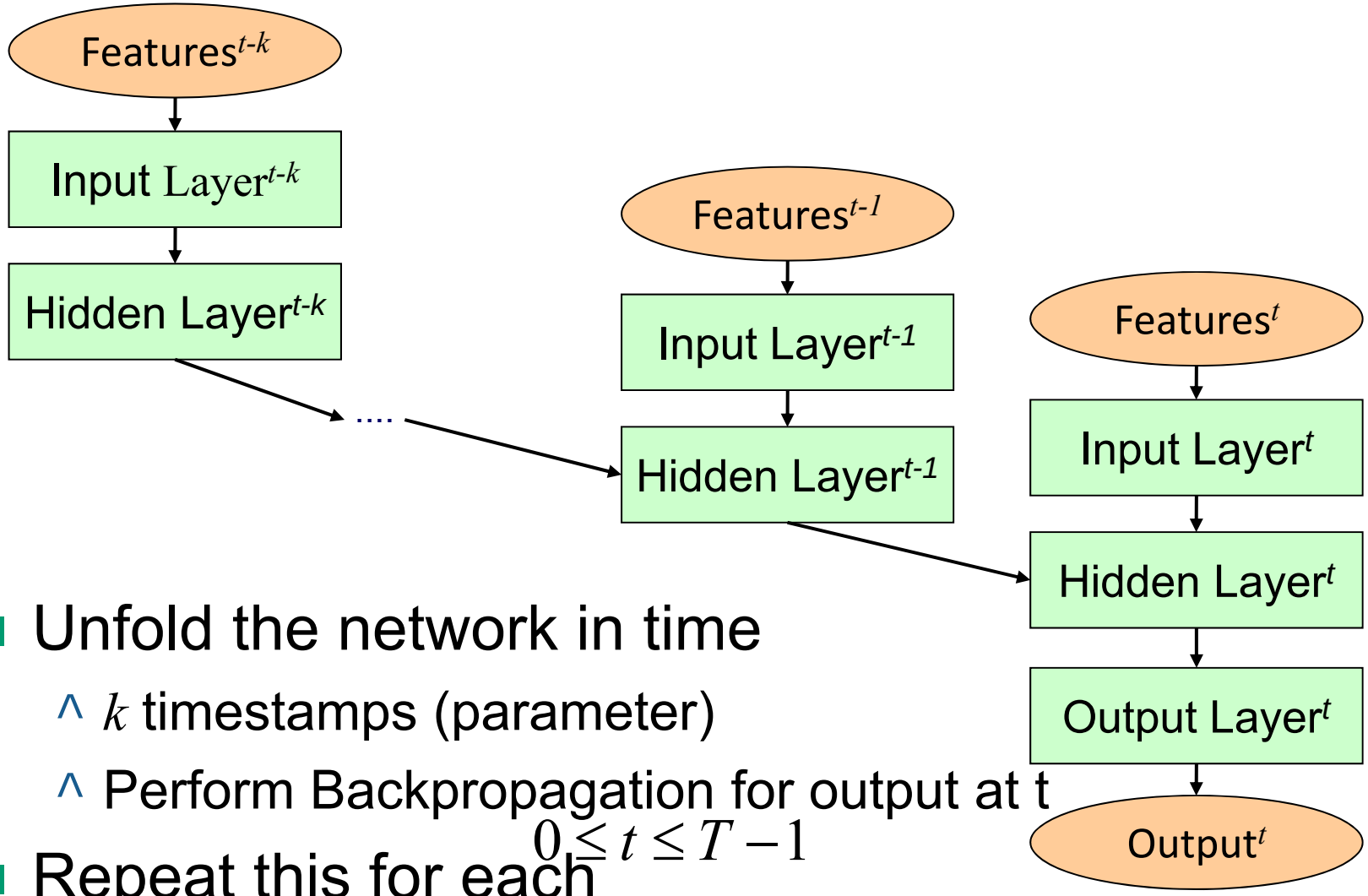


- Recurrent connections are added in order to keep information of previous time stamps in the network
- Novel equation for the activation:

$$a^t = \sum w_i x_i^t + \sum w_h b_h^{t-1}$$

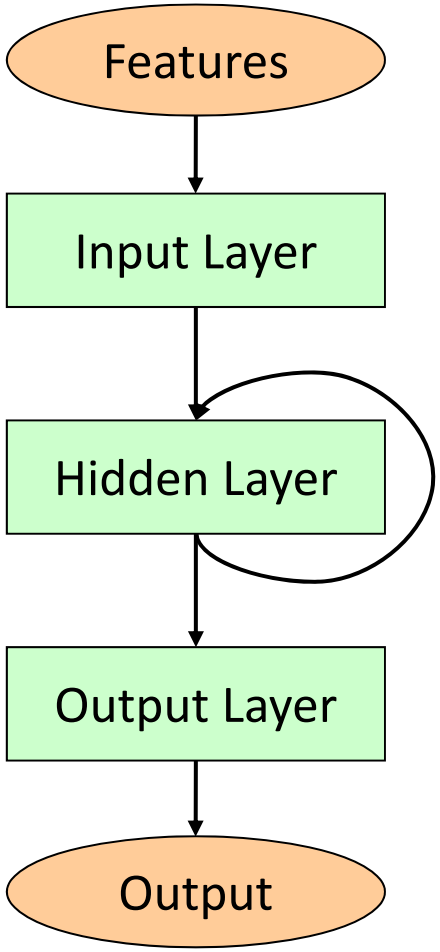
- Context information is used
- How to train those networks …?

Training of RNNs – Backpropagation Through Time



- Unfold the network in time
  - k timestamps (parameter)
  - Perform Backpropagation for output at t $0 \leq t \leq T-1$
- Repeat this for each

Recurrent Neural Networks (RNN)



- Recurrent connections are added in order to keep information of previous time stamps in the network
- Novel equation for activation:

$$a^t = \sum w_i x_i^t + \sum w_h b_h^{t-1}$$

- Can be written in matrix form

$$A^t = W_i \cdot X^t + W_h \cdot B^{t-1}$$

- Context information is used, however: impossible to store precise information over long durations

Vanishing Gradient

- Usual RNN forget information after a short period of time

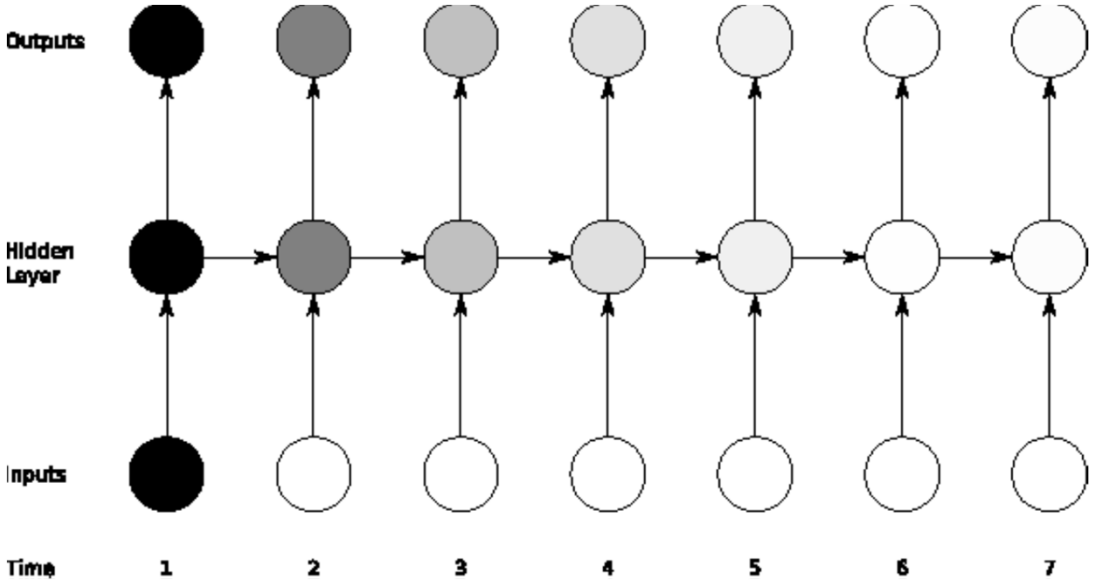
Example:

Neuron

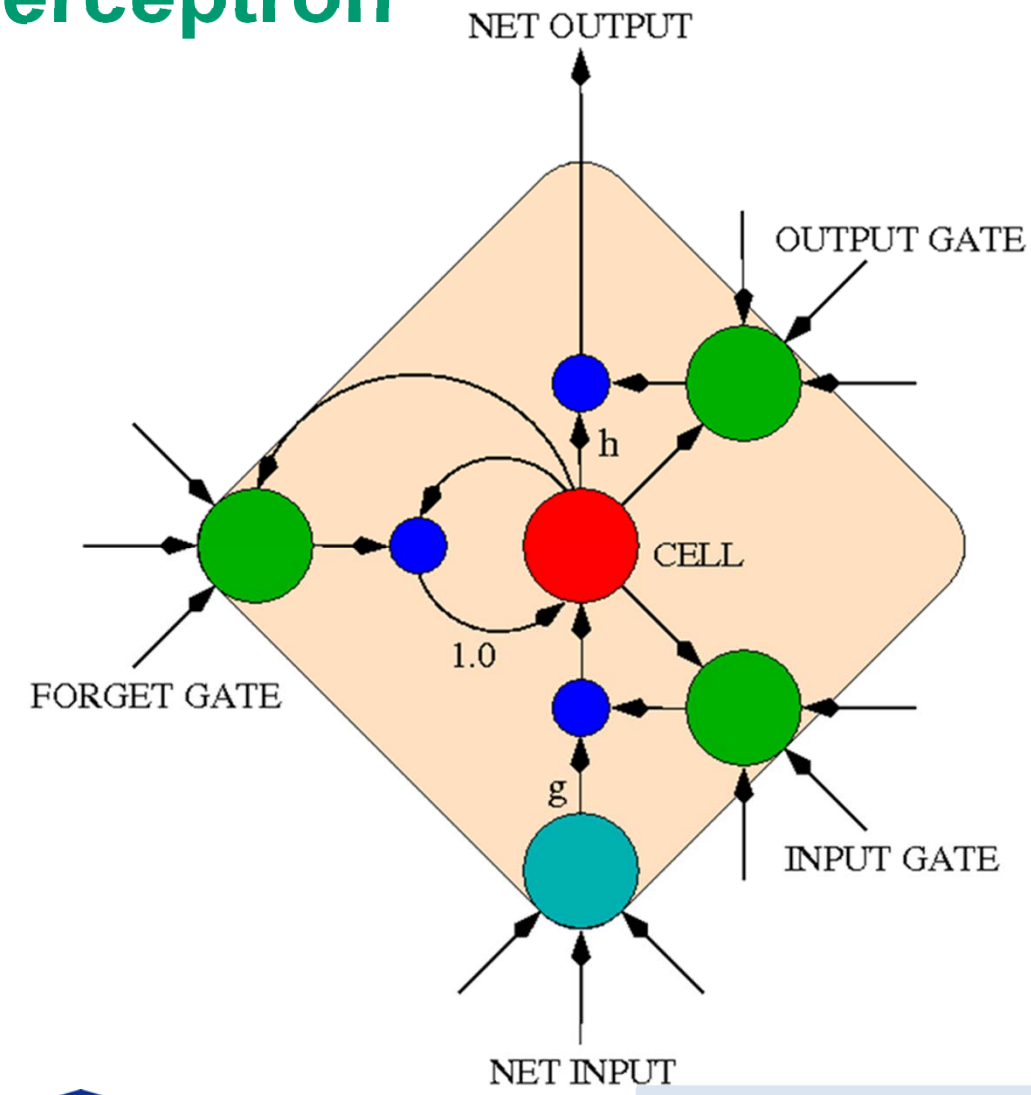
7 timestamps

Information

vanishes



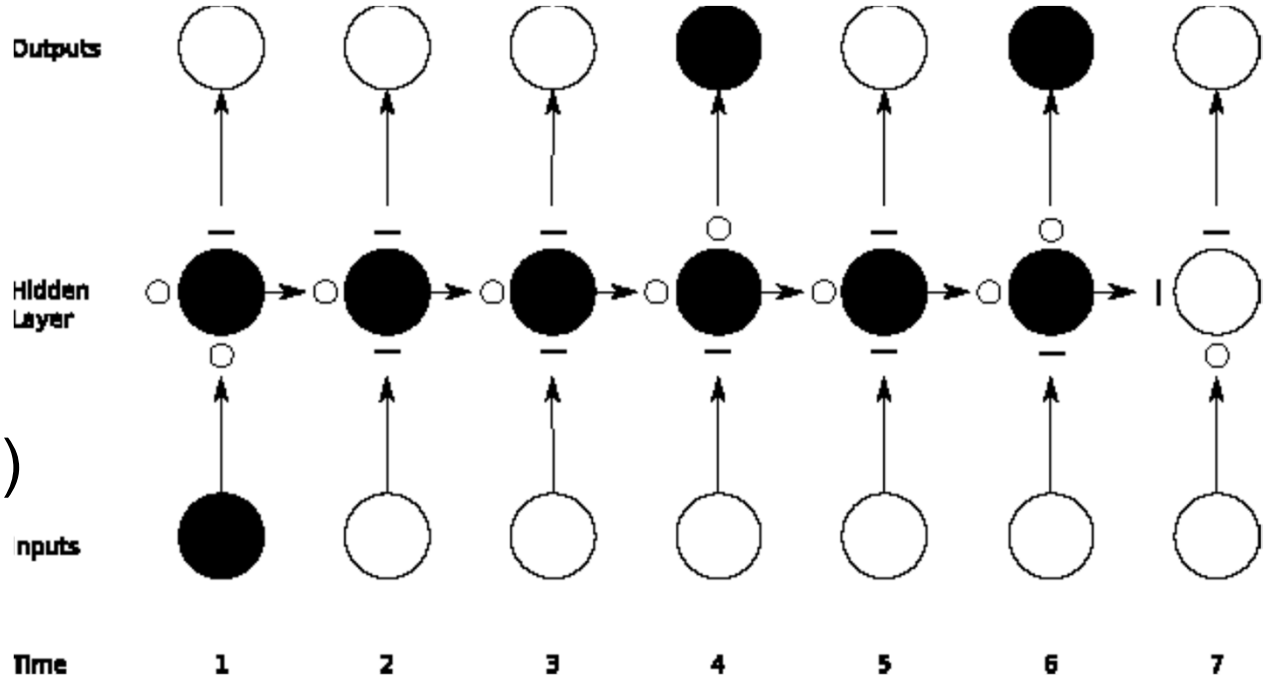
Core Idea: New Memory Cell Instead of Perceptron



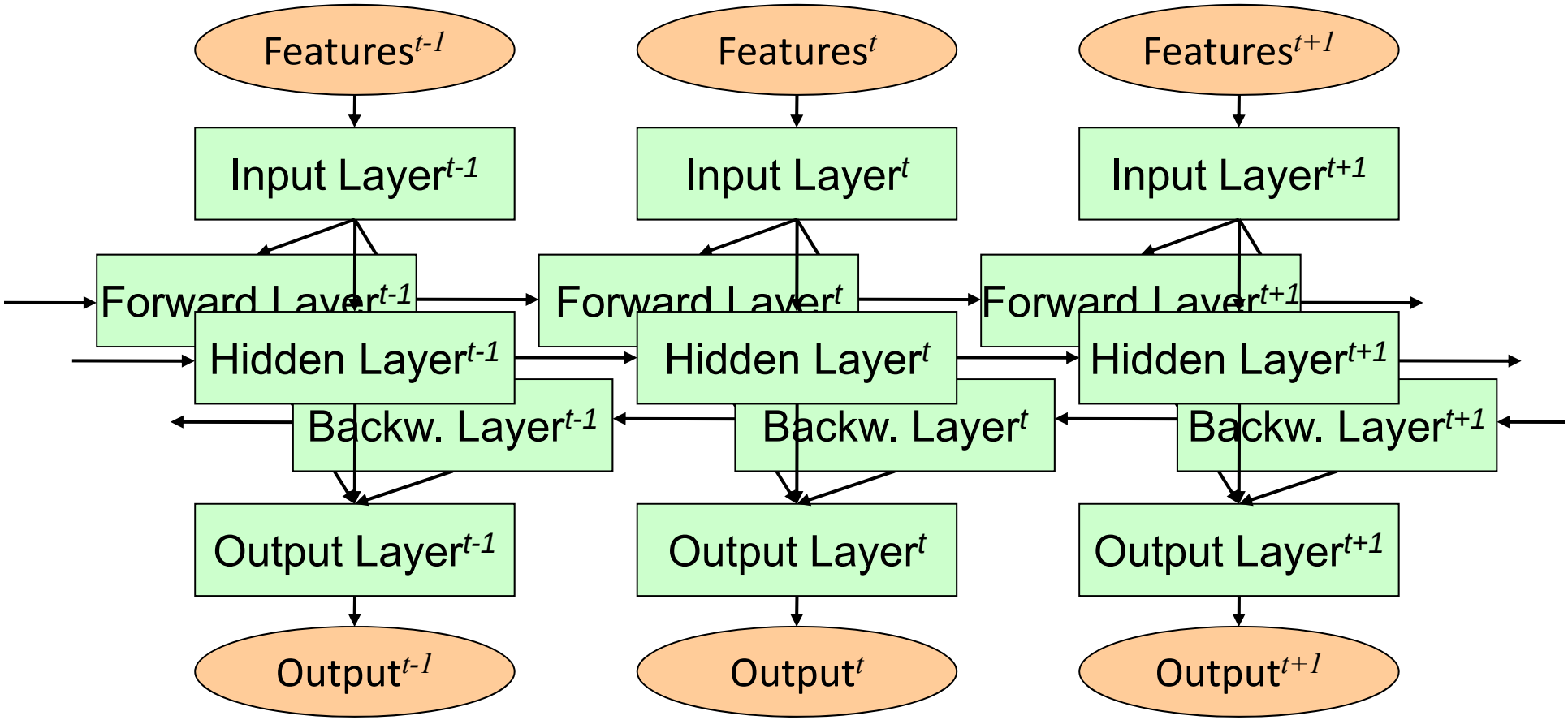
No Vanishing Gradient

$$a_{\omega}^{t} = W_{a,\omega} \cdot X^{t} + W_{h,\omega} \cdot B^{t-1} + W_{c,\omega} S_{c}^{t}$$

- Output Gate
- Output
- Neuron now
  - O : open ($\sigma=1$)
  - | : closed ($\sigma=0$)



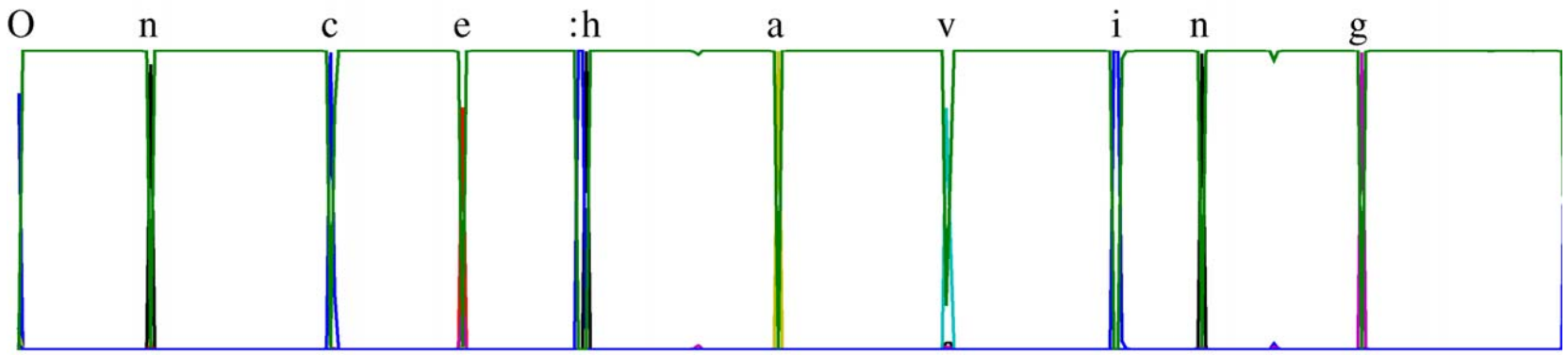
Bidirectional RNN



- Trained with backpropagation through time (forward path trough all time stamps for each hidden layer sequentially)

Connected Temporal Classification

- Additional blank label (b green)
- Allows application to whole sequences
- Output with normalized likelihood for each word

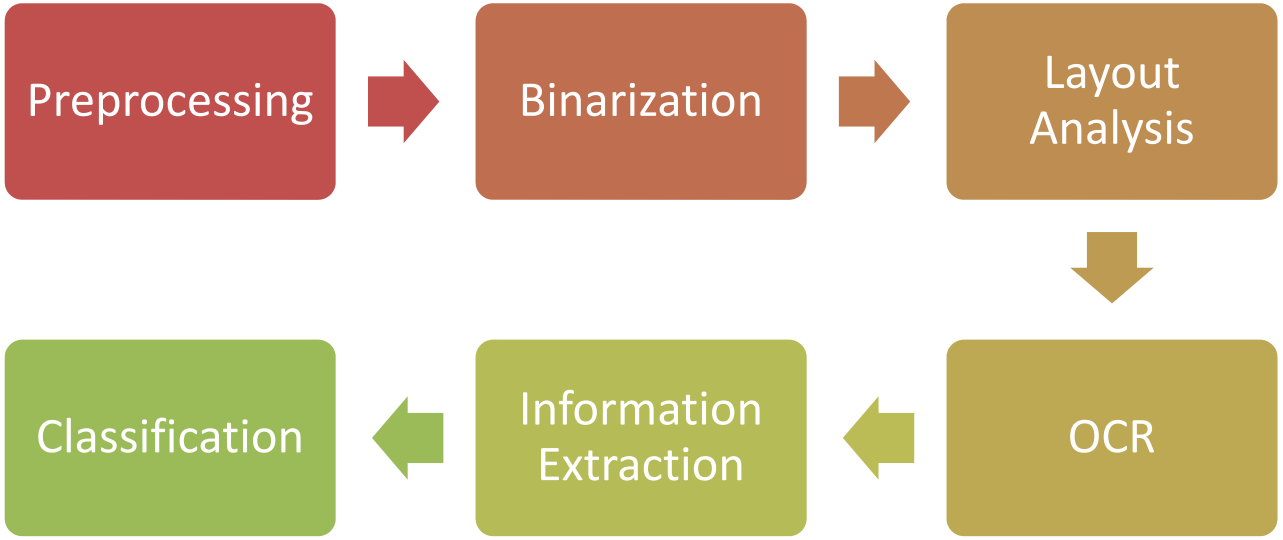


- Training: objective function is smoothed and recalculated after each iteration (details in references)
- Testing: similar to HMM Viterbi-algorithm

Processing Steps of Automatic DIA

Threshold (local, global)
- Sauvola
Classification

Top-down
- XY-cut, histograms
Bottom-up
- Connected components



HMM on features
LSTM with CTC
New: MDLSTM on pixels

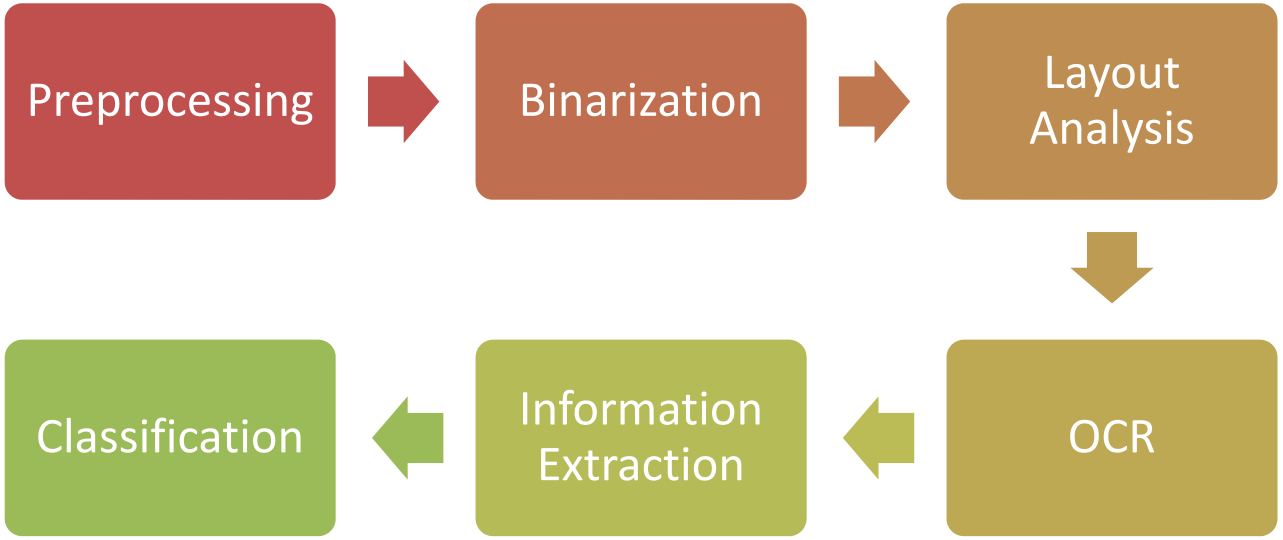
Outline

- Challenge: Why historical Documents?
- State-of-the-Art
- **Recent Trends**
- DIVAServices: Approach Towards Interoperability

Processing Steps of Automatic DIA

Threshold (local, global)
- Sauvola
Classification

Top-down
- XY-cut, histograms
Bottom-up
- Connected components



HMM on features
LSTM with CTC
New: MDLSTM on pixels

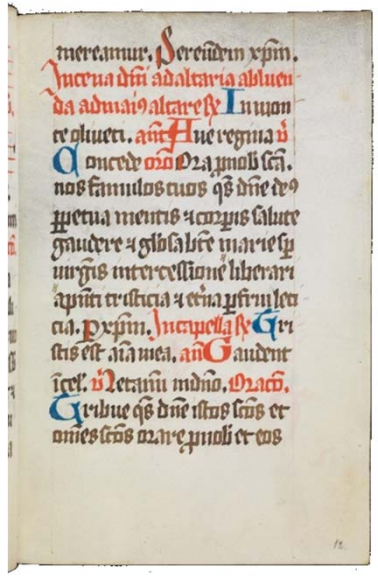
Decolorization vs. Binarization

- Instead of greyscale conversion – use color intensity
  - ^ Text is much better visible

original

greyscale

decolorized

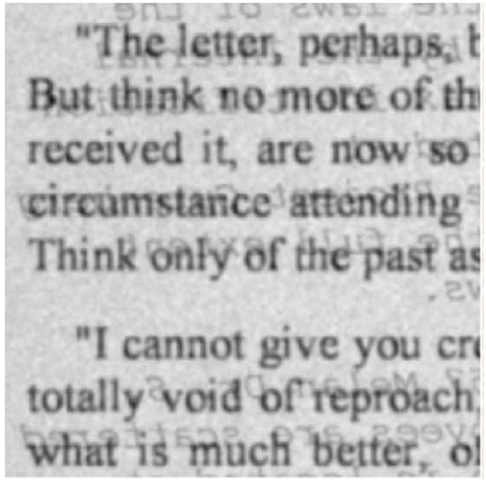


Grundland, Mark, and Neil A. Dodgson. "Decolorize: Fast, contrast enhancing, color to grayscale conversion." *Pattern Recognition* 40.11 (2007): 2891-2896.

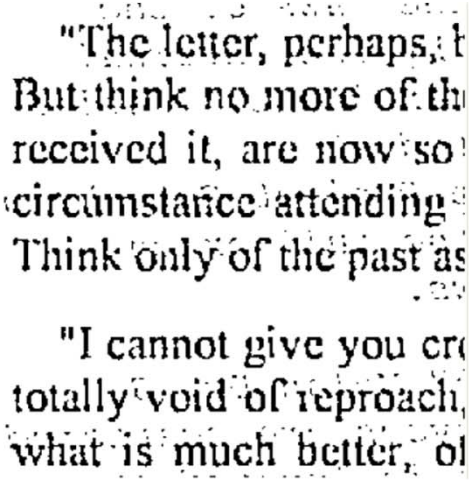
- Promising results on historical documents

Deep Learning for Binarization

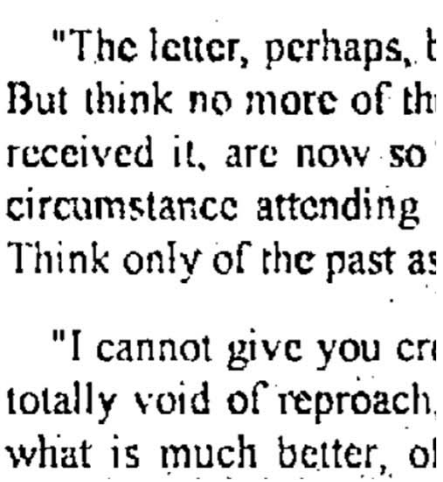
Original



Sauvola



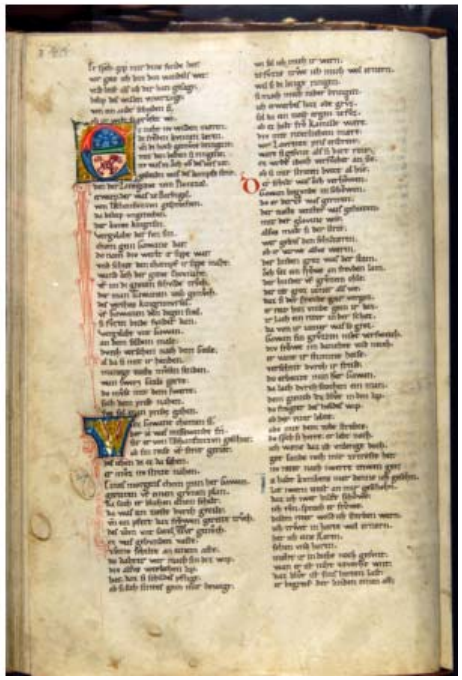
LSTM



OCR from 43% to 73%

Afzal, M. Z., et. al (2015). Document Image Binarization using LSTM : A Sequence Learning Approach. 3rd *Int. Workshop on Historical Document Imaging and Processing* (pp. 79–84).

Layout Analysis Task



Parzival (Cod. 857, page 144, Abbey Library of St. Gall, (PAR23))

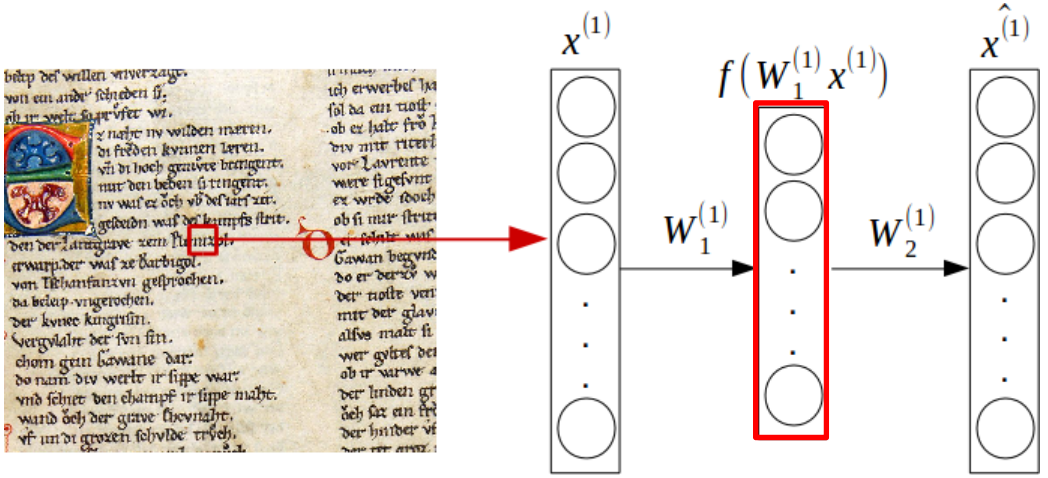
Ground Truth

Label regions (pixels) according to category:

- Text
- Decoration
- Background

Fischer, Andreas, et al. "Ground truth creation for handwriting recognition in historical documents." Proceedings of the 9th IAPR International Workshop on Document Analysis Systems. ACM, 2010.

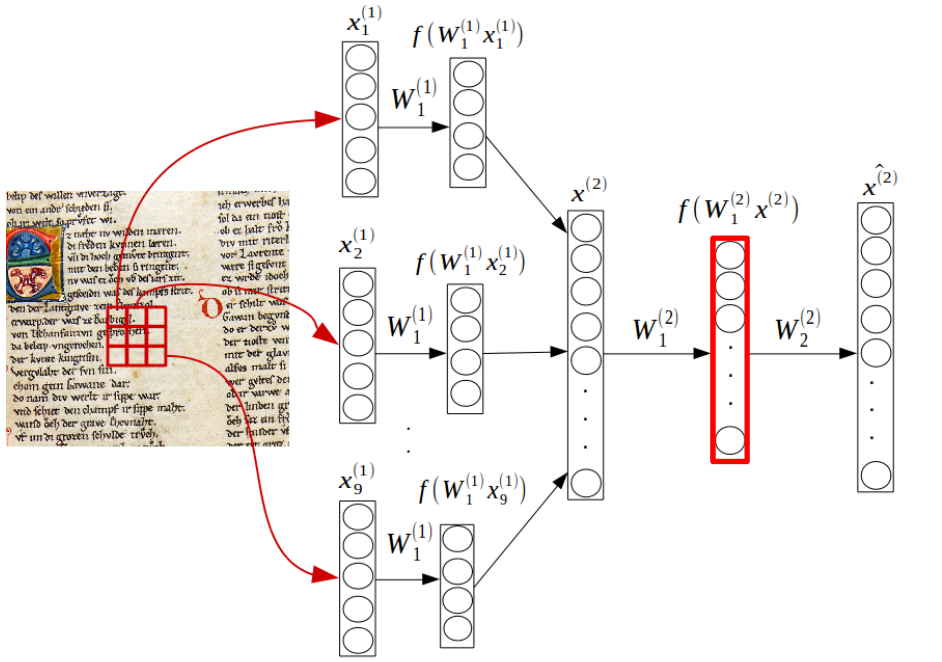
Convolutional Autoencoders (Level 1)



Feature learning from a small patch by autoencoders, Level 1.

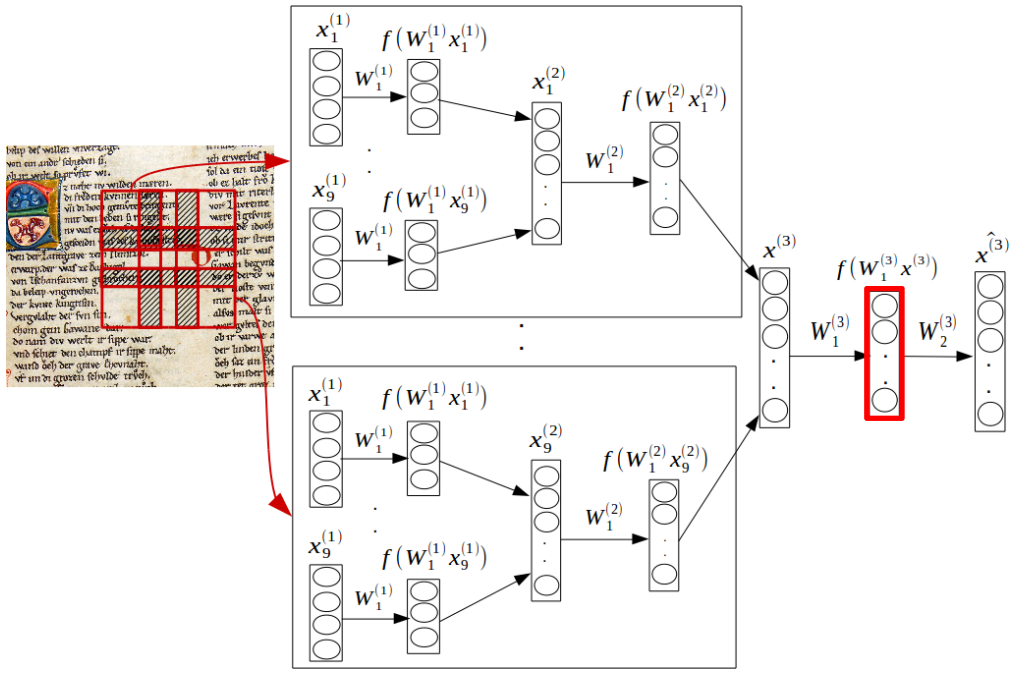
Seuret, M., Ingold, R., Liwicki, M. “A Highly-Adaptable Java Library for Document Analysis with Convolutional Auto-Encoders and Related Architectures” – to appear in ICFHR 2016
Seuret, Mathias, Alberti, Michele, Liwicki, Marcus. “N-light-N : Read The Friendly Manual“, oai:doc.rero.ch:20160809140459-BF; https://github.com/seuretm/N-light-N

Convolutional Autoencoders (Level 2)



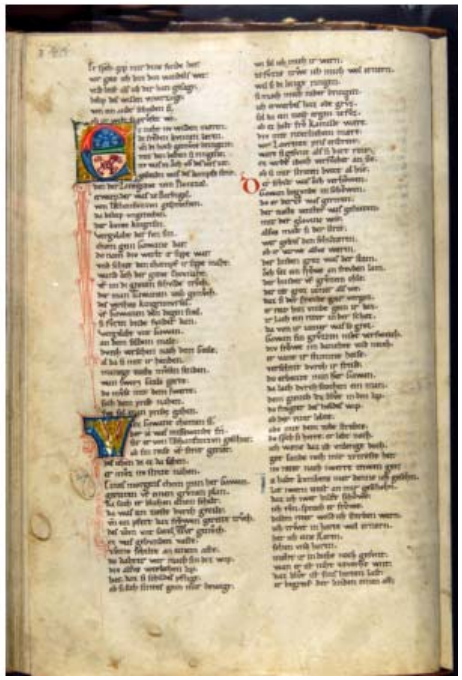
Feature learning from a medium patch by convolutional autoencoders [6], Level 2.

Convolutional Autoencoders (Level 3)



Feature learning from a big patch by convolutional autoencoders [6], Level 3.

SVM Classification Results



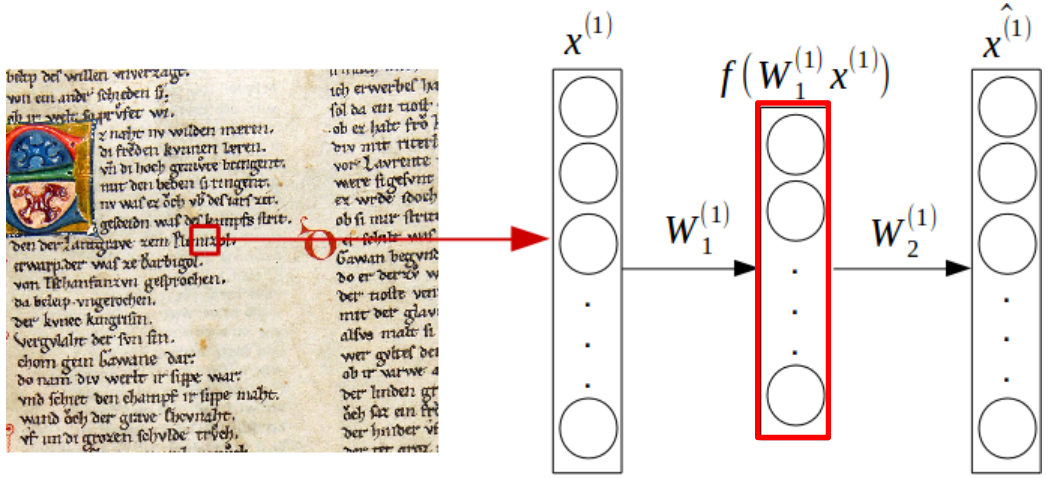
Parzival (Cod. 857, page 144, Abbey Library of St. Gall, (PAR23))

Ground Truth

Segmentation Result

Error (5%)

Understanding Auto-Encoder Features

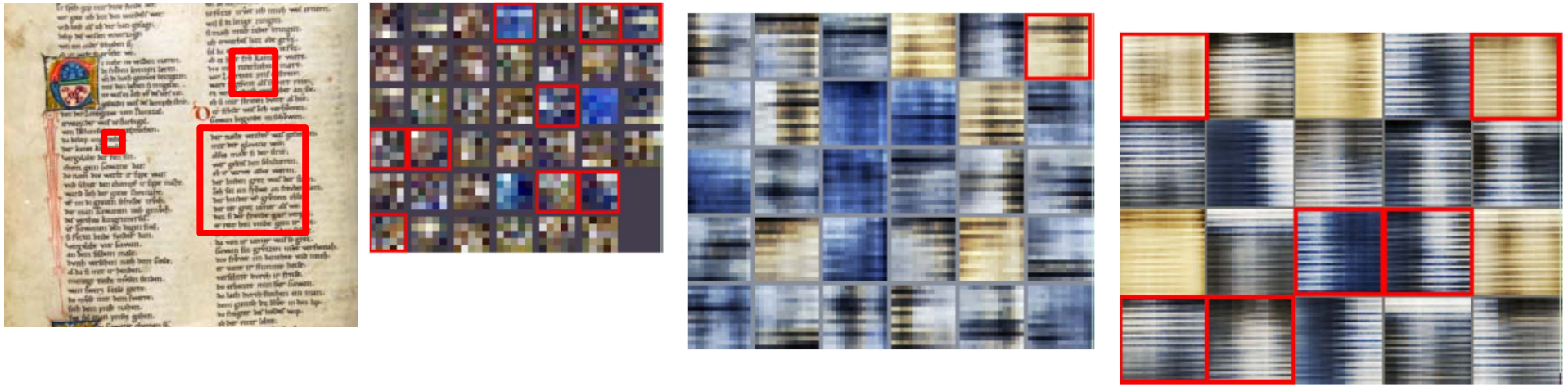


Feature learning from a small patch by autoencoders, Level 1.

Seuret, M., Ingold, R., Liwicki, M. “A Highly-Adaptable Java Library for Document Analysis with Convolutional Auto-Encoders and Related Architectures” – to appear in ICFHR 2016
Seuret, Mathias, Alberti, Michele, Liwicki, Marcus. “N-light-N : Read The Friendly Manual“, oai:doc.rero.ch:20160809140459-BF; https://github.com/seuretm/N-light-N

Visualizing CAE features

- Feature selection strategy applied

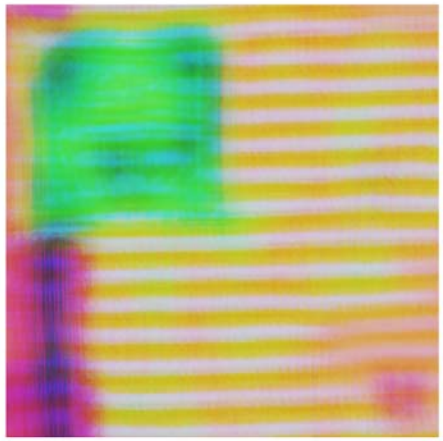


Wei, Hao, Seuret, M., Chen, K., Fischer, A., Liwicki, M., & Ingold, R. (2015). Selecting Autoencoder Features for Layout Analysis of Historical Documents. In *Third International Workshop on Historical Document Imaging and Processing* (pp. 55–62).

Network Initialization

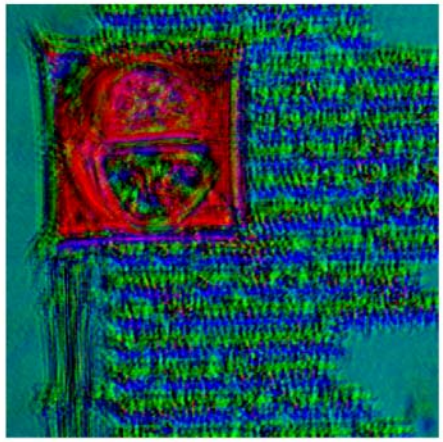
- Transfer learning
  - From AlexNet
  - Fine-tuning
- Recent trend
  - PCA
  - Transformed into encoding layer

(a) CS857 sample



(b) PCA activation

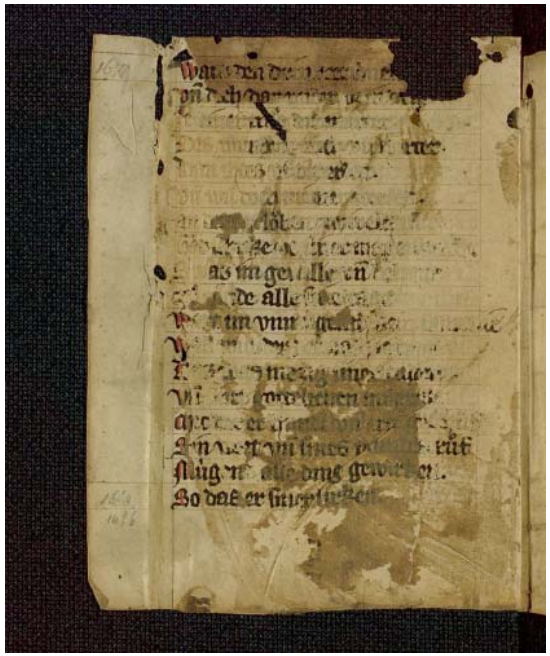
(c) Xavier activation

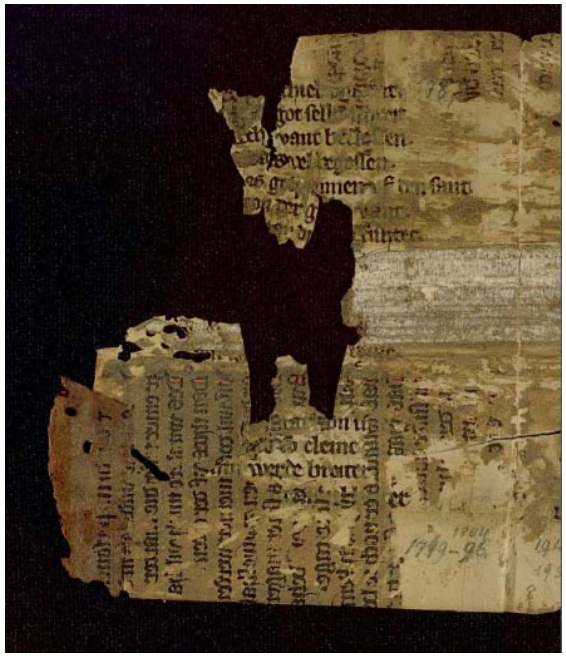


(d) Xavier enhanced

Deep Learning Needs Training Data

- Decorations
- Unique Scripts
- Degradations



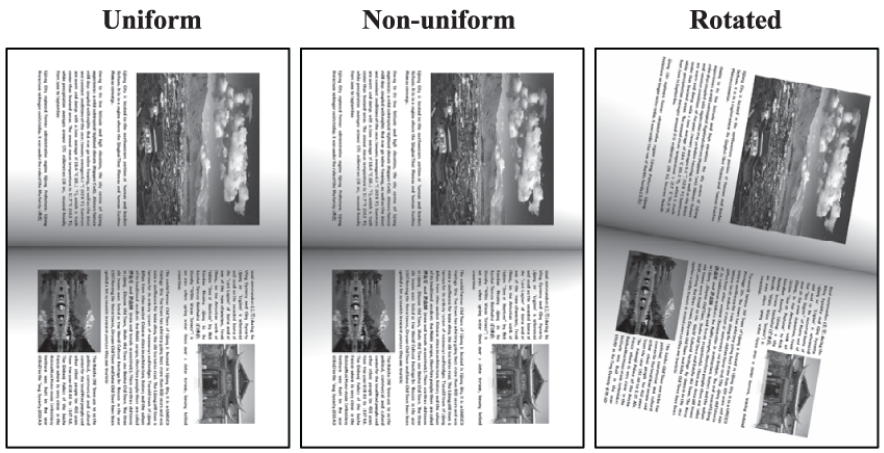


Karlsruhe, BLB, Donaueschingen A III 12

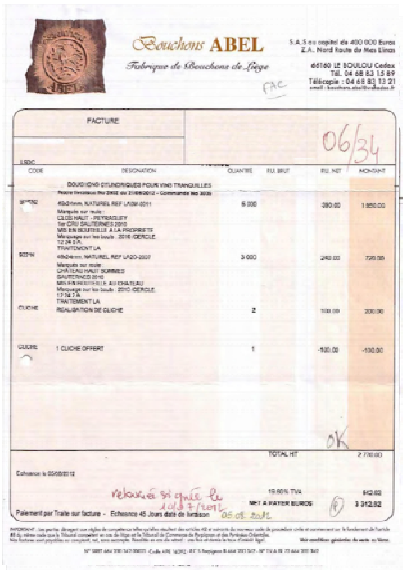
- Idea: generate synthetic training data

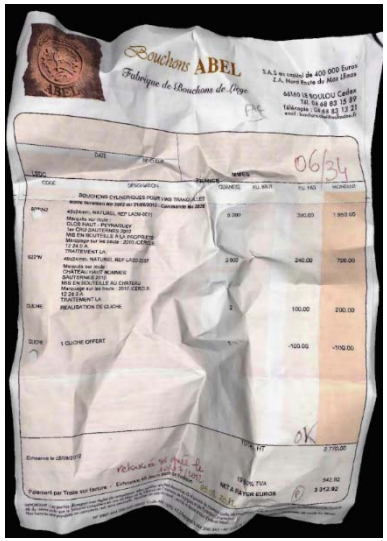
Synthetic Degradations

- Standard 2D methods, novel 3D degradation.



(b) synthetic images of texts and pictures mixed





Kieu, Van Cuong, et al. "Semi-synthetic document image generation using texture mapping on scanned 3D document shapes." 2013 12th International Conference on Document Analysis and Recognition. IEEE, 2013.

Synthetic Document Creator

Synthetic Degradations

- Working in the gradient domain
  - Learn noise from sample documents
  - Apply it on cleaner data

(a) Original document (b) Synthetic degradations

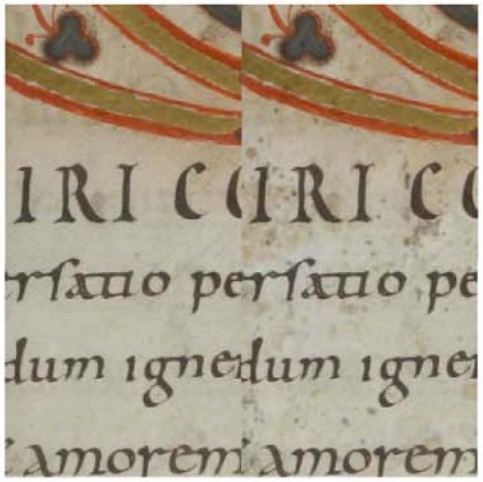
Seuret, M., Chen, K., Eichenberger, N., Liwicki, M., & Ingold, R. (2015). Gradient-Domain Degradations for Improving Historical Documents Images Layout Analysis. In 13th International Conference on Document Analysis and Recognition (pp. 1006–1010). IEEE.

Impact on DL-Based Layout Analysis

- A single labelled page used for training (→+20

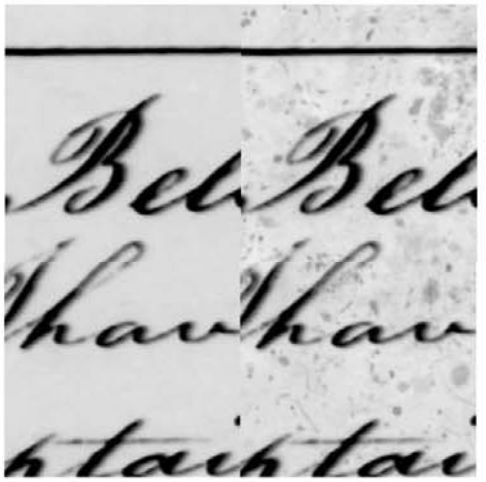
(a) Parzival

85.46 % → 92.47 %



(b) St-Gallen

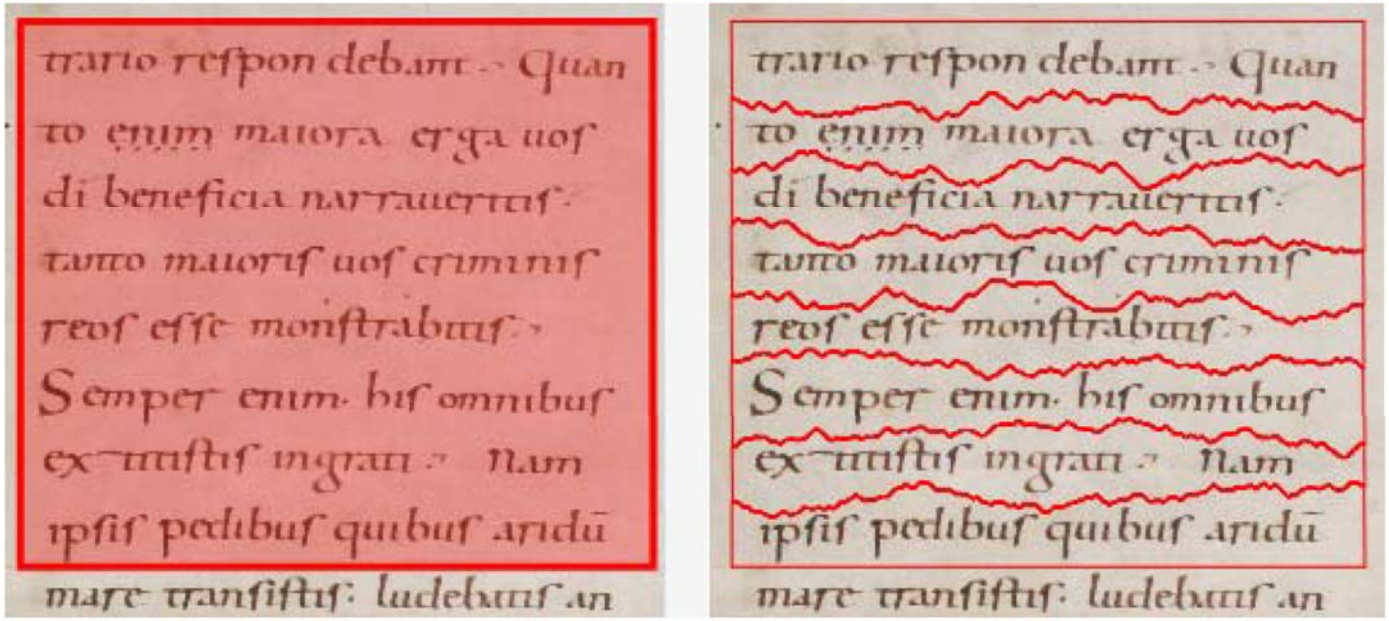
96.05 % → 94.07 %



(c) George Washington

83.71 % → 81.71 %

Text Line Segmentation – Seam Carving

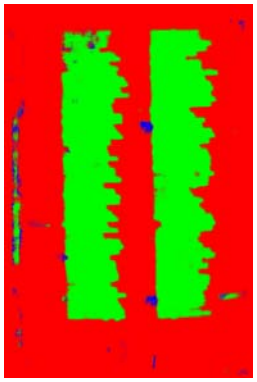


N. Arvanitopoulos Darginis and S. Süsstrunk, “Seam Carving for Text Line Extraction on Color and Grayscale Historical Manuscripts,” in 14th International Conference on Frontiers in Handwriting Recognition Conference on Frontiers in Handwriting Recognition (ICFHR), 2014.

Deep Learning for Text Line Segmentation

Input Image

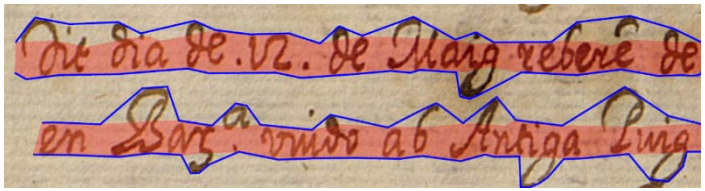
Layout Analysis (CNN)



Main Body Area (CNN)

Segmentation (watershed algorithm)

Finally: post-processing (for ascenders/descenders)



Pastor-pellicer, J., Afzal, M. Z., & Liwicki, M. (2016). Complete Text Line Extraction with Convolutional Neural Networks and Watershed Transform. In 12th IAPR Workshop on Document Analysis Systems (pp. 30–35).

OCR Error Correction using LSTM

- Given several OCR (Tesseract and ocropy), post process the errors as much as possible

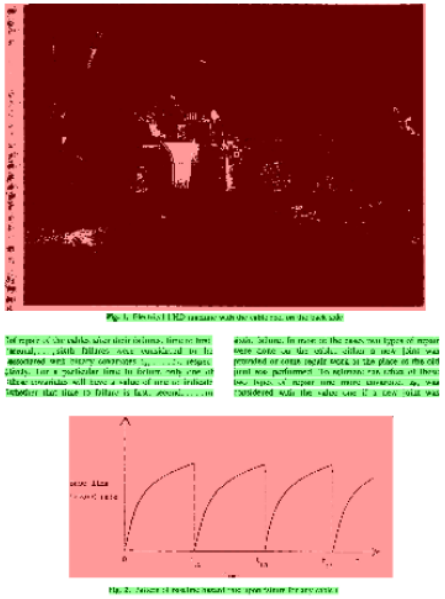
Dataset	OCR1	OCR2	ISRI[9]	Pairwise[5]	Line-Page	LSTM
English	2.0%	1.56%	1.45%	1.32%	1.26%	0.40%
Fraktur	2.7%	2.5%	2.41%	2.36%	2.31%	0.39%

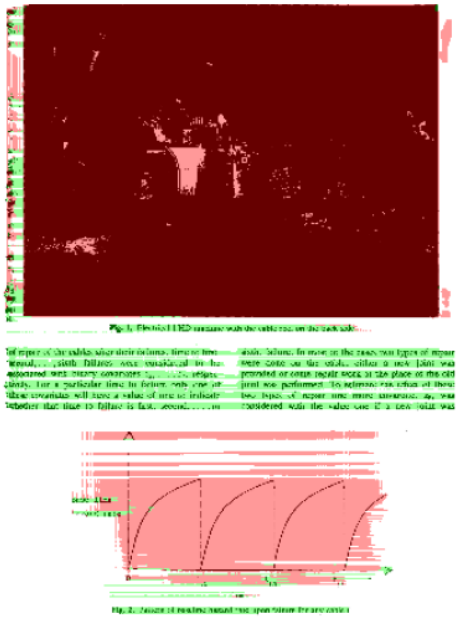
- Results on public English Dataset and Fraktur: Fontane “Wanderung durch die Mark Brandenburg” (1862-1889)

Azawi, M. Al, Liwicki, M., & Breuel, T. M. (2015). Combination of Multiple Aligned Recognition Outputs using WFST and LSTM. In *13th International Conference on Document Analysis and Recognition* (pp. 31–35). IEEE.

LSTM for Layout Analysis

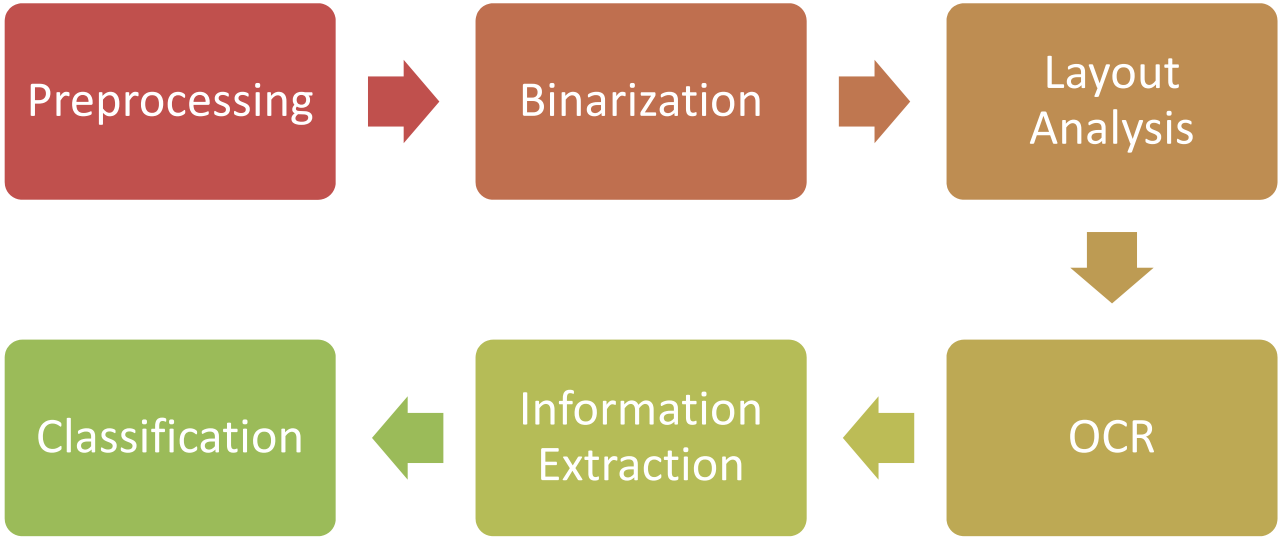
- Example on modern prints:





➔ What is Ground Truth???

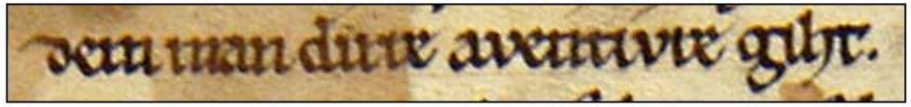
We Should Reconsider Stepwise Approach



But Document Processing can comprise much more …

Early Example: Handwriting Recognition

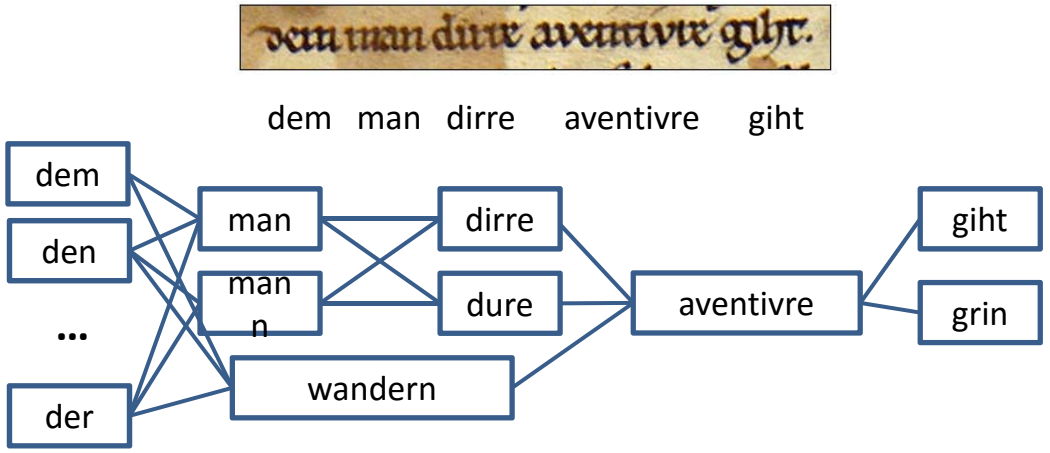
- Sayre's paradox (1973): It is impossible to recognize a handwritten word without recognizing the characters first **& vice versa**.
- State-of-the-Art: binarization- & segmentation-free whole line recognition



dem man dirre aventivre giht

Information Extraction from HWR-Results

- Using recognition lattices as intermediate step

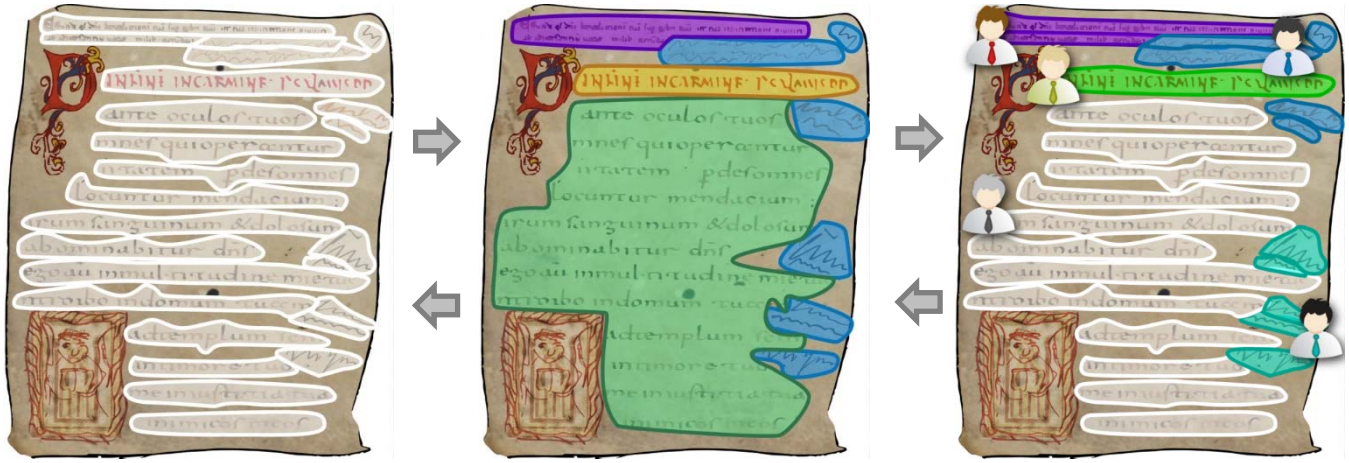


- Information extraction rate the same as on pure ASCII text

Liwicki, M., Ebert, S., & Dengel, A. (2014). Bridging the Gap Between Handwriting Recognition and Knowledge Management. *Pattern Recognition Letters*, *35*, 204–213

Fictional Example – Sequential Approach

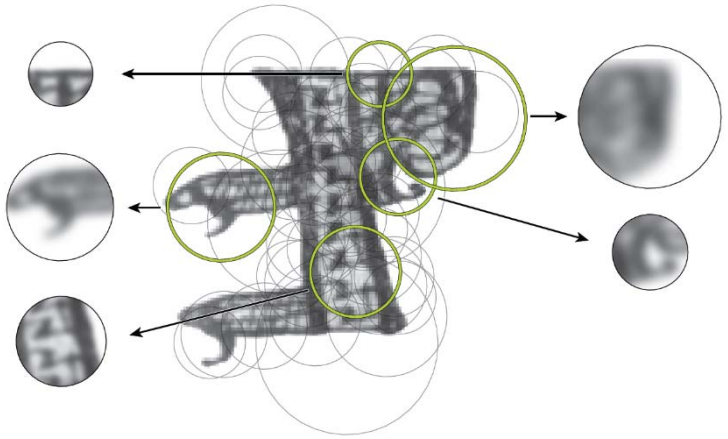
Knowledge Transfer in Iterative Approach



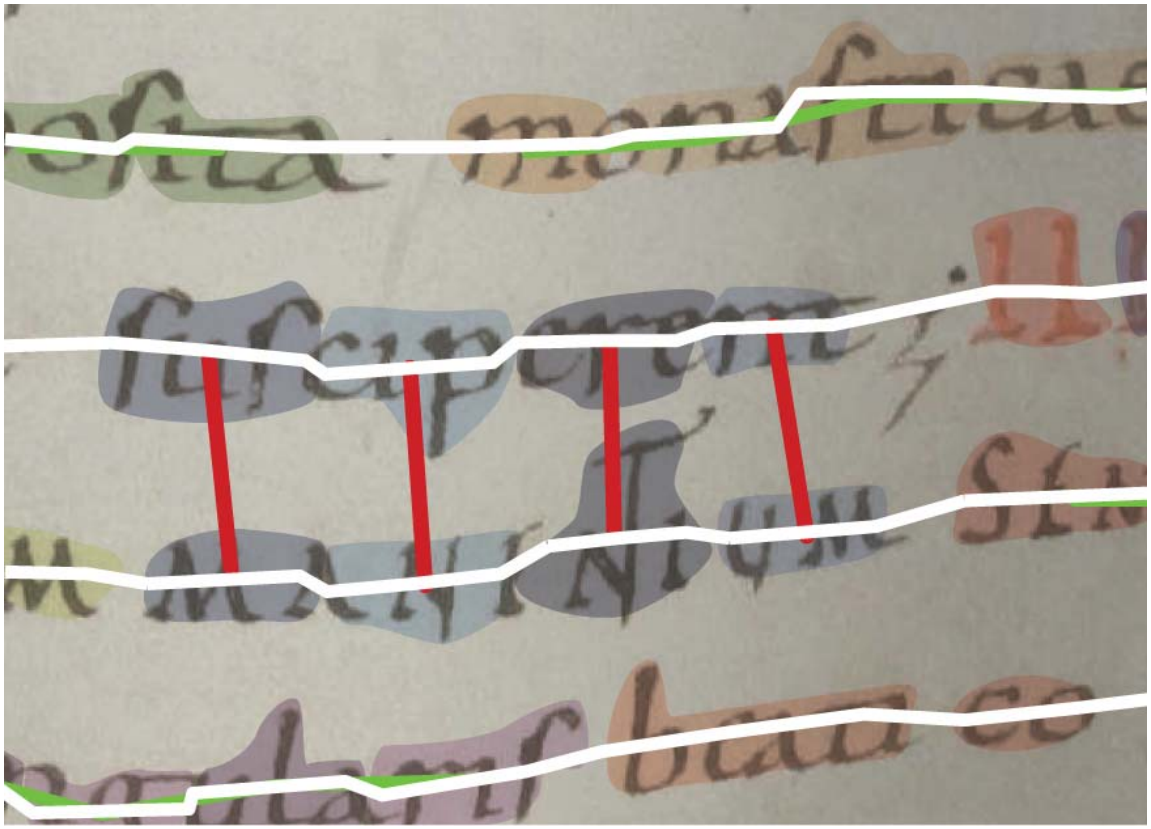
- Script discrimination
  - Prevent grouping different scripts
  - Facilitate scribe identification
- Text segmentation: statistical information about handwriting

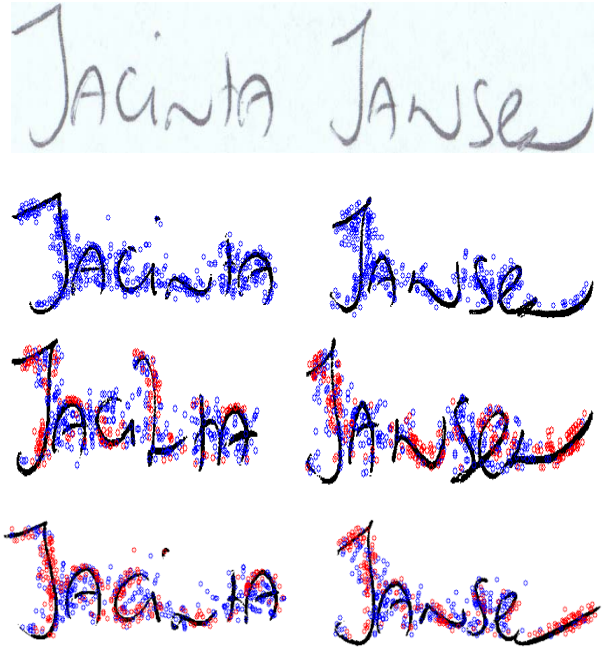
Local Features (Interest Points)

- Sparse (feature) space
- Structures dissimilar to their adjacent neighborhood
- Capable of capturing script parts
  e.g. endpoints, crossings, loops



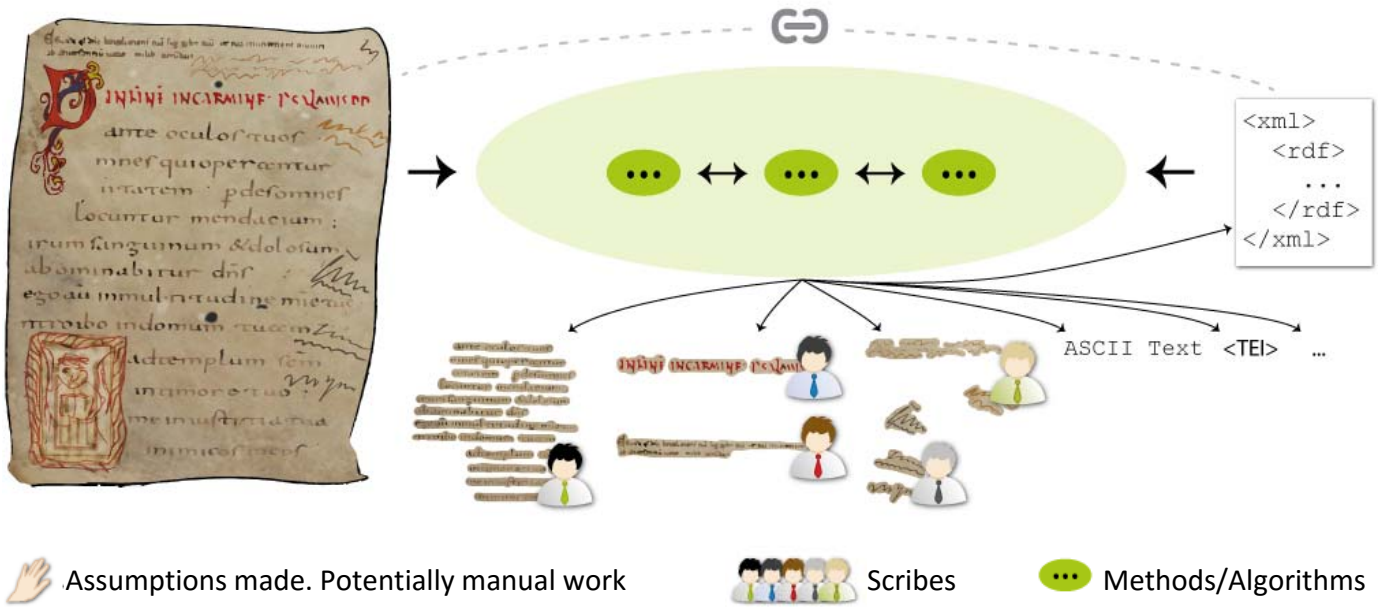
The Power of Local Features





HisDoc 2.0 in Fribourg

- **Our proposal**: Holistic approach for *Text Segmentation, Script Analysis, and Scribe Identification*



Current Trend: End-to-End

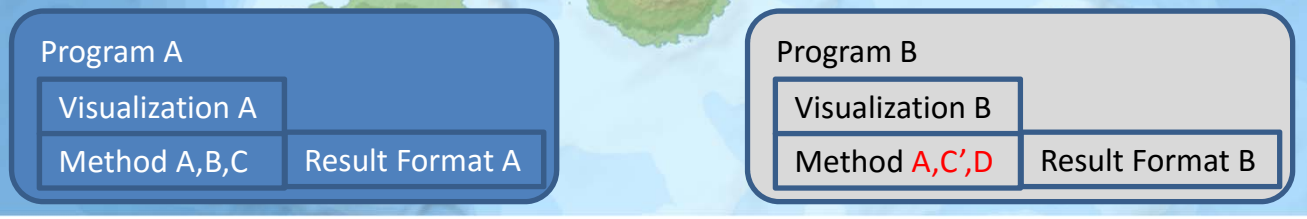
- Preprocessing-free Layout Analysis
- End-to-End OCR
  - MDLSTM over many text lines (RWTH)
- Logical Structure Recognition
  - CNN for tables
- Keyword Spotting (Fink's presentation)

Outline

- **Challenge: Why historical Documents?**
- State-of-the-Art
- Recent Trends
- DIVAServices: Approach Towards Interoperability

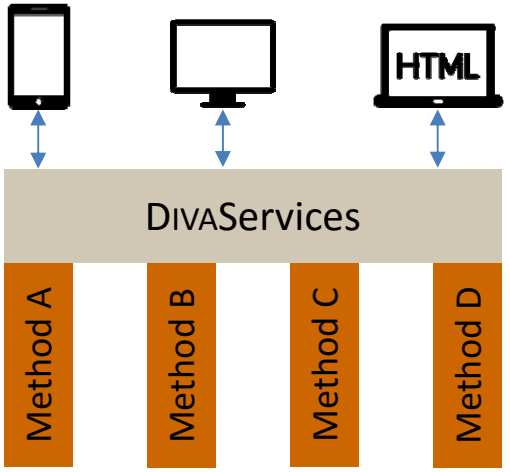
Living on Islands Can be Lonely

- Good tools are available
  - Methods coupled to the tool
- Built to solve a specific problem
- Hard to maintain
- Almost no reusability
- (Which islands did you use)



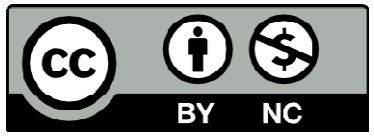
We are Building a Strong Foundation

- Accessible over the internet
- Hosted on our infrastructure
  - No computation on the client
- Defined input and output format



We are Open Source!

- Service source code
  - GPL v3.0 License
- Image Data
  - Stored on our servers for future research
  - Need to be under Creative Commons
- Methods do not need to be open source
  - But of course it would be great
- Services available
  - We also use docker



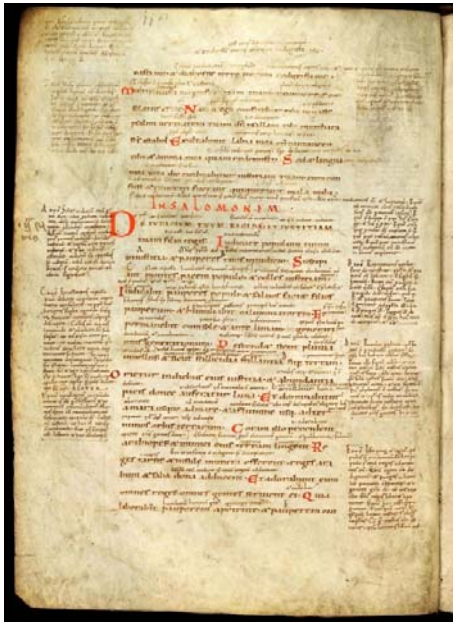
Project Website
http://bit.ly/divaservices

DIVAServices & Spotlight

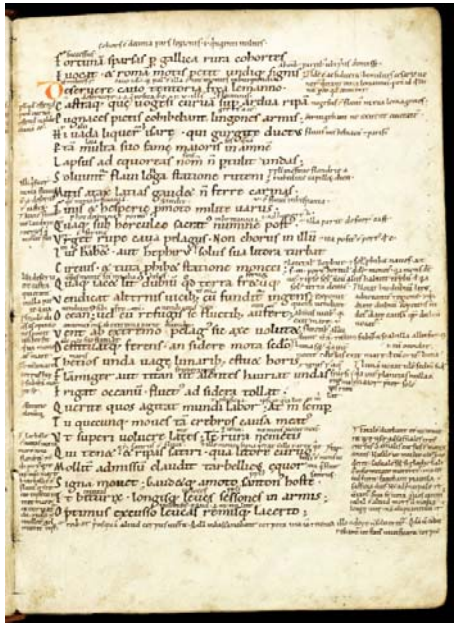
- DAS 2016 Best Paper Award
- Used in diverse tools in Fribourg
- Planned to be integrated in Hamburg (September 2016)

DivaServices-Spotlight
http://divaservices.unifr.ch/spotlight

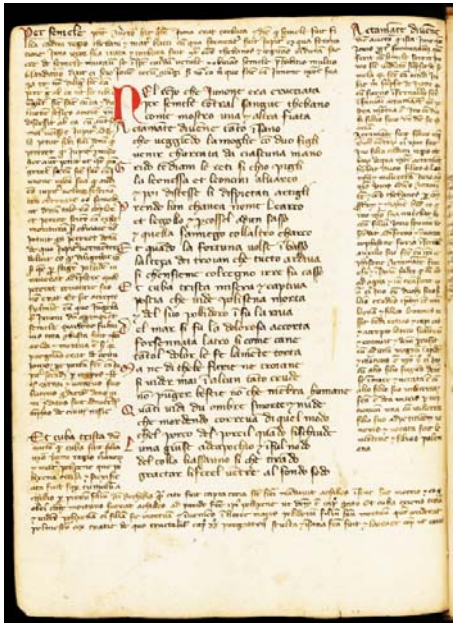
DIVA-HisDB – Challenging HWR



(a)



(b)



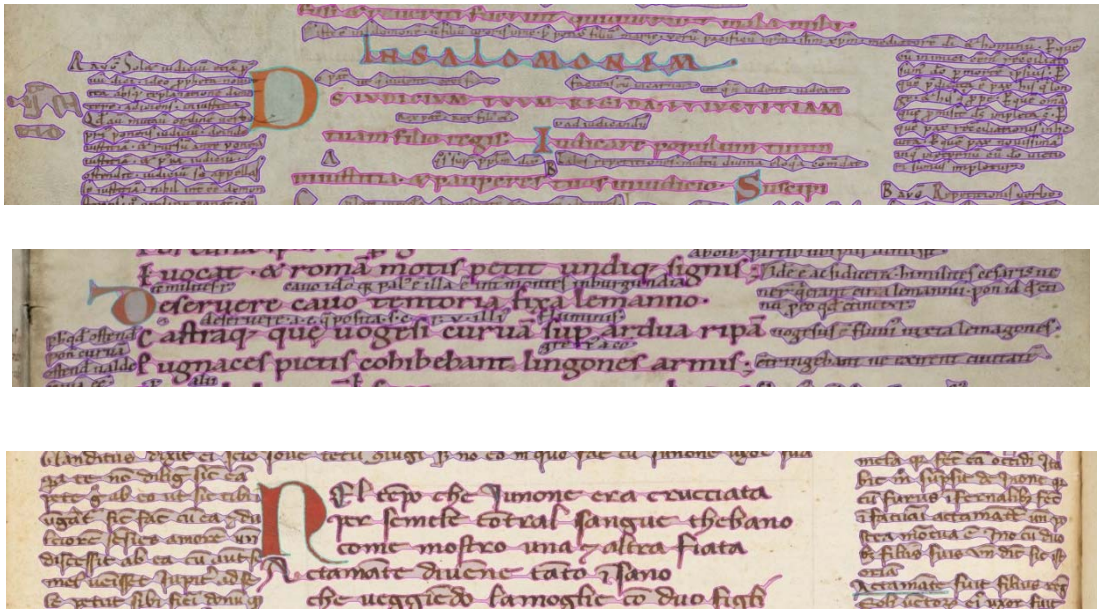
(c)

(a) St. Gallen, Stiftsbibliothek, Cod. Sang. 18, (CSG18), Latin, 10th cent.

(b) St. Gallen, Stiftsbibliothek, Cod. Sang. 863 (CSG863), Latin, 11th cent.

(c) Cologny-Geneve, Fondation Martin Bodmer, Cod. Bodmer 55 (CB55), Italian/Latin glosses, 14th cent.

GroundTruth Accurate & Multi-Labeling

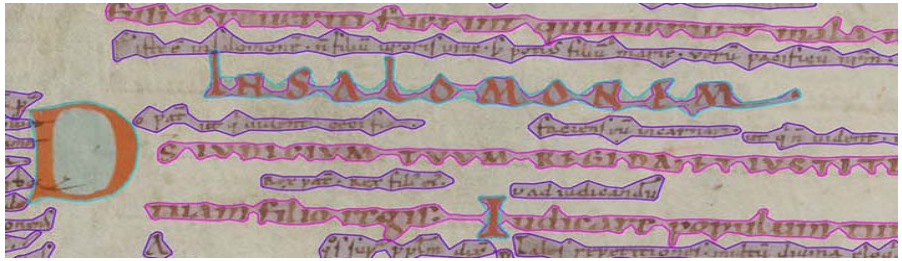


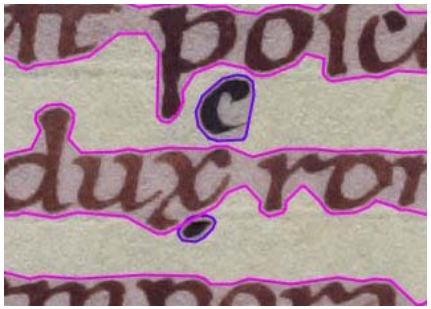
decorations

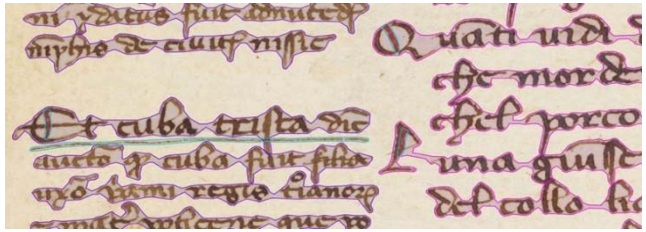
comments

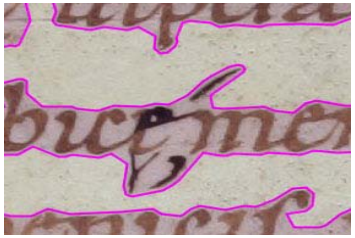
main text body

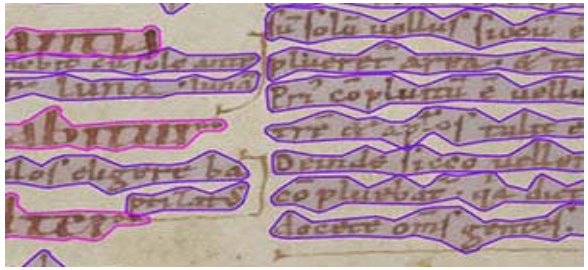
Annotation Examples (& Rules)

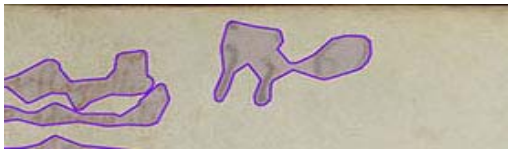










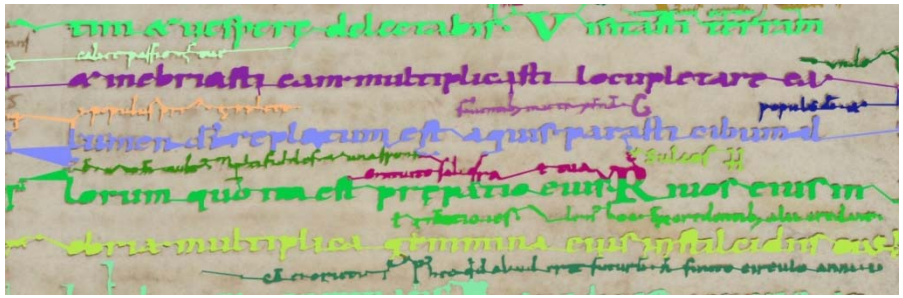


Number of Items per Annotated Category

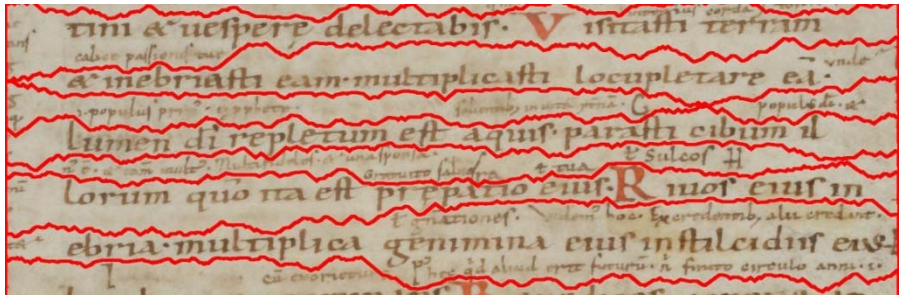
	CSG18	CSG863	CB55	total/MS	(%)	
main text body	1,353	1,538	1,486	4,377	**26.67**	29.41
decorations	672	30	835	1,537	**9.36**	0.49
comments	6,260	1,656	2,584	10,500	**63.97**	70.10
total	8,258	3,224	4,905	16,414		

Text Line Extraction is Difficult

using OCRopus



using seam carving based method



DIVAServices were used

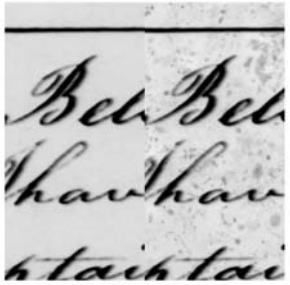
CAE Evaluation Results Using *N-light-N*

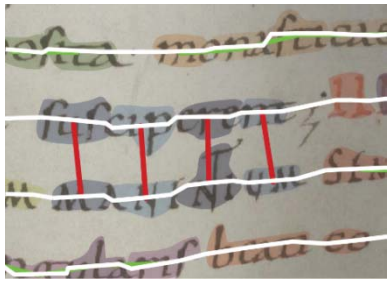
Pixel-Level (in %)	CSG18	CSG863	CB55
background	91.56	93.40	96.05
main text body	94.22	92.08	93.50
decorations	81.58	81.66	93.68
comments	91.35	99.91	90.02
average	**89.93**	**91.76**	**93.31**

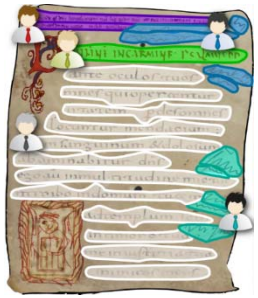
https://diuf.unifr.ch/main/hisdoc/diva-hisdb

DIVA-HisDB

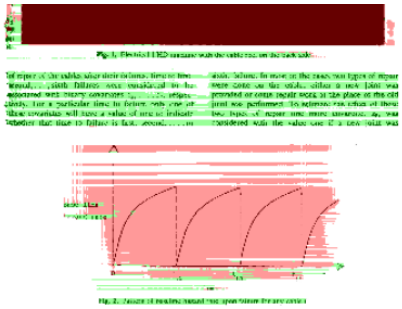
- Release Cycle
  - Every year: new manuscripts for testing
  - Benchmark competition at ICFHR/ICDAR
- License
  - Creative Commons: non commercial + attribution + ShareAlike
- Meta-Data
  - Annotators, dates
- Format: TEI







Historical Document Analysis



Marcus Liwicki
University of Fribourg
University of Kaiserslautern
Insiders Technologies GmbH
marcus.liwicki@unifr.ch