technische universität
dortmund

LAMARR
INSTITUTE FOR MACHINE LEARNING AND ARTIFICIAL INTELLIGENCE

Weakly Supervised Information Extraction from Semi-Structured Document Images

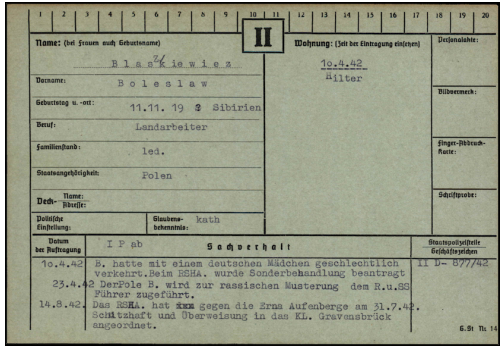
Fabian Wolf[1,2], Oliver Tüselmann[1,2], Christoph Rass[3] and Gernot A. Fink[1,2]

[1]Department of Computer Science, TU Dortmund University
[2]LAMARR, Institute for Machine Learning and Artificial Intelligence
[3]Chair of Modern History and Historical Migration Studies, Osnabrück University

27th February 2024

Semi-Structured Card File Indices

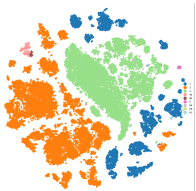
- Massive collections of structured documents
  - Gestapo card file indices
  - Immigration card file indices
  - 1950 census
- Highly valuable mass data for the humanities and social sciences
- Challenges:
  - Degradation
  - Handwritten and machine printed text
  - No annotated data available!



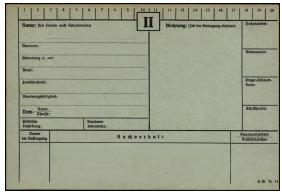
Example the Gestapo dataset from the Lower Saxony State Archive (NLA OS, Rep 439, No. 3227).

Overview

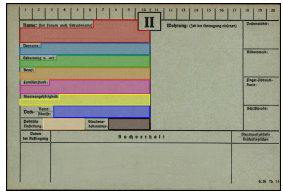
1. Template Extraction



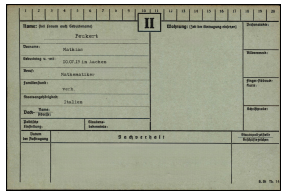
2. Template Cleaning



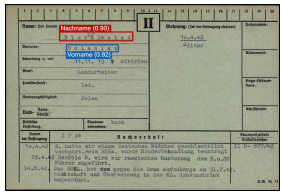
3. Template Annotation



4. Synthetic Dataset Creator



5. Information Detection

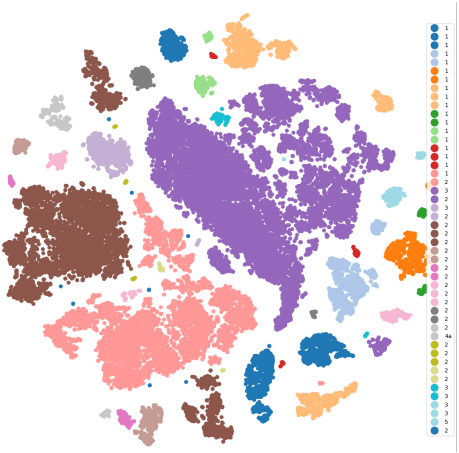


6. Text Recognition

```
fields:[
  {
    First Name: Blaskiewiez,
    Surname: Boleslaw,
    Date of birth: 11.11.19,
    Occupation: Landarbeiter,
    Nationality: Poland
  },
]
```

1. Template Extraction

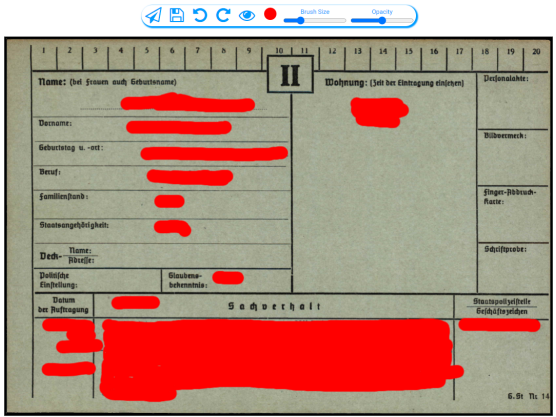
- Limited, unknown number of templates
- Semi-automatic template discovery
  1. Extract vectorial feature representation
  2. Clustering (DBSCAN [1])
  3. Manual inspection of cluster centroids
- Gestapo file indices:
  - 9 different form types
  - Accuracy 99.7%



TSNE visualization of form templates

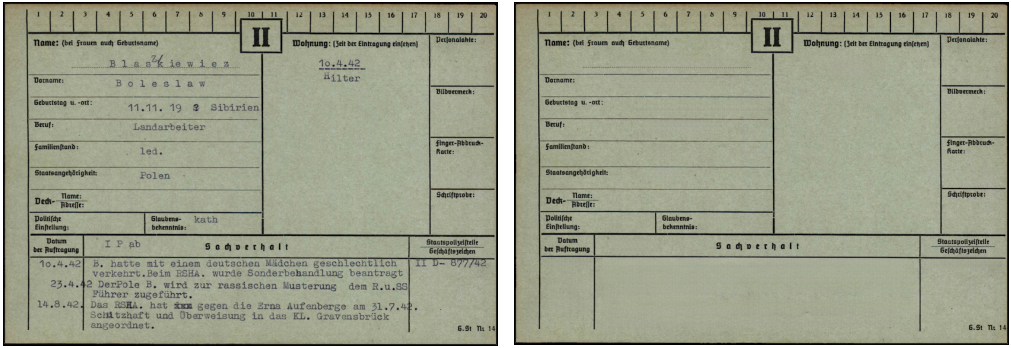
[1] M. Ester, H.-P. Kriegel, J. Sander, X. Xu et al.: *A density-based algorithm for discovering clusters in large spatial databases with noise*. In: *KDD*, Bd. 96, S. 226–231, 1996

2. Template Cleaning



[1] R. Suvorov, E. Logacheva, A. Mashikhin, A. Remizova, A. Ashukha, A. Silvestrov, N. Kong, H. Goka, K. Park und V. Lempitsky: *Resolution-robust large mask inpainting with fourier convolutions*. In: *Winter Conf. on Applications of Computer Vision*, S. 2149–2159, 2022

2. Template Cleaning

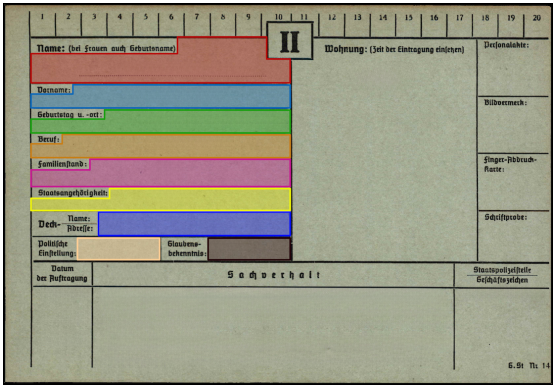


Cleaning

[1] R. Suvorov, E. Logacheva, A. Mashikhin, A. Remizova, A. Ashukha, A. Silvestrov, N. Kong, H. Goka, K. Park und V. Lempitsky: *Resolution-robust large mask inpainting with fourier convolutions*. In: *Winter Conf. on Applications of Computer Vision*, S. 2149–2159, 2022

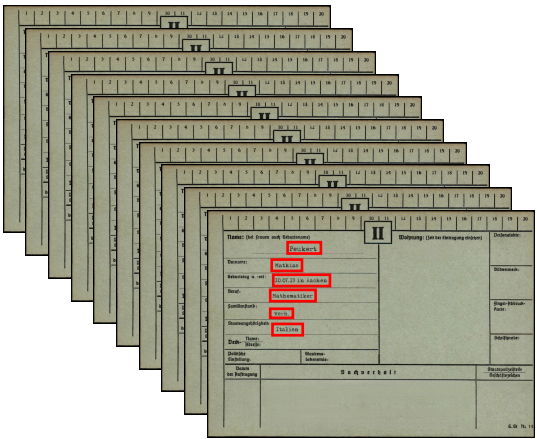
3. Template Annotation

- Manual creation of bounding box annotations for each template
- **No** annotation of individual index cards!
- Definition of field properties



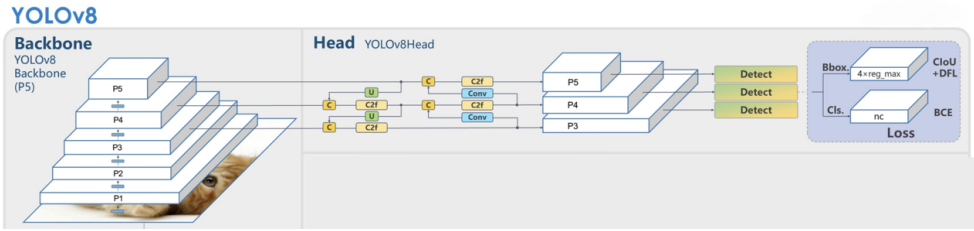
4. Synthetic Dataset Creation

- Synthesize unlimited document images
- Font based text synthesis
- Control over content and style
- Bounding Box and transcription available



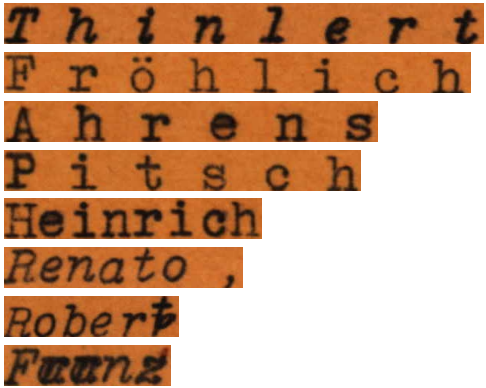
5. Object Detection Model

- Text detection with object detection network YOLOv8 [1]
- Model only trained on synthetically created document images



[1] G. Jocher, A. Chaurasia und J. Qiu: *Ultralytics YOLO*, 2023

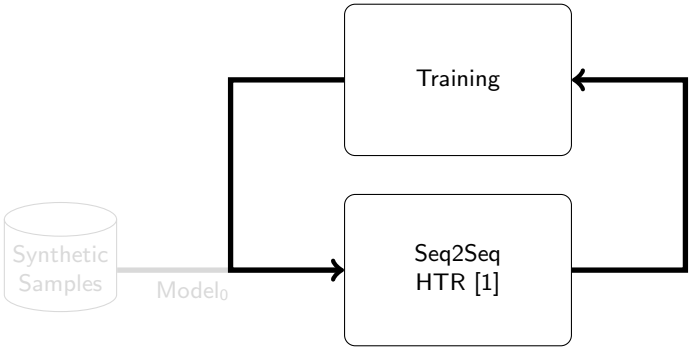
6. Text Recognition



Prediction	Annotation
Thinleft	Thinlert
Frohlich	Fröhlich
Amrens	Ahrens
Pitsch	Pitsch
Heinrich	Heinrich
Renato,	Renato
Rober#	Robert
Foods	Franz

[1] M. Li, T. Lv, J. Chen, L. Cui, Y. Lu, D. A. F. Florêncio, C. Zhang, Z. Li und F. Wei: *TrOCR: Transformer-Based Optical Character Recognition with Pre-trained Models*. In: *Conf. on Artificial Intelligence, AAAI*, S. 13094–13102, 2023

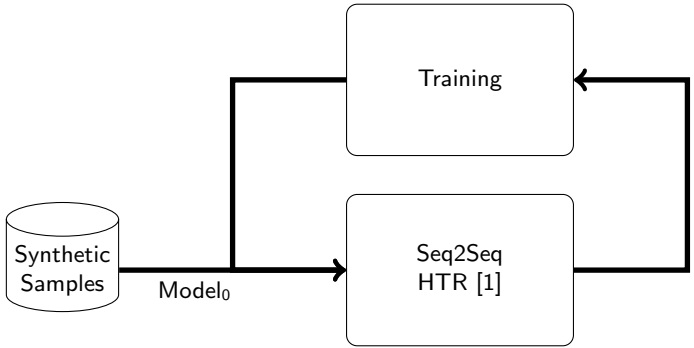
Self-Training



[1] L. Kang, J. I. Toledo, P. Riba, M. Villegas, A. Fornés und M. Rusiñol: *Convolve, Attend and Spell: An Attention-based Sequence-to-Sequence Model for Handwritten Word Recognition*. In: *German Conference on Pattern Recognition*, Bd. 11269, S. 459–472, Stuttgart, Germany, 2018

[2] F. Wolf und G. A. Fink: *Self-Training of Handwritten Word Recognition for Synthetic-to-Real Adaptation*. In: *Proc. Int. Conf. on Pattern Recognition*, S. 3885–3892, Montreal, Canada, 2022. IEEE

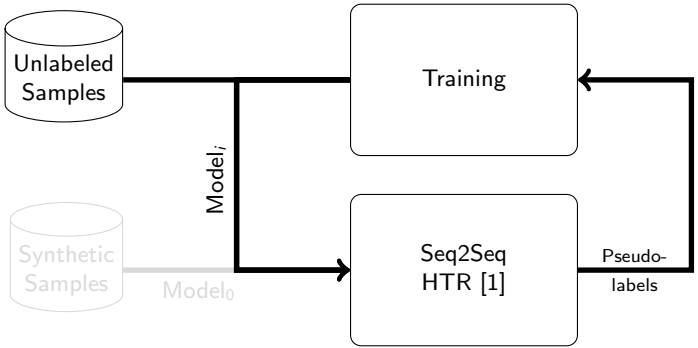
Self-Training



[1] L. Kang, J. I. Toledo, P. Riba, M. Villegas, A. Fornés und M. Rusiñol: *Convolve, Attend and Spell: An Attention-based Sequence-to-Sequence Model for Handwritten Word Recognition*. In: *German Conference on Pattern Recognition*, Bd. 11269, S. 459–472, Stuttgart, Germany, 2018

[2] F. Wolf und G. A. Fink: *Self-Training of Handwritten Word Recognition for Synthetic-to-Real Adaptation*. In: *Proc. Int. Conf. on Pattern Recognition*, S. 3885–3892, Montreal, Canada, 2022. IEEE

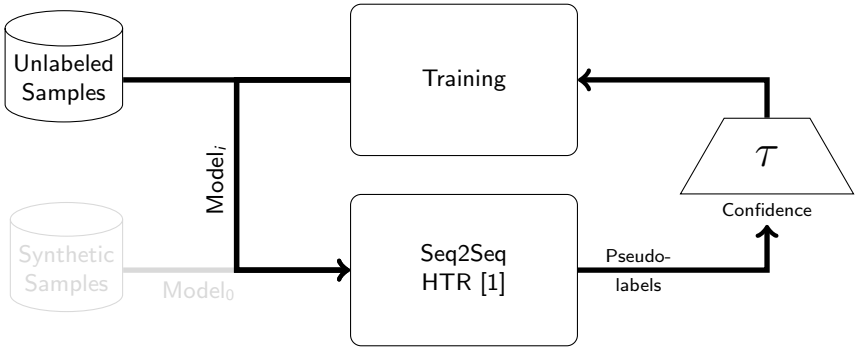
Self-Training



[1] L. Kang, J. I. Toledo, P. Riba, M. Villegas, A. Fornés und M. Rusiñol: *Convolve, Attend and Spell: An Attention-based Sequence-to-Sequence Model for Handwritten Word Recognition*. In: *German Conference on Pattern Recognition*, Bd. 11269, S. 459–472, Stuttgart, Germany, 2018

[2] F. Wolf und G. A. Fink: *Self-Training of Handwritten Word Recognition for Synthetic-to-Real Adaptation*. In: *Proc. Int. Conf. on Pattern Recognition*, S. 3885–3892, Montreal, Canada, 2022. IEEE

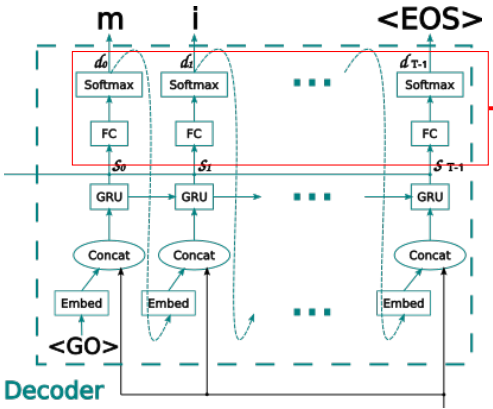
Self-Training



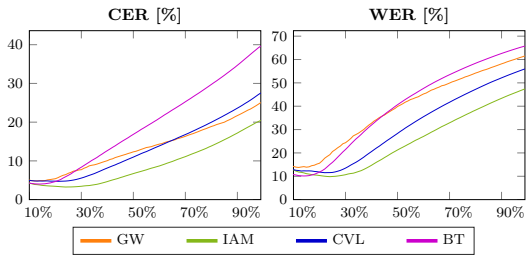
[1] L. Kang, J. I. Toledo, P. Riba, M. Villegas, A. Fornés und M. Rusiñol: *Convolve, Attend and Spell: An Attention-based Sequence-to-Sequence Model for Handwritten Word Recognition*. In: *German Conference on Pattern Recognition*, Bd. 11269, S. 459–472, Stuttgart, Germany, 2018

[2] F. Wolf und G. A. Fink: *Self-Training of Handwritten Word Recognition for Synthetic-to-Real Adaptation*. In: *Proc. Int. Conf. on Pattern Recognition*, S. 3885–3892, Montreal, Canada, 2022. IEEE

Confidence Estimation



$$c = \frac{1}{T} \sum_0^T \max(d_t)$$



[1] F. Wolf und G. A. Fink: *Self-Training of Handwritten Word Recognition for Synthetic-to-Real Adaptation*. In: *Proc. Int. Conf. on Pattern Recognition*, S. 3885–3892, Montreal, Canada, 2022. IEEE

Self-Training Benchmark Results

Selection	GW		IAM		CVL		BT	
	CER	WER	CER	WER	CER	WER	CER	WER
Synthetic	17.8	45.0	21.1	48.4	28.2	58.1	23.5	52.3
Self-Train (ST)	14.2	35.1	13.0	35.9	10.5	31.1	14.7	38.5
ST (Confidence)	**12.6**	**31.2**	**10.2**	**27.7**	**7.5**	**24.6**	**10.2**	**27.7**

Datasets: George Washington (GW), IAM database, CVL database, Bentham (BT)

[1] F. Wolf und G. A. Fink: *Self-Training of Handwritten Word Recognition for Synthetic-to-Real Adaptation*. In: *Proc. Int. Conf. on Pattern Recognition*, S. 3885–3892, Montreal, Canada, 2022. IEEE

Self-Training Benchmark Results

Selection	GW		IAM		CVL		BT	
	CER	WER	CER	WER	CER	WER	CER	WER
Synthetic	17.8	45.0	21.1	48.4	28.2	58.1	23.5	52.3
Self-Train (ST)	14.2	35.1	13.0	35.9	10.5	31.1	14.7	38.5
ST (Confidence)	**12.6**	**31.2**	**10.2**	**27.7**	**7.5**	**24.6**	**10.2**	**27.7**

Datasets: George Washington (GW), IAM database, CVL database, Bentham (BT)

[1] F. Wolf und G. A. Fink: *Self-Training of Handwritten Word Recognition for Synthetic-to-Real Adaptation*. In: *Proc. Int. Conf. on Pattern Recognition*, S. 3885–3892, Montreal, Canada, 2022. IEEE

Self-Training Benchmark Results

Selection	GW		IAM		CVL		BT	
	CER	WER	CER	WER	CER	WER	CER	WER
Synthetic	17.8	45.0	21.1	48.4	28.2	58.1	23.5	52.3
Self-Train (ST)	14.2	35.1	13.0	35.9	10.5	31.1	14.7	38.5
ST (Confidence)	12.6	31.2	10.2	27.7	7.5	24.6	10.2	27.7

Datasets: George Washington (GW), IAM database, CVL database, Bentham (BT)

[1] F. Wolf und G. A. Fink: *Self-Training of Handwritten Word Recognition for Synthetic-to-Real Adaptation*. In: *Proc. Int. Conf. on Pattern Recognition*, S. 3885–3892, Montreal, Canada, 2022. IEEE

Self-Training Benchmark Results

Selection	GW		IAM		CVL		BT	
	CER	WER	CER	WER	CER	WER	CER	WER
Synthetic	17.8	45.0	21.1	48.4	28.2	58.1	23.5	52.3
Self-Train (ST)	14.2	35.1	13.0	35.9	10.5	31.1	14.7	38.5
ST (Confidence)	**12.6**	**31.2**	**10.2**	**27.7**	**7.5**	**24.6**	**10.2**	**27.7**

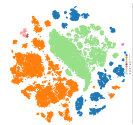
Datasets: George Washington (GW), IAM database, CVL database, Bentham (BT)

[1] F. Wolf und G. A. Fink: *Self-Training of Handwritten Word Recognition for Synthetic-to-Real Adaptation*. In: *Proc. Int. Conf. on Pattern Recognition*, S. 3885–3892, Montreal, Canada, 2022. IEEE

Conclusions

- Annotation-free pipeline for information extraction
- Application to new semi-structured document image collections without the manual creation of annotations
- Model training on synthetic data and adaptation with self-training

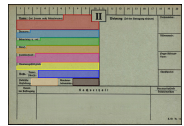
1. Template Extraction



2. Template Cleaning



3. Template Annotation



4. Data Synthesis



5. Information Detection



6. Text Recognition

