Post-Retrieval Semantic Re-Ranking via Zero-shot LLMs for Segmentation-Free Document Image Keyword Spotting

Stergios Papazis

Thursday, 10 July 2025

Department of Computer Science & Engineering
School of Engineering
University of Ioannina

Table of contents

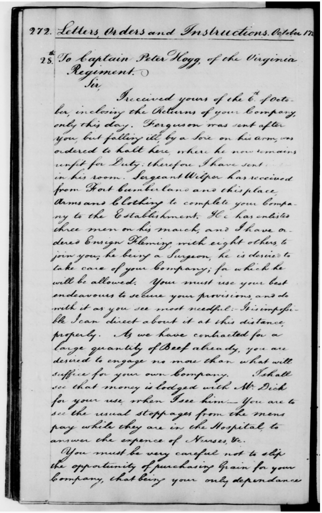
Keyword Spotting Overview

Document preservation

- digitization → collections of document images

Document image analysis

- search
- summarize
- automatic translation
- question answering
- named entity linking


272. Letters, Orders and Instructions. October 17[illegible]

28th To Captain Peter Hogg, of the Virginia Regiment.

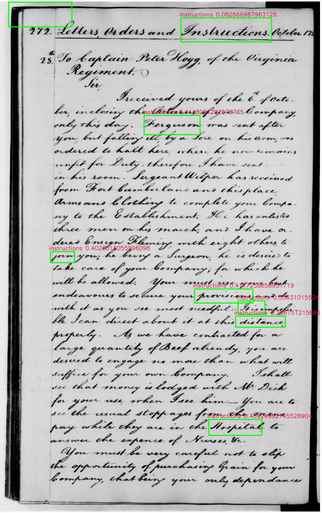
Sir,

I received yours of the 6th of October, inclosing the Returns of your Company, only this day. Ferguson was sent after you, but falling ill, by a Sore on his arm, was ordered to halt here, where he now remains unfit for Duty: therefore I have sent in his room. Sergeant Wilper has received from Fort Cumberland and this place, Arms and Clothing to complete your Company to the Establishment. He has enlisted three men on his march, and I have ordered Ensign Fleming, with eight others, to join you; he being a Surgeon, he is desired to take care of your Company; for which he will be allowed. Your must use your best endeavours to secure your provisions, and do with it as you see most needful:- It is impossible I can direct about it at this distance, properly. As we have contracted for a large quantity of Beef already, you are desired to engage no more than what will suffice for your own Company. I shall see that money is lodged with Mr. Dick for your use, when I see him — You are to see the usual stoppages from the mens pay while they are in the Hospital, to answer the expence of Nurses, &c.

You must be very careful not to slip the opportunity of purchasing Grain for your Company, that being your only dependance

Problem

- efficient retrieval of keywords in digitized documents


272. Letters, Orders and Instructions. October 17

28th To Captain Peter Hogg, of the Virginia Regiment.

Sir,

I received yours of the 6th of October, inclosing the Return of your Company, only this day. Ferguson was sent after you, but falling ill, by a Sore on his arm, was ordered to halt here, where he now remains unfit for Duty: therefore I have sent in his room. Sergeant Wilper has received from Fort Cumberland and this place Arms and Clothing to complete your Company to the Establishment. He has enlisted three men on his march, and I have ordered Ensign Fleming with eight others to join you; he being a Surgeon, he is desired to take care of your Company; for which he will be allowed. You must use your best endeavours to secure your provisions with it as you see most needful: It is impossible I can direct about it at this distance, properly. As we have contracted for a large quantity of Beef already, you are desired to engage no more than what will suffice for your own Company. I shall see that money is lodged with Mr. Dick for your use when I see him — You are to see the usual stoppages from pay while they are in the Hospital to answer the expence of Nurses, &c.

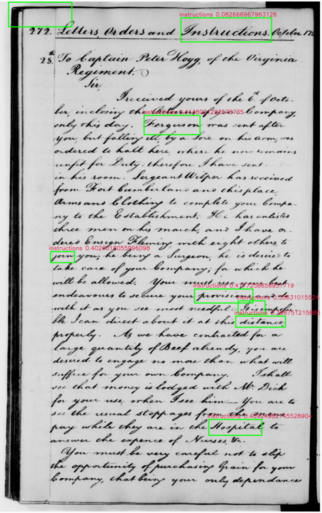
You must be very careful not to slip the opportunity of purchasing Grain for your Company, that being your only dependance

Problem

- efficient retrieval of keywords in digitized documents

Naive approach

- text transcription (e.g., OCR)
- retrieval on text



272. Letters, Orders and Instructions. October 17[illegible]

28th To Captain Peter Hogg, of the Virginia Regiment.

Sir,

I received yours of the 6th of October, inclosing the [illegible] Company, only this day. Ferguson was sent after you, but falling ill, by a Sore on his arm, was ordered to halt here, where he now remains unfit for Duty: therefore I have sent [illegible] in his room. Sergeant Wilper has received from Fort Cumberland and this place, Arms and Clothing to complete your Company to the Establishment. He has enlisted three men on his march, and I have ordered Ensign Fleming with eight others, to join you; he being a Surgeon, he is desired to take care of your Company; for which he will be allowed. You must use your best endeavours to secure your provisions and do with it as you see most needful: It is impossible I can direct about it at this distance, properly. As we have contracted for a large quantity of Beef already, you are desired to engage no more than what will suffice for your own Company. I shall see that money is lodged with Mr. Dick for your use, when I see him. — You are to see the usual stoppages from the mens pay while they are in the Hospital to answer the expence of Nurses, &c.

You must be very careful not to slip the opportunity of purchasing Grain for your Company, that being your only dependance

Problem

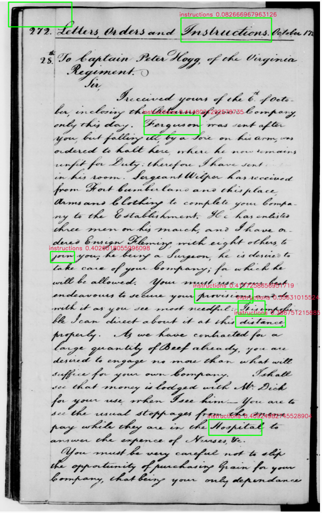
- efficient retrieval of keywords in digitized documents

Naive approach

- text transcription (e.g., OCR)
- retrieval on text
- recognition errors

Challenges

- intra/inter- writer variability
- physical deterioration: ink fading, staining, paper degradation


272. Letters, Orders and Instructions. October 17[illegible]

28th To Captain Peter Hogg, of the Virginia Regiment.

Sir,

I received yours of the 6th of October, inclosing the [illegible] Company only this day. Ferguson was sent after you, but falling ill, by a Sore on his arm, was ordered to halt here, where he now remains unfit for Duty: therefore I have sent [illegible] in his room. Sergeant Wilper has received from Fort Cumberland and this place, Arms and Clothing to complete your Company to the Establishment. He has enlisted three men on his march, and I have ordered Ensign Fleming with eight others, to join you; he being a Surgeon, he is desired to take care of your Company; for which he will be allowed. Your must use your best endeavours to secure your provisions and do with it as you see most needful: It is impossible I can direct about it at this distance properly. As we have contracted for a large quantity of Beef already, you are desired to engage no more than what will suffice for your own Company. I shall see that money is lodged with Mr. Dick for your use, when I see him — You are to see the usual stoppages from the men's pay while they are in the Hospital to answer the expence of Nurses, &c.

You must be very careful not to slip the opportunity of purchasing Grain for your Company, that being your only dependance

Keyword Spotting

KWS: *recognition-free* image retrieval

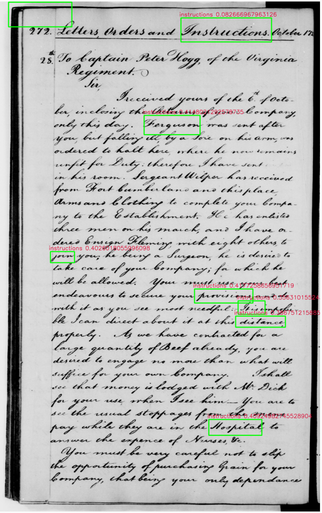
- given a keyword find all instances in document corpus
- returned them sorted by relevance

Enables

- efficient keyword-based search
- improved browsing experience

Avoids

- automatic transcription → no recognition errors
- manual annotation → cost effective


272. Letters, Orders and Instructions October 17..

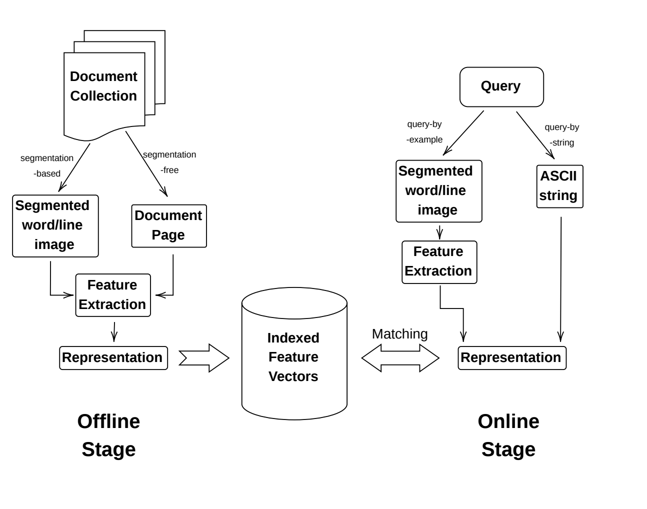
28th To Captain Peter Hogg, of the Virginia Regiment.

Sir,

I received yours of the 6th of October, inclosing the Return of your Company only this day. Ferguson was sent after you, but falling ill, by a Sore on his arm, was ordered to halt here, where he now remains unfit for Duty: therefore I have sent in his room. Sergeant Wilper has received from Fort Cumberland and this place Arms and Clothing to complete your Company to the Establishment. He has enlisted three men on his march, and I have ordered Ensign Fleming with eight others, to join you; he being a Surgeon, he is desired to take care of your Company; for which he will be allowed. You must use your best endeavours to secure your provisions with it as you see most needful: It is impossible I can direct about it at this distance properly. As we have contracted for a large quantity of Beef already, you are desired to engage no more than what will suffice for your own Company. I shall see that money is lodged with Mr Dick for your use when I see him — You are to see the usual stoppages from pay while they are in the Hospital to answer the expence of Nurses, &c.

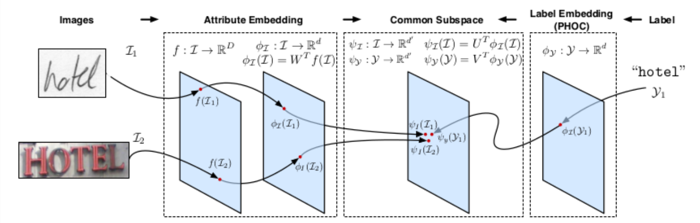
You must be very careful not to slip the opportunity of purchasing Grain for your Company, that being your only dependance

Keyword Spotting: General Architecture



Giotis et al. [1]

Keyword Spotting: General Architecture



Figures: Almazán et al. [2] and Wilkinson, Lindström, and Brun [3]

Keyword Spotting: Taxonomies

KWS approaches classified by assumptions

- Input Modularity
  - query-by-example: image example
  - query-by-string: alphanumeric string
- Segmentation
  - segmentation-based: word images or line images
  - segmentation-free: document images
- Machine Learning
  - learning-free: handcraft features, no manual annotations
  - learning-based: superior performance (CNN, HMM, SVM)
- Similarity
  - verbatim: visual resemblance
  - semantic: semantic resemblance
  - hybrid: verbatim & semantic

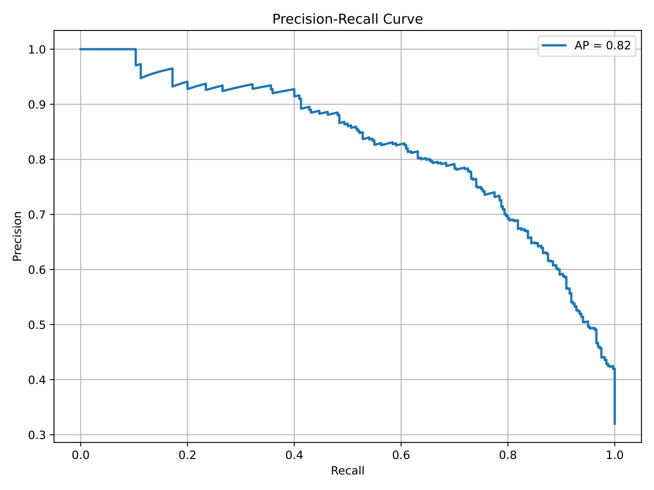
Keyword Spotting: Metrics

Information Retrieval Metrics

- Precision $= \frac{\text{number of } \textbf{relevant retrieved} \text{ instances}}{\text{number of } \textbf{retrieved} \text{ instances}} = \frac{TP}{TP+FP}$
- Recall $= \frac{\text{number of } \textbf{relevant retrieved} \text{ instances}}{\text{number of } \textbf{relevant} \text{ instances}} = \frac{TP}{TP+FN}$
- Average Precision $= \frac{1}{R}\sum_{k=1}^{n} \mathrm{P@}k \cdot rel(k)$
  - Precision at rank k: $\mathrm{P@}k =$ precision of top k retrievals
  - $rel(k) = \begin{cases} 1, & \text{if k-th instance relevant} \\ 0, & \text{otherwise} \end{cases}$
  - R total relevant items
- Mean Average Precision (mAP):
  mean of Average Precision computed over all queries

Keyword Spotting: Metrics

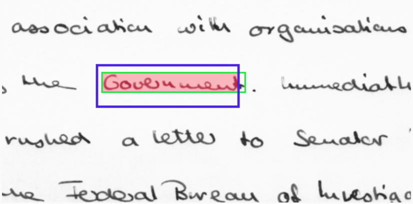
Average Precision estimates area under Precision-Recall curve.



Keyword Spotting: Metrics

When is an instance relevant?

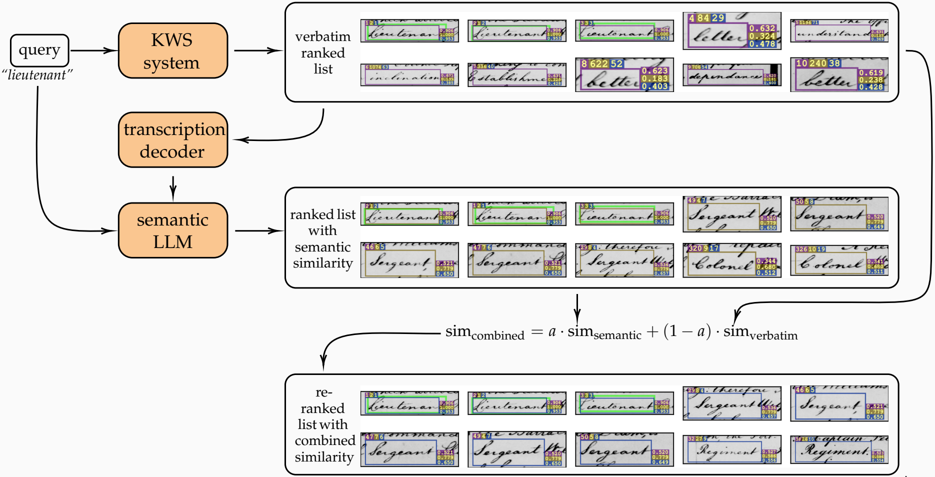
- Intersection over Union (IoU)



- $\text{IoU}(\text{prediction}, \text{ground-truth}) \geq \text{threshold}$
- typical thresholds: $\begin{cases} 0.25 & \rightarrow \quad \text{mAP@25} \\ 0.5 & \rightarrow \quad \text{mAP@50} \end{cases}$
- at most one instance considered relevant

Proposed System

Post-retrieval semantic re-ranking via zero-shot LLMs for segmentation-free KWS



Proposed System: Overview

Post-retrieval semantic re-ranking via zero-shot LLMs for segmentation-free KWS

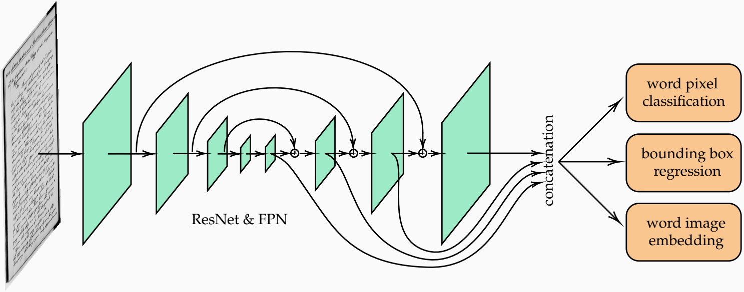
Objectives

1. improve exact retrieval (increase mAP)
2. enhance semantic quality of search

Pipeline variants

- **KWS Baselines**: WordRetrievalNet, KWS-Simplified
- **Decoders**: KWS-Simplified-based, TrOCR (visual-textual transformer)
- **Semantic LLM**: Roberta, MPNet, MiniLM

Baseline: WordRetrievalNet

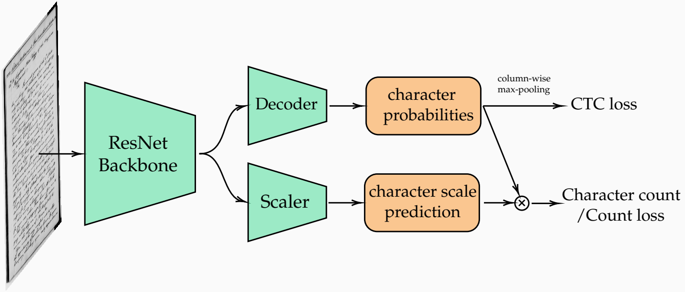


Loss

- dice coefficient loss: $1 - \frac{2|X \cap Y|}{|X| + |Y|}$
- distance-IoU loss: IoU loss penalizing center distances
- cosine loss

Zhao et al. [4]

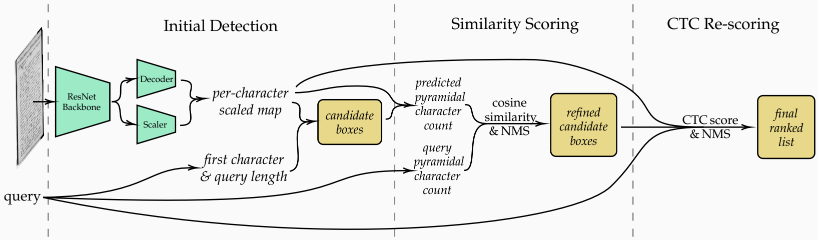
Baseline: Segmentation-free KWS Simplified



Offline stage

- character c, image coordinates (i,j), $F(i,j,c)$ probability, $S(i,j)$ scale
- $y_c = \sum_{i=s_1}^{e_1} \sum_{j=s_2}^{e_2} F(i,j,c) \cdot S(i,j)$ (regression)
- Connectionist Temporal Classification (CTC)
- training and pre-computing of F, S

Retsinas, Sfikas, and Nikou [5]



Online stage

- estimate first character location based on query first char
- estimate bounding box with query length
- pyramidal counting (like PHOC)
- CTC-based re-scoring

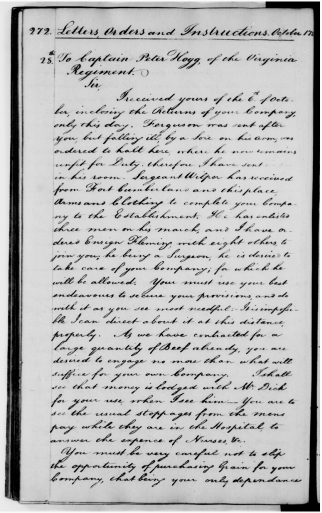
Retsinas, Sfikas, and Nikou [5]

Experimental Evaluation

George Washington Dataset (GW)

- 20 handwritten letters
- $4,860$ annotated words
- few writers (single-writer)
- 4-fold cross validation
- 14 train/1 validate/5 test
- test queries: all unique transcriptions test pages

Lavrenko, Rath, and Manmatha [6]



272. Letters, Orders and Instructions. October 17

28th To Captain Peter Hogg, of the Virginia Regiment.

Sir,

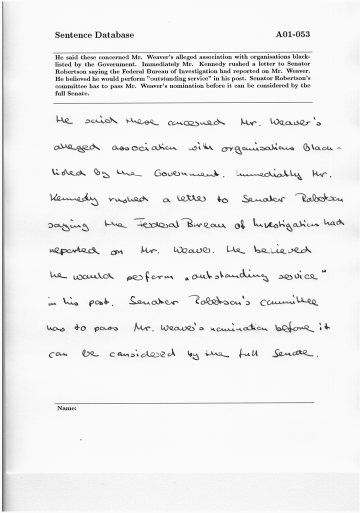
I received yours of the 6th of October, inclosing the Returns of your Company only this day. Ferguson was sent after you, but falling ill, by a Sore on his arm, was ordered to halt here, where he now remains unfit for Duty: therefore I have sent in his room. Sergeant Wilper has received from Fort Cumberland and this place Arms and Clothing to complete your Company to the Establishment. He has enlisted three men on his march, and I have ordered Ensign Fleming, with eight others, to join you; he being a Surgeon, he is desired to take care of your Company; for which he will be allowed. You must use your best endeavours to secure your provisions, and do with it as you see most needful: It is impossible I can direct about it at this distance properly. As we have contracted for a large quantity of Beef already, you are desired to engage no more than what will suffice for your own Company. I shall see that money is lodged with Mr Dick for your use, when I see him — You are to see the usual stoppages from the mens pay while they are in the Hospital, to answer the expence of Nurses, &c.

You must be very careful not to slip the opportunity of purchasing Grain for your Company, that being your only dependance

IAM Handwriting Database

- $1,539$ forms
- $115,320$ annotated words
- 657 writers
- official split (1,087/220/232)
- test queries: all unique transcriptions test pages, excluding stopwords, non alphanumeric

Marti and Bunke [7]



Sentence Database A01-053

He said these concerned Mr. Weaver's alleged association with organisations black-listed by the Government. Immediately Mr. Kennedy rushed a letter to Senator Robertson saying the Federal Bureau of Investigation had reported on Mr. Weaver. He believed he would perform "outstanding service" in his post. Senator Robertson's committee has to pass Mr. Weaver's nomination before it can be considered by the full Senate.

He said these concerned Mr. Weaver's
alleged association with organisations black-
listed by the Government. Immediatly Mr.
Kennedy rushed a letter to Senator Robertson
saying the Federal Bureau of Investigation had
reported on Mr. Weaver. He believed
he would perform „outstanding service"
in his post. Senator Robertson's committee
has to pass Mr. Weaver's nomination before it
can be considered by the full Senate.

Name:

Experimental Evaluation: Implementation

Training on GW ($\times 4$)

- WordRetrievalNet 120 epochs
- KWS-Simplified 200 epochs

Training on IAM

- WordRetrievalNet 80 epochs

Fine-tuned on GW ($\times 4$)

- TrOCR 20 epochs

RoBERTa, MPNet, MiniLM: pre-trained

Experimental Evaluation: WordRetrievalNet

Semantic weight	Semantic LLM	George Washington Dataset				IAM Handwriting Database			
		KWS-Simplified decoder		TrOCR decoder		KWS-Simplified decoder		TrOCR decoder	
		mAP@25	mAP@50	mAP@25	mAP@50	mAP@25	mAP@50	mAP@25	mAP@50
0.0*	all -MiniLM -L12 -v2	94.31 ± 1.8	88.29 ± 4.0	94.31 ± 1.8	88.29 ± 4.0	79.15	72.85	79.15	72.85
0.1		95.80 ± 1.5	89.51 ± 3.8	**94.51 ± 1.8**	**88.43 ± 4.0**	**80.60**	**73.98**	**82.04**	**75.40**
0.2		96.10 ± 1.5	89.75 ± 3.8	93.96 ± 2.0	87.84 ± 4.0	79.29	72.62	80.59	73.77
0.3		**96.30 ± 1.3**	**89.79 ± 3.7**	93.20 ± 2.1	87.07 ± 3.8	75.87	69.30	77.68	71.05
0.4		96.04 ± 1.5	89.64 ± 3.6	91.71 ± 2.5	85.73 ± 3.9	71.36	65.13	74.04	67.66
0.5		95.34 ± 1.4	89.01 ± 3.5	89.05 ± 2.9	83.19 ± 3.8	65.95	60.19	70.68	64.57
0.6		94.39 ± 1.5	88.20 ± 3.5	84.82 ± 4.2	79.41 ± 4.2	60.26	54.99	67.17	61.33
0.7		93.25 ± 1.6	87.37 ± 3.5	79.04 ± 4.8	74.03 ± 3.7	55.95	50.99	64.47	58.88
0.8		91.92 ± 1.9	86.26 ± 3.5	73.12 ± 6.1	68.42 ± 4.5	53.12	48.38	62.88	57.42
0.9		90.93 ± 2.0	85.42 ± 3.6	69.68 ± 6.6	65.20 ± 4.8	51.46	46.88	62.22	56.83
1.0		89.83 ± 2.2	84.28 ± 3.7	64.58 ± 7.5	60.07 ± 6.0	47.96	43.71	54.34	49.48
0.0*	all -mpnet -base -v2	94.31 ± 1.8	88.29 ± 4.0	94.31 ± 1.8	88.29 ± 4.0	79.15	72.85	79.15	72.85
0.1		95.90 ± 1.4	89.67 ± 3.8	**94.68 ± 1.6**	**88.62 ± 3.9**	**80.84**	**74.18**	**82.12**	**75.43**
0.2		96.27 ± 1.3	89.86 ± 3.8	94.09 ± 1.7	87.98 ± 3.8	80.01	73.21	80.77	74.02
0.3		96.50 ± 1.4	89.97 ± 4.0	93.07 ± 2.1	87.02 ± 3.7	77.55	70.82	77.87	71.29
0.4		**96.56 ± 1.5**	**90.08 ± 3.9**	91.78 ± 2.4	85.85 ± 3.7	73.86	67.22	74.73	68.33
0.5		96.00 ± 1.3	89.62 ± 3.6	89.42 ± 2.9	83.85 ± 3.8	69.24	62.95	71.23	65.12
0.6		95.04 ± 1.2	88.82 ± 3.5	85.13 ± 3.6	79.84 ± 3.8	63.70	57.97	67.84	61.99
0.7		93.83 ± 1.6	87.86 ± 3.4	79.10 ± 5.3	74.10 ± 4.1	58.18	52.97	65.43	59.78
0.8		92.57 ± 1.7	86.71 ± 3.4	74.06 ± 6.2	69.30 ± 4.6	53.92	49.10	63.80	58.25
0.9		91.53 ± 1.8	85.84 ± 3.6	70.50 ± 6.4	65.90 ± 4.6	51.43	46.89	62.78	57.36
1.0		90.21 ± 2.1	84.45 ± 3.7	64.91 ± 7.5	60.31 ± 6.0	49.26	44.84	54.99	50.11
0.0*	stsb -roberta -base	94.31 ± 1.8	88.29 ± 4.0	94.31 ± 1.8	88.29 ± 4.0	79.15	72.85	79.15	72.85
0.1		95.83 ± 1.6	89.60 ± 3.9	**94.58 ± 2.0**	**88.55 ± 4.2**	**81.16**	**74.49**	**81.88**	**75.18**
0.2		96.19 ± 1.3	89.82 ± 3.7	94.01 ± 2.1	87.91 ± 4.2	80.39	73.55	80.97	74.22
0.3		**96.59 ± 1.3**	**90.17 ± 3.7**	92.87 ± 2.9	86.92 ± 4.5	77.40	70.68	77.51	70.90
0.4		96.44 ± 1.6	89.99 ± 3.7	90.75 ± 3.2	84.96 ± 4.4	73.54	66.94	74.29	67.87
0.5		96.32 ± 1.6	89.90 ± 3.7	87.91 ± 3.5	82.32 ± 4.3	69.45	63.25	71.28	65.11
0.6		96.04 ± 1.7	89.57 ± 3.8	84.15 ± 4.3	78.83 ± 4.9	64.83	59.05	68.70	62.85
0.7		95.27 ± 1.7	88.91 ± 4.0	80.37 ± 4.4	75.25 ± 4.7	61.28	55.73	66.71	61.03
0.8		94.37 ± 1.8	88.08 ± 3.9	77.09 ± 4.4	72.16 ± 4.3	58.23	53.05	65.16	59.66
0.9		93.54 ± 2.0	87.46 ± 4.0	74.05 ± 4.8	69.22 ± 4.1	56.22	51.22	64.13	58.74
1.0		92.32 ± 2.2	86.19 ± 3.9	67.98 ± 5.4	63.34 ± 4.3	53.99	49.11	56.03	51.24

*This is essentially the baseline model. It does not use a decoder.

Experimental Evaluation: KWS-Simplified

Semantic weight	Semantic LLM	George Washington Dataset				IAM Handwriting Database			
		KWS-Simplified decoder		TrOCR decoder		KWS-Simplified decoder		TrOCR decoder	
		mAP@25	mAP@50	mAP@25	mAP@50	mAP@25	mAP@50	mAP@25	mAP@50
0.0*	all-MiniLM-L12-v2	89.74 ± 0.7	72.29 ± 3.0	89.74 ± 0.7	72.29 ± 3.0	86.40	63.73	86.40	63.73
0.1		90.62 ± 0.5	72.77 ± 3.0	90.27 ± 0.8	72.53 ± 3.1	86.71	63.86	87.49	64.22
0.2		90.68 ± 0.4	72.82 ± 3.0	**90.38 ± 0.7**	**72.57 ± 3.1**	**86.79**	**63.94**	87.84	64.47
0.3		90.70 ± 0.5	72.84 ± 3.0	90.35 ± 0.7	72.55 ± 3.2	86.57	63.72	88.01	64.58
0.4		90.72 ± 0.5	72.84 ± 3.0	90.30 ± 0.7	72.51 ± 3.2	86.11	63.36	**88.06**	**64.60**
0.5		90.79 ± 0.4	72.90 ± 3.0	90.32 ± 0.6	72.55 ± 3.2	85.65	63.01	87.74	64.43
0.6		90.85 ± 0.5	72.93 ± 3.0	90.32 ± 0.6	72.54 ± 3.2	84.66	62.38	87.28	64.17
0.7		**90.91 ± 0.5**	**72.97 ± 3.0**	90.26 ± 0.7	72.51 ± 3.2	83.32	61.47	86.30	63.56
0.8		90.83 ± 0.4	72.90 ± 2.9	90.12 ± 0.6	72.42 ± 3.1	81.31	60.20	84.66	62.49
0.9		90.82 ± 0.4	72.89 ± 2.9	89.93 ± 0.6	72.29 ± 3.0	77.61	57.85	81.51	60.47
1.0		90.43 ± 0.7	72.73 ± 2.7	87.16 ± 0.6	70.39 ± 2.1	71.97	54.39	76.74	57.24
0.0*	all-mpnet-base-v2	89.74 ± 0.7	72.29 ± 3.0	89.74 ± 0.7	72.29 ± 3.0	86.40	63.73	86.40	63.73
0.1		90.63 ± 0.5	72.78 ± 3.0	90.22 ± 0.7	72.49 ± 3.1	86.69	**63.88**	87.53	64.25
0.2		90.69 ± 0.5	72.83 ± 3.0	90.31 ± 0.7	72.51 ± 3.1	**86.70**	63.87	87.89	64.51
0.3		90.72 ± 0.5	72.84 ± 3.0	90.29 ± 0.6	72.49 ± 3.1	86.56	63.75	88.16	64.80
0.4		90.81 ± 0.5	72.91 ± 3.1	**90.32 ± 0.7**	**72.54 ± 3.1**	86.32	63.59	**88.25**	**64.88**
0.5		90.86 ± 0.5	72.96 ± 3.1	90.29 ± 0.7	72.53 ± 3.2	85.91	63.33	87.97	64.69
0.6		90.87 ± 0.4	72.94 ± 3.0	90.26 ± 0.6	72.49 ± 3.1	85.03	62.71	87.52	64.37
0.7		**90.94 ± 0.5**	**72.97 ± 3.1**	90.10 ± 0.7	72.39 ± 3.1	84.15	62.10	86.60	63.77
0.8		90.91 ± 0.5	72.96 ± 3.0	90.06 ± 0.6	72.35 ± 3.0	82.56	61.09	85.07	62.78
0.9		90.85 ± 0.4	72.92 ± 3.0	89.74 ± 0.4	72.09 ± 2.9	79.47	59.19	81.97	60.86
1.0		90.47 ± 0.5	72.78 ± 2.8	87.06 ± 0.7	70.23 ± 2.1	74.04	55.89	77.30	57.78
0.0*	stsb-roberta-base	89.74 ± 0.7	72.29 ± 3.0	89.74 ± 0.7	72.29 ± 3.0	86.40	63.73	86.40	63.73
0.1		90.62 ± 0.5	72.78 ± 3.0	90.21 ± 0.6	72.46 ± 3.0	86.58	63.82	87.35	64.13
0.2		90.63 ± 0.5	72.79 ± 3.0	90.27 ± 0.6	72.47 ± 3.0	**86.78**	**63.85**	87.78	64.52
0.3		90.66 ± 0.5	72.80 ± 3.0	90.27 ± 0.6	72.47 ± 3.1	86.72	63.74	88.04	64.72
0.4		90.76 ± 0.5	72.88 ± 3.0	90.32 ± 0.6	72.52 ± 3.1	86.45	63.59	**88.12**	**64.84**
0.5		90.89 ± 0.5	**72.97 ± 3.1**	**90.33 ± 0.5**	**72.54 ± 3.1**	85.87	63.29	87.62	64.60
0.6		90.90 ± 0.4	**72.97 ± 3.1**	90.30 ± 0.4	72.50 ± 3.0	85.09	62.79	86.78	64.16
0.7		**90.92 ± 0.4**	**72.97 ± 3.1**	90.13 ± 0.4	72.41 ± 2.9	83.82	61.99	85.44	63.32
0.8		90.86 ± 0.4	72.92 ± 3.0	89.96 ± 0.5	72.28 ± 2.9	81.96	60.79	83.06	61.79
0.9		90.74 ± 0.4	72.86 ± 3.0	89.75 ± 0.5	72.12 ± 2.8	79.45	59.12	80.17	59.88
1.0		90.33 ± 0.5	72.71 ± 2.8	87.08 ± 0.9	70.24 ± 2.0	76.10	57.18	76.49	57.40

*This is essentially the baseline model. It does not use a decoder.

Baseline: WordRetrievalNet > KWS-Simplified

	GW (mAP@25/50)	IAM (mAP@25/50)
WordRetrievalNet	+2.3/+0.9	+3.0/+2.6
KWS-Simplified	+1.2/+0.7	+1.9/1.2

- WordRetrievalNet: query-independent pre-compute word locations
- WordRetrievalNet more candidates → more efficient re-ranking
- KWS-Simplified fewer candidates to pick from (no query expansion)
- small gains mAP@50 on GW

Experimental Evaluation: Results Overview

Decoder

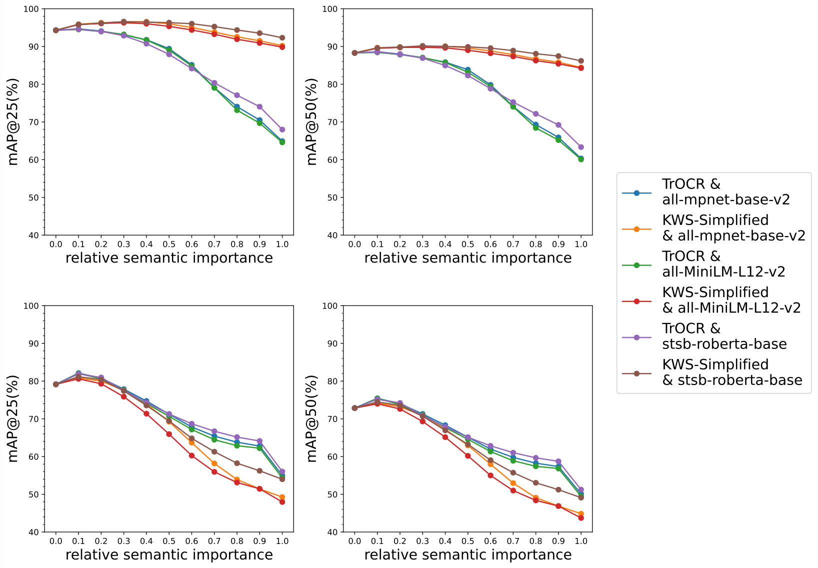
- on GW: KWS-Simplified $>$ TrOCR
  - mAP@25%: $+2.3\% > +0.6\%$
  - mAP@50%: $+0.9\% > +0.3\%$
- on IAM: KWS-Simplified $<$ TrOCR
  - mAP@25%: $+2.0\% < +3.0\%$
  - mAP@50%: $+1.6\% < +2.6\%$
- final performance shaped by accuracy of initial bounding boxes, not only decoder
- importance of holistic design (end-to-end)

Experimental Evaluation: Results Overview

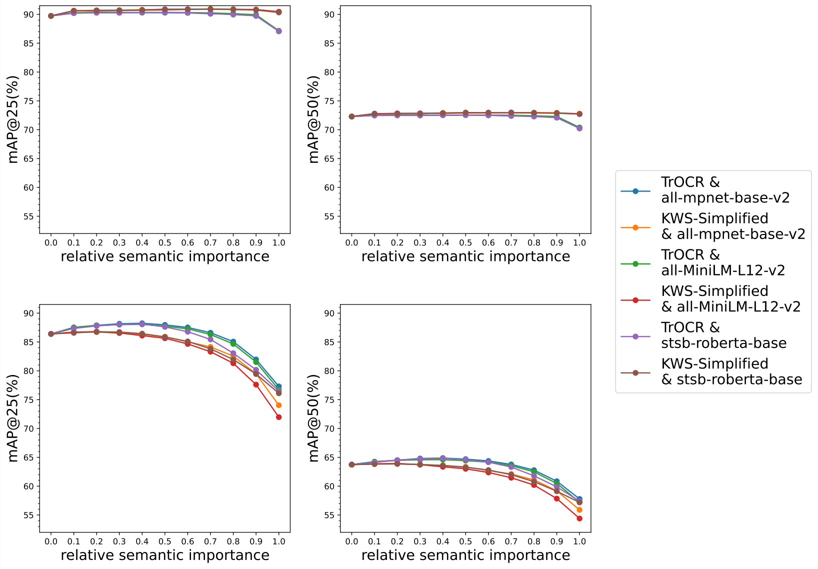
Semantic LLMs

- indistinguishable (usually less 0.3%, at most 0.5%)
- robustness and adaptability

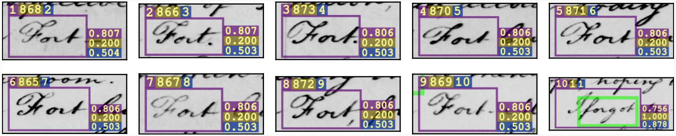
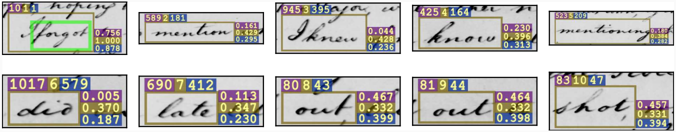
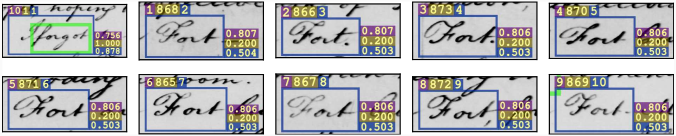
WordRetrievalNet Results



KWS-Simplified Results



Visual Examples

Similarity	Top-10 Ranked Lists
Verbatim	
Semantic	
Combined	

Visual Examples

Similarity	Top-21 Ranked Lists
Verbatim	
Semantic	
Combined	

Conclusion

Conclusions

- promising semantic re-ranking: modular, extension KWS system
- effective generalization across datasets
- consistent gains, adaptable to different LLMs
- low variability
- WordRetrievalNet on IAM
  - $+3.0\%$ (mAP@25: from 79.15% to 82.12%)
  - $+2.6\%$ (mAP@50: from 72.85% to 75.43%)
- KWS-Simplified on IAM
  - $+1.85\%$ (mAP@25: from 86.40% to 88.25%)
  - $+1.15\%$ (mAP@50: from 63.73% to 64.88%)
- WordRetrievalNet on GW: $+2.3\%$ (mAP@25: from 94.31% to 96.59%)
- KWS-Simplified on GW: $+1.2\%$ (mAP@25: from 89.74% to 90.94%)
- qualitive improvements

Future Directions

End-to-end semantic representations for segm-free KWS, too

- search
- summarize
- automatic translation
- question answering
- named entity linking

Thanks for your attention!

References

[1] Angelos P. Giotis et al. "A survey of document image word spotting techniques". In: *Pattern Recognition* 68 (2017), pp. 310–332. DOI: 10.1016/j.patcog.2017.02.023 (cit. on p. 9).

[2] J. Almazán et al. "Word Spotting and Recognition with Embedded Attributes". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 36.12 (2014), pp. 2552–2566. DOI: 10.1109/TPAMI.2014.2339814 (cit. on p. 10).

[3] Tomas Wilkinson, Jonas Lindström, and Anders Brun. "Neural Ctrl-F: Segmentation-Free Query-by-String Word Spotting in Handwritten Manuscript Collections". In: *Proceedings of the 16^{th} IEEE International Conference on Computer Vision (ICCV)*. 2017, pp. 4443–4452. DOI: 10.1109/ICCV.2017.475 (cit. on p. 10).

[4] Peng Zhao et al. "Query by Strings and Return Ranking Word Regions with Only One Look". In: *Proceedings of the Asian Conference on Computer Vision (ACCV)*. 2020, pp. 3–18. DOI: 10.1007/978-3-030-69544-6_1 (cit. on p. 18).

[5] George Retsinas, Giorgos Sfikas, and Christophoros Nikou. "Keyword Spotting Simplified: A Segmentation-Free Approach Using Character Counting and CTC Re-scoring". In: *Document Analysis and Recognition - ICDAR 2023*. Springer Nature Switzerland, 2023, pp. 446–464. DOI: 10.1007/978-3-031-41676-7_26 (cit. on pp. 19, 20).

[6] V. Lavrenko, T. M. Rath, and R. Manmatha. "Holistic word recognition for handwritten historical documents". In: *Proceedings of the 1^{st} International Workshop on Document Image Analysis for Libraries*. 2004, pp. 278–287. DOI: 10.1109/DIAL.2004.1263256 (cit. on p. 22).

References

[7] U.-V. Marti and H. Bunke. "The IAM-database: an English sentence database for offline handwriting recognition". In: *International Journal on Document Analysis and Recognition* 5.1 (Nov. 2002), pp. 39–46. DOI: 10.1007/s100320200071 (cit. on p. 23).

[8] Tomas Wilkinson and Anders Brun. "Semantic and Verbatim Word Spotting Using Deep Neural Networks". In: *15th International Conference on Frontiers in Handwriting Recognition (ICFHR)*. Shenzhen, China: IEEE, 2016, pp. 307–312. DOI: 10.1109/ICFHR.2016.0065.

[9] Sounak Dey et al. "Evaluation of word spotting under improper segmentation scenario". In: *International Journal on Document Analysis and Recognition (IJDAR)* 22 (2019), pp. 361–374. DOI: 10.1007/s10032-019-00338-9.

[10] Praveen Krishnan and CV Jawahar. "Bringing semantics into word image representation". In: *Pattern Recognition* 108 (2020), p. 107542. DOI: 10.1016/j.patcog.2020.107542.

[11] Praveen Krishnan and C. V. Jawahar. "HWNet v2: an efficient word image representation for handwritten documents". In: *International Journal on Document Analysis and Recognition (IJDAR)* 22.4 (2019), pp. 387–405. DOI: 10.1007/s10032-019-00336-x.

[12] Oliver Tüselmann, Fabian Wolf, and Gernot A Fink. "Identifying and Tackling Key Challenges in Semantic Word Spotting". In: *2020 17th International Conference on Frontiers in Handwriting Recognition (ICFHR)*. IEEE. 2020, pp. 55–60. DOI: 10.1109/ICFHR2020.2020.00021.

[13] Oliver Tüselmann, Fabian Wolf, and Gernot A. Fink. "Are End-to-End Systems Really Necessary for NER on Handwritten Document Images?" In: *Document Analysis and Recognition – ICDAR 2021*. Springer International Publishing, 2021, pp. 808–822. DOI: 10.1007/978-3-030-86331-9_52.

References

[14] Oliver Tüselmann and Gernot A. Fink. "Named Entity Linking on Handwritten Document Images". In: *Document Analysis Systems*. Springer International Publishing, 2022, pp. 199–213. DOI: 10.1007/978-3-031-06555-2_14.

[15] Oliver Tüselmann et al. "Recognition-Free Question Answering on Handwritten Document Collections". In: *Frontiers in Handwriting Recognition*. Springer International Publishing, 2022, pp. 259–273. DOI: 10.1007/978-3-031-21648-0_18.

[16] Oliver Tüselmann and Gernot A Fink. "Exploring Semantic Word Representations for Recognition-Free NLP on Handwritten Document Images". In: *International Conference on Document Analysis and Recognition*. Springer. 2023, pp. 85–100. DOI: 10.1007/978-3-031-41685-9_6.

[17] Oliver Tüselmann and Gernot A. Fink. "Neural models for semantic analysis of handwritten document images". In: *International Journal on Document Analysis and Recognition (IJDAR)* 27.3 (2024), pp. 245–263. DOI: 10.1007/s10032-024-00477-8.

[18] George Washington. *George Washington Papers, Series 2, Letterbooks 1754-1799: Letterbook 1, Aug. 11, 1754 - Dec. 25, 1755*. Last accessed 30 April 2025, pp. 270–279, 300–309. URL: https://www.loc.gov/item/mgw2.001.

[19] Jon Almazán et al. "Segmentation-free word spotting with exemplar SVMs". In: *Pattern Recognition* 47.12 (2014), pp. 3967–3978. DOI: 10.1016/j.patcog.2014.06.005.

[20] Leonard Rothacker and Gernot A. Fink. "Segmentation-free Query-by-String Word Spotting with Bag-of-Features HMMs". In: *Proceedings of the 13^{th} International Conference on Document Analysis and Recognition (ICDAR)*. 2015, pp. 661–665. DOI: 10.1109/ICDAR.2015.7333844.

References

[21] Tomas Wilkinson, Jonas Lindström, and Anders Brun. *Neural Word Search in Historical Manuscript Collections*. 2020. DOI: `10.48550/arXiv.1812.02771`.

[22] Ilya Loshchilov and Frank Hutter. *Decoupled Weight Decay Regularization*. 2019. DOI: `10.48550/arXiv.1711.05101`.

[23] Minghao Li et al. "TrOCR: Transformer-based Optical Character Recognition with Pre-trained Models". In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 37. 11. 2023, pp. 13094–13102. DOI: `10.1609/aaai.v37i11.26538`.

[24] Daniel Cer et al. "SemEval-2017 Task 1: Semantic Textual Similarity Multilingual and Crosslingual Focused Evaluation". In: *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*. Aug. 2017, pp. 1–14. DOI: `10.18653/v1/S17-2001`.

[25] Sebastian Sudholt and Gernot A. Fink. "Evaluating Word String Embeddings and Loss Functions for CNN-Based Word Spotting". In: *Proceedings of the 14^{th} IAPR International Conference on Document Analysis and Recognition (ICDAR)*. 2017, pp. 493–498. DOI: `10.1109/ICDAR.2017.87`.

[26] Arjun Sharma and S. K. Pramod. "Adapting off-the-shelf CNNs for word spotting & recognition". In: *Proceedings of the 13^{th} International Conference on Document Analysis and Recognition (ICDAR)*. 2015, pp. 986–990. DOI: `10.1109/ICDAR.2015.7333909`.

[27] George Retsinas et al. "Efficient Learning-Free Keyword Spotting". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41.7 (2019), pp. 1587–1600. DOI: `10.1109/TPAMI.2018.2845880`.

References

[28] P. Krishnan, K. Dutta, and C. V. Jawahar. “Deep Feature Embedding for Accurate Recognition and Retrieval of Handwritten Text”. In: *Proceedings of the 15^{th} International Conference on Frontiers in Handwriting Recognition (ICFHR)*. 2016, pp. 289–294. DOI: 10.1109/ICFHR.2016.0062.

[29] Jon Almazán et al. “Efficient Exemplar Word Spotting”. In: *Proceedings of the British Machine Vision Conference*. 2012, pp. 67.1–67.11. DOI: 10.5244/C.26.67.

[30] Fabian Wolf, Kai Brandenbusch, and Gernot A. Fink. “Improving Handwritten Word Synthesis for Annotation-free Word Spotting”. In: *Proceedings of the 17^{th} International Conference on Frontiers in Handwriting Recognition (ICFHR)*. 2020, pp. 61–66. DOI: 10.1109/ICFHR2020.2020.00022.

[31] Fatemeh Daraee, Saeed Mozaffari, and Seyyed Mohammad Razavi. “Handwritten keyword spotting using deep neural networks and certainty prediction”. In: *Computers & Electrical Engineering* 92 (2021), pp. 107–111. DOI: 10.1016/j.compeleceng.2021.107111.

[32] Mohamed Mhiri, Christian Desrosiers, and Mohamed Cheriet. “Word spotting and recognition via a joint deep embedding of image and text”. In: *Pattern Recognition* 88 (2019), pp. 312–320. DOI: 10.1016/j.patcog.2018.11.017.

[33] M. Rusiñol et al. “Browsing Heterogeneous Document Collections by a Segmentation-Free Word Spotting Method”. In: *Proceedings of the 11^{th} International Conference on Document Analysis and Recognition (ICDAR)*. 2011, pp. 63–67. DOI: 10.1109/ICDAR.2011.22.

[34] A. Kovalchuk, L. Wolf, and N. Dershowitz. “A Simple and Fast Word Spotting Method”. In: *Proceedings of the 14^{th} International Conference on Frontiers in Handwriting Recognition (ICFHR)*. 2014, pp. 3–8. DOI: 10.1109/ICFHR.2014.9.

References

[35] Praveen Krishnan and Cheerakkuzhi Veluthemana Jawahar. "Bringing Semantics in Word Image Retrieval". In: *2013 12th International Conference on Document Analysis and Recognition*. 2013, pp. 733–737. DOI: 10.1109/ICDAR.2013.150.

[36] L. Rothacker, M. Rusiñol, and G. A. Fink. "Bag-of-Features HMMs for Segmentation-Free Word Spotting in Handwritten Documents". In: *Proceedings of the 12^{th} International Conference on Document Analysis and Recognition (ICDAR)*. 2013, pp. 1305–1309. DOI: 10.1109/ICDAR.2013.264.

[37] Tomas Mikolov et al. *Efficient Estimation of Word Representations in Vector Space*. 2013. DOI: 10.48550/arXiv.1301.3781.

[38] Piotr Bojanowski et al. "Enriching Word Vectors with Subword Information". In: *Transactions of the Association for Computational Linguistics* 5 (June 2017), pp. 135–146. DOI: 10.1162/tacl_a_00051.

[39] Jeffrey Pennington, Richard Socher, and Christopher Manning. "GloVe: Global Vectors for Word Representation". In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Oct. 2014, pp. 1532–1543. DOI: 10.3115/v1/D14-1162.

[40] Jacob Devlin et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*. 2019, pp. 4171–4186. DOI: 10.48550/arXiv.1810.04805.

[41] Yinhan Liu et al. "RoBERTa: A Robustly Optimized BERT Pretraining Approach". In: (2019). DOI: 10.48550/arXiv.1907.11692.

References

[42] Kaitao Song et al. "MPNet: Masked and Permuted Pre-training for Language Understanding". In: *Proceedings of the 34th International Conference on Neural Information Processing Systems*. 2020, pp. 16857–16867. DOI: arXiv:2004.09297.

[43] Wenhui Wang et al. "MiniLM: Deep Self-Attention Distillation for Task-Agnostic Compression of Pre-Trained Transformers". In: *Proceedings of the 34th International Conference on Neural Information Processing Systems*. 2020, pp. 5776–5788. DOI: arXiv:2002.10957.

[44] George A Miller. "WordNet: a lexical database for English". In: *Communications of the ACM* 38.11 (1995), pp. 39–41. DOI: 10.1145/219717.219748.

[45] Albert Gordo et al. "LEWIS: Latent Embeddings for Word Images and Their Semantics". In: *Proceedings of the IEEE International Conference on Computer Vision*. 2015, pp. 1242–1250. DOI: 10.1109/ICCV.2015.147.

[46] Minesh Mathew, Dimosthenis Gomez Lluisand Karatzas, and C. V. Jawahar. "Asking questions on handwritten document collections". In: *International Journal on Document Analysis and Recognition (IJDAR)* 24.3 (Sept. 2021), pp. 235–249. DOI: 10.1007/s10032-021-00383-3.

[47] Chandranath Adak et al. "Detecting Named Entities in Unstructured Bengali Manuscript Images". In: *2019 International Conference on Document Analysis and Recognition (ICDAR)*. 2019, pp. 196–201. DOI: 10.1109/ICDAR.2019.00040.

[48] Praveen Krishnan and CV Jawahar. "Generating synthetic data for text recognition". In: *arXiv preprint arXiv:1608.04224* (2016). DOI: 10.48550/arXiv.1608.04224.

[49] Yonghui Wu et al. "Google's neural machine translation system: Bridging the gap between human and machine translation". In: *arXiv preprint arXiv:1609.08144* (2016). DOI: 10.48550/arXiv.1609.08144.

References

[50] Alexey Dosovitskiy et al. “An image is worth 16x16 words: Transformers for image recognition at scale”. In: *arXiv preprint arXiv:2010.11929* (2020). DOI: 10.48550/arXiv.2010.11929.

[51] Nils Reimers and Iryna Gurevych. “Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks”. In: (Nov. 2019). DOI: 10.18653/v1/D19-1410.

[52] G. Louloudis, A.L. Kesidis, and B. Gatos. “Efficient Word Retrieval Using a Multiple Ranking Combination Scheme”. In: *Proceedings of the 10^{th} IAPR International Workshop on Document Analysis Systems (DAS)*. 2012, pp. 379–383. DOI: 10.1109/DAS.2012.34.

[53] Jinchuan Tian et al. “Bayes Risk CTC: Controllable CTC Alignment in Sequence-to-Sequence Tasks”. In: *Proceedings of the International Conference on Learning Representations (ICLR)*. Available at: https://openreview.net/forum?id=Bd7GueaTxUz. 2023. DOI: 10.48550/arXiv.2210.07499.

[54] Kaiming He et al. “Deep Residual Learning for Image Recognition”. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016, pp. 770–778. DOI: 10.1109/CVPR.2016.90.

[55] Tsung-Yi Lin et al. “Feature Pyramid Networks for Object Detection”. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE. 2017, pp. 936–944. DOI: 10.1109/CVPR.2017.106.

[56] Wenhai Wang et al. “Shape Robust Text Detection with Progressive Scale Expansion Network”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019, pp. 9336–9345. DOI: 10.1109/CVPR.2019.00956.

[57] Zhaohui Zheng et al. “Distance-IoU Loss: Faster and Better Learning for Bounding Box Regression”. In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 34. 2020, pp. 12993–13000. DOI: 10.1609/aaai.v34i07.6999.

References

[58] Alex Graves et al. "Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks". In: *Proceedings of the 23rd international conference on Machine learning*. 2006, pp. 369–376. DOI: 10.1145/1143844.1143891.

[59] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. "V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation". In: *2016 fourth international conference on 3D vision (3DV)*. Ieee. 2016, pp. 565–571. DOI: 10.1109/3DV.2016.79.

[60] J.R. Rohlicek et al. "Continuous hidden Markov modeling for speaker-independent word spotting". In: *International Conference on Acoustics, Speech, and Signal Processing*, vol. 1. 1989, pp. 627–630. DOI: 10.1109/ICASSP.1989.266505.

[61] F.R. Chen, L.D. Wilcox, and D.S. Bloomberg. "Word spotting in scanned images using hidden Markov models". In: *1993 IEEE International Conference on Acoustics, Speech, and Signal Processing*. Vol. 5. 1993, pp. 1–4. DOI: 10.1109/ICASSP.1993.319732.

[62] R. Manmatha, Chengfeng Han, and E.M. Riseman. "Word spotting: a new approach to indexing handwriting". In: *Proceedings CVPR IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. 1996, pp. 631–637. DOI: 10.1109/CVPR.1996.517139.

[63] Rothacker, Leonard and Sudholt, Sebastian and Rusakov, Eugen and Kasperidus, Matthias and Fink, Gernot A. "Word Hypotheses for Segmentation-Free Word Spotting in Historic Document Images". In: *Proceedings of the 14^{th} IAPR International Conference on Document Analysis and Recognition (ICDAR)*. 2017, pp. 1174–1179. DOI: 10.1109/ICDAR.2017.194.

References

[64] Suman K. Ghosh and Ernest Valveny. “R-PHOC: Segmentation-Free Word Spotting Using CNN”. In: *Proceedings of the 14^{th} IAPR International Conference on Document Analysis and Recognition (ICDAR)*. 2017, pp. 801–806. DOI: 10.1109/ICDAR.2017.136.

[65] Irina Rabaev, Klara Kedem, and Jihad El-Sana. “Keyword Retrieval Using Scale-Space Pyramid”. In: *Proceedings of the 12^{th} IAPR Workshop on Document Analysis Systems (DAS)*. 2016, pp. 144–149. DOI: 10.1109/DAS.2016.16.

[66] Konstantinos Zagoris, Ioannis Pratikakis, and Basilis Gatos. “Unsupervised Word Spotting in Historical Handwritten Document Images Using Document-Oriented Local Features”. In: *IEEE Transactions on Image Processing* 26.8 (2017), pp. 4032–4041. DOI: 10.1109/TIP.2017.2700721.

[67] Yann Leydier et al. “Towards an Omnilingual Word Retrieval System for Ancient Manuscripts”. In: *Pattern Recognition* 42.9 (2009), pp. 2089–2105.

[68] B. Gatos and I. Pratikakis. “Segmentation-free word spotting in historical printed documents”. In: *Proceedings of the 10^{th} International Conference on Document Analysis and Recognition (ICDAR)*. 2009, pp. 271–275.

[69] Xi Zhang and Chew Lim Tan. “Handwritten word image matching based on Heat Kernel Signature”. In: *Pattern Recognition* 48.11 (2015), pp. 3346–3356.

[70] M. Rusiñol et al. “Efficient segmentation-free keyword spotting in historical document collections”. In: *Pattern Recognition* 48.2 (2015), pp. 545–555.

[71] H. Jegou, M. Douze, and C. Schmid. “Product Quantization for Nearest Neighbor Search”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 33.1 (2011), pp. 117–128.

References

[72] Pau Riba, Josep Llados, and Alicia Fornes. “Handwritten Word Spotting by Inexact Matching of Grapheme Graphs”. In: *Proceedings of the 13^{th} International Conference on Document Analysis and Recognition (ICDAR)*. 2015, pp. 781–785.

[73] Huaigu Cao, Venu Govindaraju, and Anurag Bhardwaj. “Unconstrained handwritten document retrieval”. In: *International Journal on Document Analysis and Recognition* 14.2 (2011), pp. 145–157.

[74] A. L. Kesidis et al. “A word spotting framework for historical machine-printed documents”. In: *International Journal on Document Analysis and Recognition* 14.2 (2011), pp. 131–144.

[75] T. Konidaris et al. “Keyword-guided word spotting in historical printed documents using synthetic data and user feedback”. In: *International Journal on Document Analysis and Recognition* 9.2-4 (2007), pp. 167–177.

[76] Marcal Rusiñol and Josep Llados. “Boosting the handwritten word spotting experience by including the user in the loop”. In: *Pattern Recognition* 47.3 (2014), pp. 1063–1072.

[77] Fabian Wolf, Philipp Oberdiek, and Gernot A. Fink. “Exploring Confidence Measures for Word Spotting in Heterogeneous Datasets”. In: *CoRR* abs/1903.10930 (2019). arXiv: `1903.10930`. URL: `http://arxiv.org/abs/1903.10930`.

[78] A. Bhardwaj, D. Jose, and V. Govindaraju. “Script independent word spotting in multilingual documents”. In: *Proceedings of the 2^{nd} Workshop on Cross Lingual Information Access (CLIA)*. 2008, pp. 48–54.

[79] Ekta Vats, Anders Hast, and Alicia Fornés. “Training-Free and Segmentation-Free Word Spotting using Feature Matching and Query Expansion”. In: *Proceedings of the 15^{th} International Conference on Document Analysis and Recognition (ICDAR)*. 2019, pp. 1294–1299. DOI: `10.1109/ICDAR.2019.00209`.

References

[80] R. Shekhar and C.V. Jawahar. “Word Image Retrieval Using Bag of Visual Words”. In: *Proceedings of the 10^{th} IAPR International Workshop on Document Analysis Systems (DAS)*. 2012, pp. 297–301.

[81] SumanK. Ghosh and Ernest Valveny. “A Sliding Window Framework for Word Spotting Based on Word Attributes”. In: *Proceedings of the 7^{th} Iberian Conference on Pattern Recognition and Image Analysis (PRAI)*. 2015, pp. 652–661.