

Wordspotting in historischen Dokumenten

Gernot A. Fink

TU Dortmund, Fakultät für Informatik

Herne, 18. Februar 2017

- ▶ Einführung *Warum Wordspotting?*
- ▶ Wordspotting *Grundlagen und Methoden*
- ▶ Tiefe neuronale Netze *... der Stand-der-Kunst*
- ▶ Zusammenfassung *... und ein Anwendungsbeispiel*

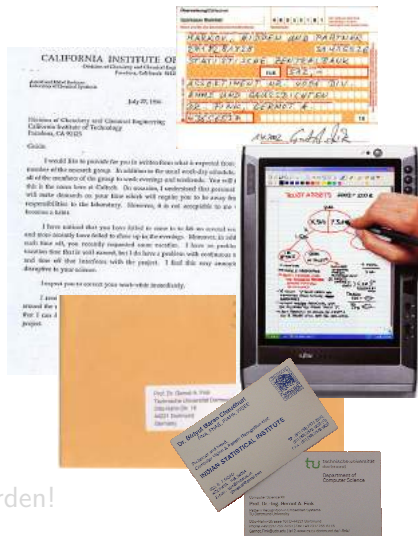
Mit Beiträgen von:

Leonard Rothacker, Sebastian Sudholt

Einführung: Systeme zum Maschinellen Lesen

Stand des Maschinellen Lesens:

- ▶ Eines der ersten Anwendungsgebiete der Informatik
- ▶ Ex. OCR-Systeme für maschinell gedruckte Texte (inkl. Fraktur)
- ▶ Ex. *online* Handschrifterkennung (auf SmartPhones, Tablets, ...)
- ▶ *Offline* Handschrifterkennung: *offenes Forschungsproblem*



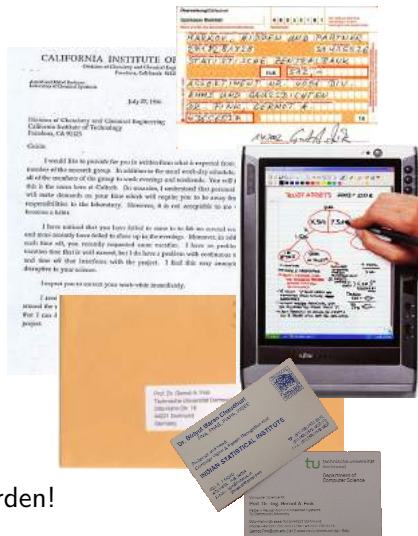
Generelle Methodik/Problematik:

Transkriptionsmodelle müssen auf *beträchtlichen* Mengen von *annotierten* Beispieldaten trainiert werden!

Einführung: Systeme zum Maschinellen Lesen

Stand des Maschinellen Lesens:

- ▶ Eines der ersten Anwendungsgebiete der Informatik
- ▶ Ex. OCR-Systeme für maschinell gedruckte Texte (inkl. Fraktur)
- ▶ Ex. *online* Handschrifterkennung (auf Smartphones, Tablets, ...)
- ▶ *Offline* Handschrifterkennung: *offenes Forschungsproblem*



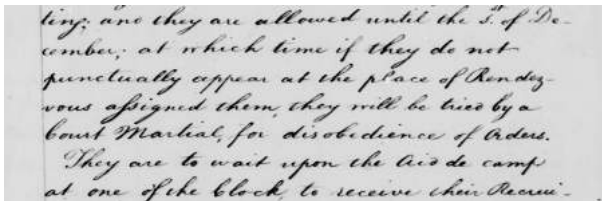
Generelle Methodik/Problematik:

Transkriptionsmodelle müssen auf *beträchtlichen* Mengen von *annotierten* Beispieldaten trainiert werden!

Einführung: Warum Wordspotting?

Was wenn die automatische Transkription von Handschrift nicht länger möglich ist?

... lassen Sie es uns etwas einfacher machen!



Alternative: Suche ("Retrieval") einzelner Wörter im Gegensatz zu Transkription ("query-by-example")

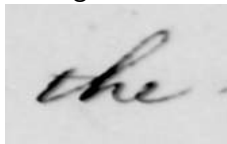
Bilder aus *The George Washington Papers at the Library of Congress, 1741-1799*

Einführung: Warum Wordspotting?

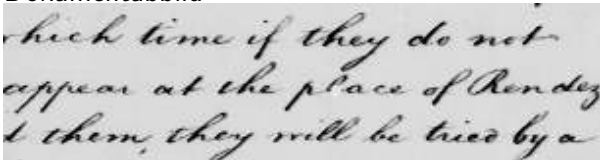
Was wenn die automatische Transkription von Handschrift nicht länger möglich ist?

Alternative: Suche ("Retrieval") einzelner Wörter im Gegensatz zu Transkription ("query-by-example")

Anfrage-Wortabbild



Dokumentabbild



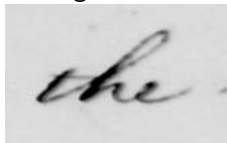
Bilder aus *The George Washington Papers at the Library of Congress, 1741-1799*

Einführung: Warum Wordspotting?

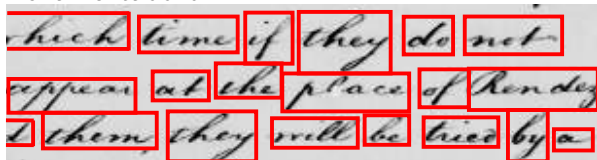
Was wenn die automatische Transkription von Handschrift nicht länger möglich ist?

Alternative: Suche ("Retrieval") einzelner Wörter im Gegensatz zu Transkription ("query-by-example")

Anfrage-Wortabbild



Dokumentabbild

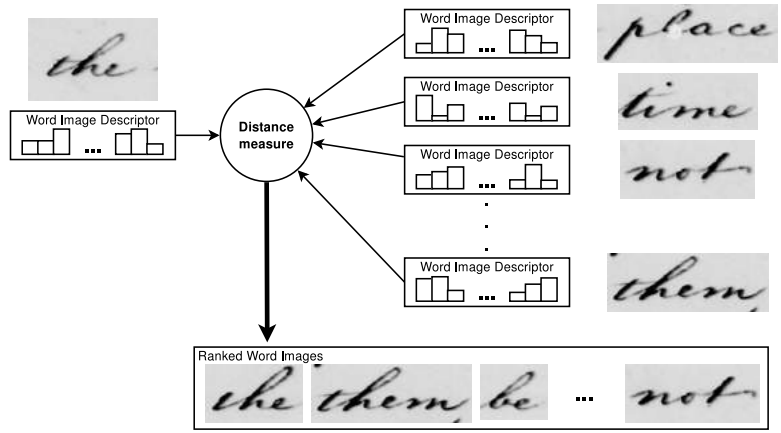


Bilder aus *The George Washington Papers at the Library of Congress, 1741-1799*

Einführung: Warum Wordspotting?

Query-by-example Wordspotting

Woher kennen wir das?



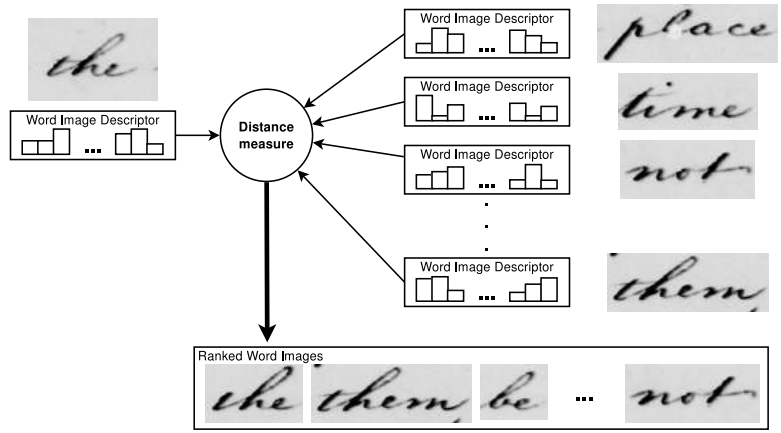
Bilder aus *The George Washington Papers at the Library of Congress, 1741-1799*

Nach [Rath & Manmatha, IJRR'07]

Einführung: Warum Wordspotting?

Query-by-example Wordspotting

Woher kennen wir das?



Bilder aus *The George Washington Papers at the Library of Congress, 1741-1799*

Nach [Rath & Manmatha, IJRR'07]

Einführung: Warum Wordspotting?

QbE Wordspotting \approx Spezialfall der Inhaltsbasierten Bildsuche

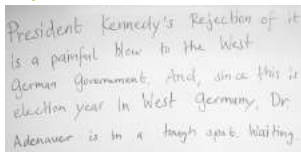


(Quelle: The Chinese University of Hong Kong)

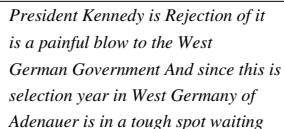
Einführung: Transkription vs. Suche

Einführung: Transkription vs. Suche

Transkription von Dokumenten (= "klassische" Erkennung)



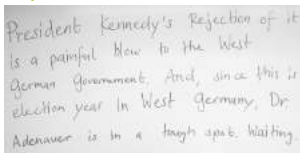
President Kennedy's Rejection of it
is a painful blow to the West
German Government. And, since this is
election year in West Germany, Dr.
Adenauer is in a tough spot, waiting.



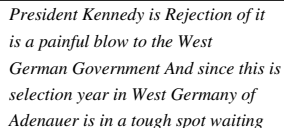
*President Kennedy is Rejection of it
is a painful blow to the West
German Government And since this is
selection year in West Germany of
Adenauer is in a tough spot waiting*

Einführung: Transkription vs. Suche

Transkription von Dokumenten (= "klassische" Erkennung)

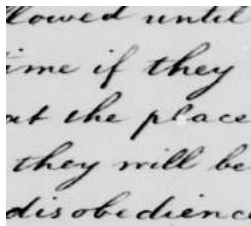


President Kennedy's Rejection of it is a painful blow to the West German Government. And, since this is election year in West Germany, Dr Adenauer is in a tough spot waiting.

President Kennedy is Rejection of it is a painful blow to the West German Government And since this is selection year in West Germany of Adenauer is in a tough spot waiting

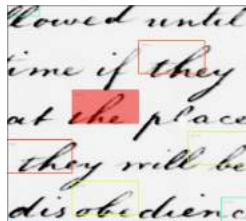
Suche in Dokumenten (aka "Wordspotting")



lowed until
time if they
at the place
they will be
disobedienc



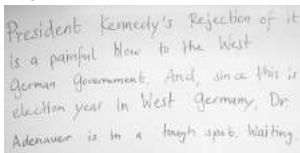
the

lowed until
time if they
at the place
they will be
disobedienc

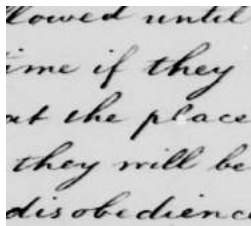
Einführung: Transkription vs. Suche

Transkription von Dokumenten (= "klassische" Erkennung)



President Kennedy is Rejection of it is a painful blow to the West German Government And since this is selection year in West Germany of Adenauer is in a tough spot waiting

Suche in Dokumenten (aka "Wordspotting")

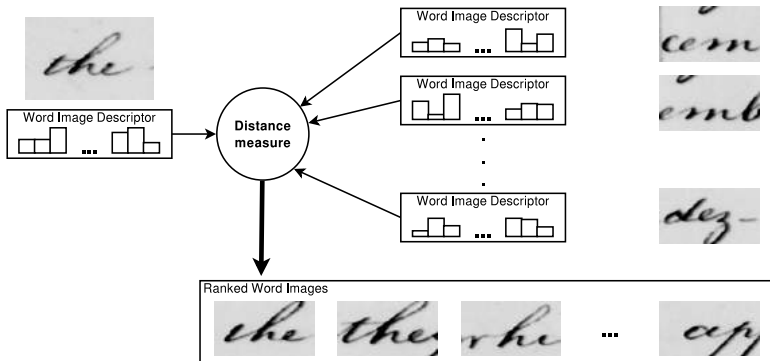
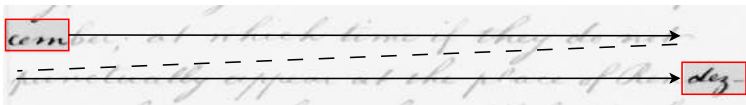


the



Wordspotting: Segmentierung?

Segmentierungsfreies Wordspotting mit gleitendem Suchfenster



Wordspotting: Schriftsysteme?

Prinzipielle Methodik praktisch universell einsetzbar!

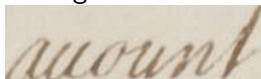
Anfragebild

Dokumentabbild

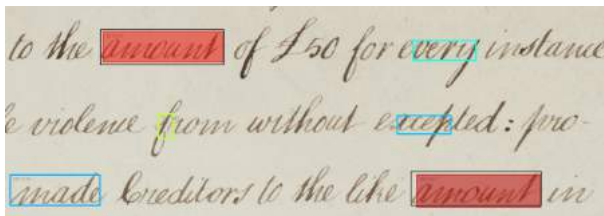
Wordspotting: Schriftsysteme?

Prinzipielle Methodik praktisch universell einsetzbar!

Anfragebild



Dokumentabbild

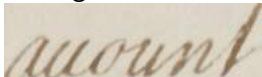


(Quelle: UCL Special Collections, Bentham Papers)

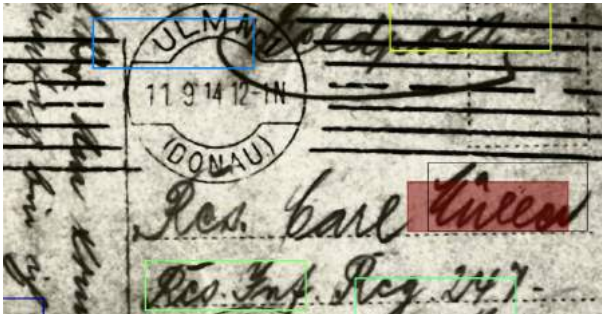
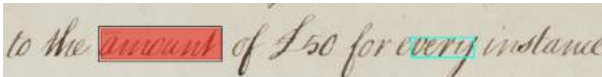
Wordspotting: Schriftsysteme?

Prinzipielle Methodik praktisch universell einsetzbar!

Anfragebild



Dokumentabbild



(Quelle: Feldpostkarte, Privatsammlung Dr. Britta Bley, Dortmund)

Wordspotting: Schriftsysteme?

Prinzipielle Methodik praktisch universell einsetzbar!

Anfragebild

Dokumentabbild

প্রবাহিত হইতেছে। পশ্চিম এবং দক্ষিণ বঙ্গে
 ঠর সময়ে যে শিবের "গাজন" হইয়া থাকে, তাহাঁহ
 বা" নামে পরিচিত। এই গাজন বা গম্ভীরায়

(Quelle: H. Palit. Aadyer Gambhira. Krishnacharan Sarkar, Maldaha, 1913)

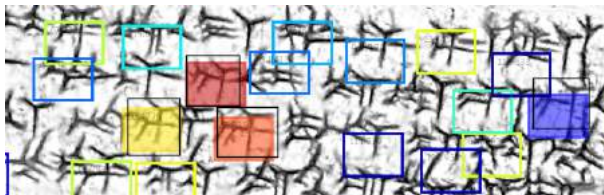
Wordspotting: Schriftsysteme?

Prinzipielle Methodik praktisch universell einsetzbar!

Anfragebild

Dokumentabbild

(Quelle: H. Palit. Aadyer Gambhira. Krishnacharan Sarkar, Maldaha, 1913)



(Keilschrifttafel, Hattusa, 30.000 v. Chr., Quelle: Hethitologie Portal, Mainz)

Wordspotting: Zwischenfazit

Methode betrachtet bisher: *Query-by-Example Wordspotting*

Wesentlicher Vor- / Nachteil:

- ✓ Erfordert kein Trainingsmaterial / keine Annotationen!
- ⚡ Anfragen können nur *ausgewählt* werden.

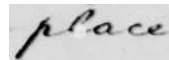
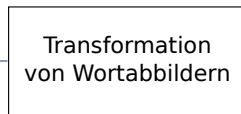
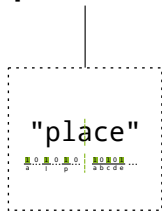
Was ist mit beliebigen Anfragen?

Anforderung: Erzeugung eines Schriftmodells aus einer *textuellen* Eingabe (= "string")
⇒ *Query-by-String Wordspotting*

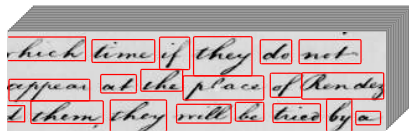
Wordspotting: Query-by-String

Idee: Gemeinsame Repräsentation für Schrift und Bild

"place"



↑ maschinelles Lernen

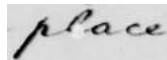
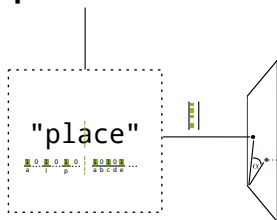


annotierte Stichprobe

Wordspotting: Query-by-String

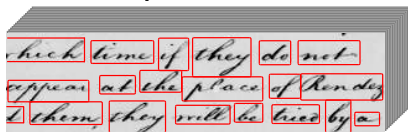
Idee: Gemeinsame Repräsentation für Schrift und Bild

"place"



Transformation
von Wortabbildern

↑ maschinelles Lernen



annotierte Stichprobe

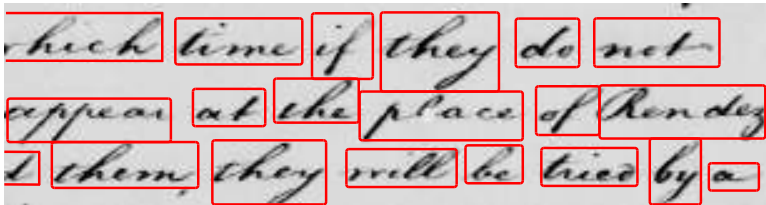
- ✓ Vergleich einfach möglich
- ⚡ Transformation für Bilder muss aus Beispieldaten gelernt werden

⇒ Annotierte Stichprobe erforderlich!

Wordspotting: Query-by-String

Idee: Gemeinsame Repräsentation für Schrift und Bild

Beispielhafte Annotation auf Wortebene

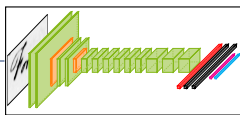
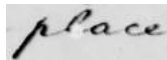
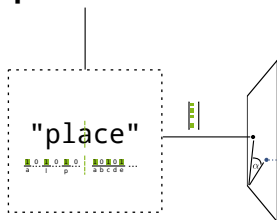


which time if they do not
appear at the place of Rendez
them they will be tied by a

Wordspotting: Query-by-String

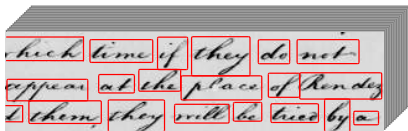
Idee: Gemeinsame Repräsentation für Schrift und Bild

"place"



↑ maschinelles Lernen

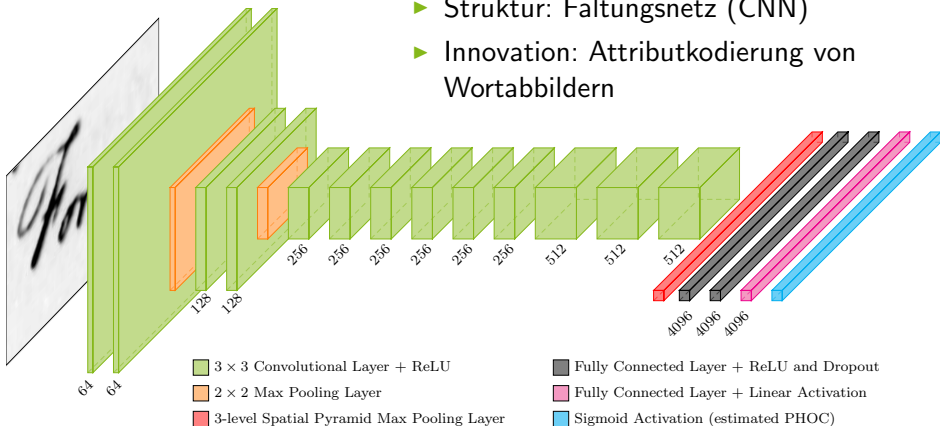
Lösung: Training eines tiefen neuronalen Netzwerks!



annotierte Stichprobe

Wordspotting: Tiefe Neuronale Netze

- ▶ Struktur: Faltungsnetz (CNN)
- ▶ Innovation: Attributkodierung von Wortabbildern



(Sudholt & Fink: ICFHR, 2016, Gewinner des *Best Paper Award!*)

Anwendungsbeispiel Wordspotting

Activities Main.py Wed 12:39 Wordspotting - washington Julian K. rby

File

2700270.png

2710271.png

270 *Litters Orders and Instructions, October 1755.*

only for the public use, until by particu- lar orders from me. You are to send down a Barrel of Hens with the Arms to Winchester, and about two thousand weight of Flour, for the two Companies of Rangers; twelve hundred of which to be delivered **Captain** . . . tibly one company, at the Plantation of Charles Litters - the rest to your **into company** at . . . Nicholas Reames.

October 26. 1755

Winchester, October 28 1755. **Parde Hampton:**

The officers who come down from Fort Crumland with Colonel Washington, are immediately to go down - they are allowed until the 1. of December, at which time if they do not punctually appear at the place of Rendezvous assigned them they will be tried by a Court-Martial for disobedience of Orders. They are to wait upon the first of each at one of the blocks to receive their Respective Instructions. Each Officer present to give in a Return immediately of the number of men he has **at block** - line 24 below

Query Info Query Results

Captain 2700270

Captain's 2710271

Captain 2710271

e captain 2710271

Captain 2710271

Company 2700270

Company 2710271

Show Result Patches

Show Heatmap

2 of 2 Documents processed

Zusammenfassung

Grundmethode: *Query-by-Example Wordspotting*

- ✓ Gute Leistungsfähigkeit auf **isogenen** Dokumenten
- ✓ Kein Annotationsaufwand erforderlich!
- ⚡ Schlechte Generalisierung auf verschiedene Schriftstile

Alternative: *Query-by-String Wordspotting*

- ✓ Jede (textuelle) Anfrage kann verwendet werden
- ⊛ Erfordert **annotiertes** Trainingsmaterial!
- ✓ Stand-der-Technik: *Tiefe Neuronale Netze*

Zusammenfassung

Grundmethode: *Query-by-Example Wordspotting*

- ✓ Gute Leistungsfähigkeit auf **isogenen** Dokumenten
- ✓ Kein Annotationsaufwand erforderlich!
- ⚡ Schlechte Generalisierung auf verschiedene Schriftstile

Alternative: *Query-by-String Wordspotting*

- ✓ Jede (textuelle) Anfrage kann verwendet werden
- ⊘ Erfordert **annotiertes** Trainingsmaterial!
- ✓ Stand-der-Technik: *Tiefe Neuronale Netze*

