Digitalisierung

Vorlesung *Einführung in die Digital Humanities*
MSc Digital Humanities | Wintersemester 2020/21

Prof. Dr. Christof Schöch

Universität Trier

Einstieg

Sitzungsüberblick

1. Digitalisierung allgemein
2. Digitalisierung von Bildern
3. Digitalisierung von Text (OCR)

1. Digitalisierung allgemein

Was ist Digitalisierung?

Der Begriff Digitalisierung bezeichnet allgemein die Veränderungen von Prozessen, Objekten und Ereignissen, die bei einer zunehmenden Nutzung digitaler Geräte erfolgt. Im ursprünglichen und engeren Sinne ist dies die Erstellung digitaler Repräsentationen von physischen Objekten, Ereignissen oder analogen Medien. ("Digitalisierung", Wikipedia)

Digitization -- the conversion of an analogue signal or code into a digital signal or code -- is the bedrock of both digital library holdings and digital humanities research. (Melissa Terras 2012)

Warum überhaupt digitalisieren?

- Bessere Verfügbarkeit (ortsunabhängig, gleichzeitig)
- Bessere Zugänglichkeit (umfangreiche Bände, große Karten)
- virtuelle Zusammenführung dislozierter Bestände
- Konservierung und Schonung fragiler/wertvoller Objekte
- Aufbereitung des digitalen Bildes (bspw. Bildschärfe, Kontrast, Zoom etc.)
- Verbesserte Analysemöglichkeiten (bspw. Volltext)
- Integration in Unterrichtsmaterial
- Verlustfreies Kopieren möglich

Warum *nicht* digitalisieren?

- Kosten für Digitalisierung (Sach- und Personalkosten)
- Vollständige Digitalisierung der menschlichen Überlieferung unrealistisch
- Eventuelle Beschädigung fragiler / wertvoller Objekte
- Digitalisate können Original nicht immer ersetzen
- Folgekosten für Archivierung / Langzeitverfügbarkeit

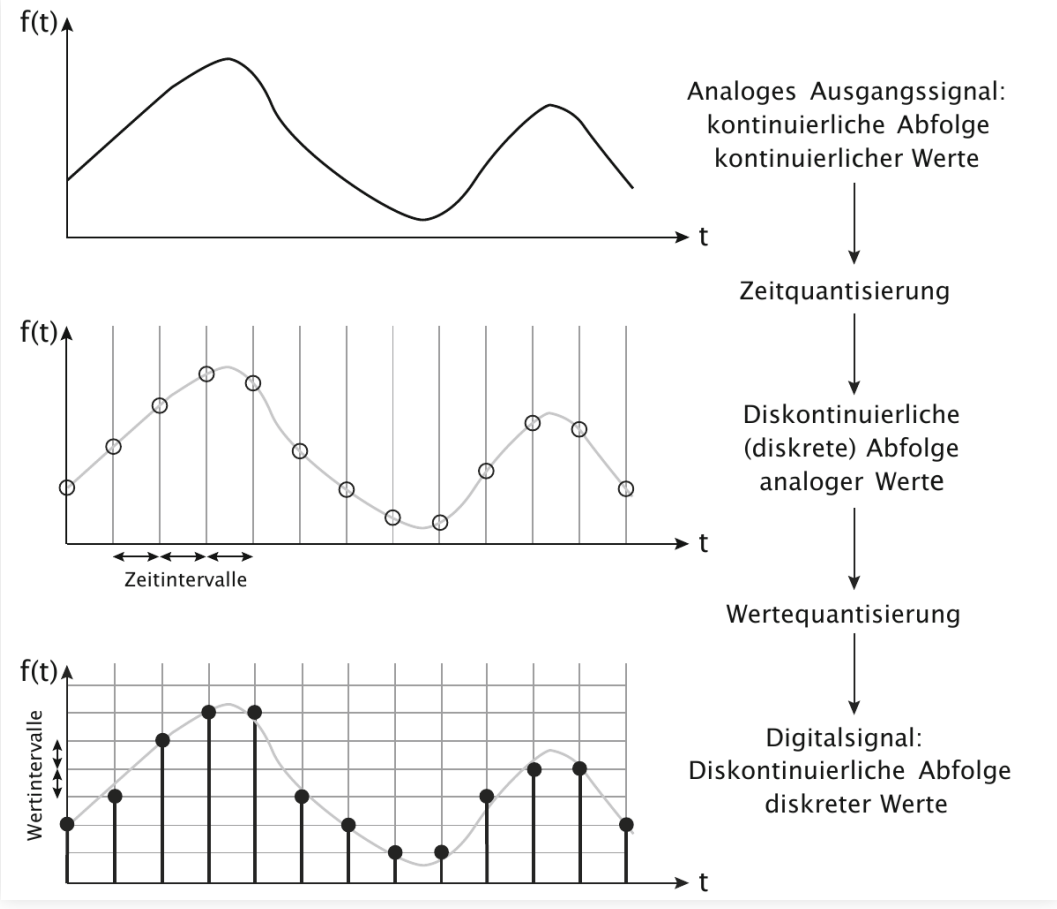
Begrifflichkeit

- Digitalisat
- Digitale Repräsentationen
- Digitales Abbild, Modell, Surrogat
- analog (von gr. *analogos*, verhältnismäßig)
- digital (von lat. *digitus*, Finger)
- Analog-Digital-Wandler

Analog vs. digital

- analog = "zeit-/raum- und wertekontinuierlich"
- digital = zeit-/raum- und wertediskret"
- Digitalisierung
  - Zeit-/Raumquantisierung (mit bestimmter Auflösung)
  - Wertquantisierung (mit bestimmter Abtastrate)

Quantisierung



(Bildquelle: Rehbein 2017)
Beispiele: Tonhöhenverlauf; Dicke eine Vase; Schwarzwertverteilung einer Druckseite

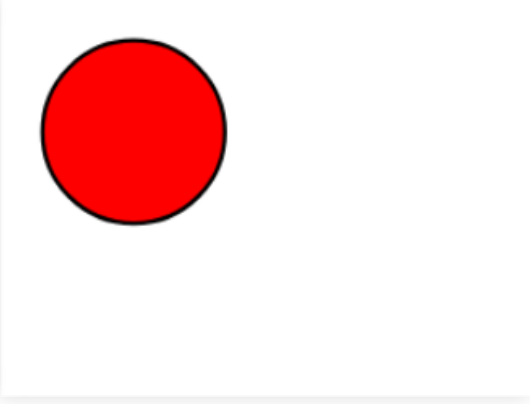
2. Digitalisierung von Bildern

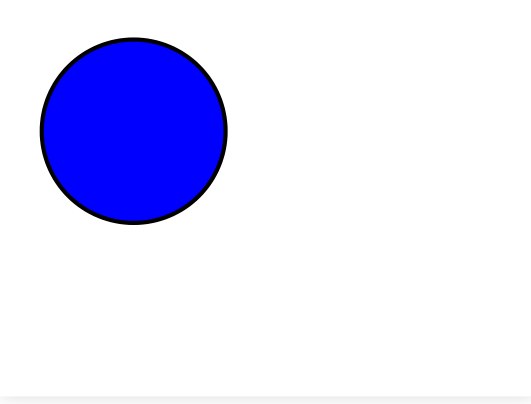
(A) Digitale Bildformate

Digitale Bildformate

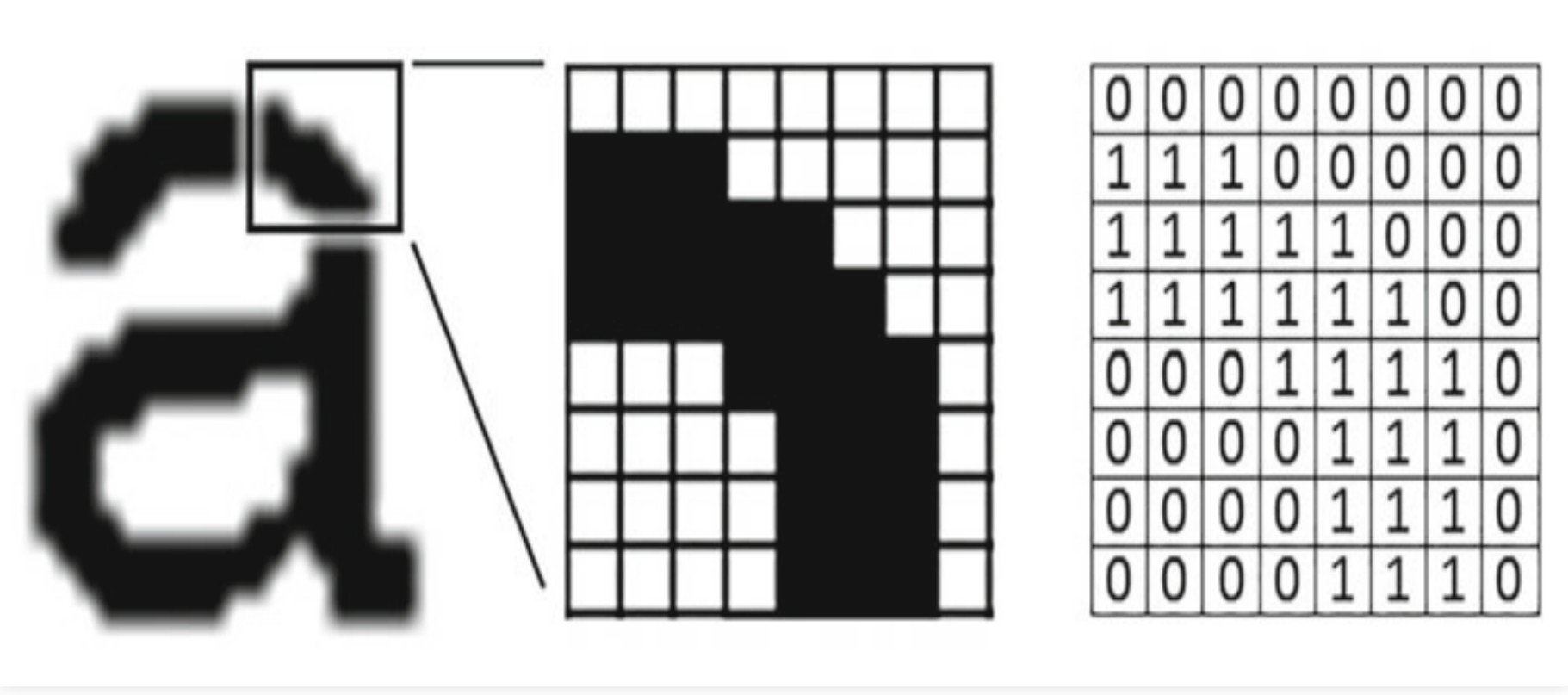
- **Rastergrafik**: "zweidimensionale Matrix der Bildpunkte (Pixel)"
  vs.
- **Vektorgrafik**: Repräsentation als Menge "elementarer Zeichenroutinen"

Raster- und Vektorgrafik





Prinzip der Rastergrafik



(Bildquelle: Rehbein 2017)
Parameter: Bildgröße (Pixel) und Farbtiefe (Bit)

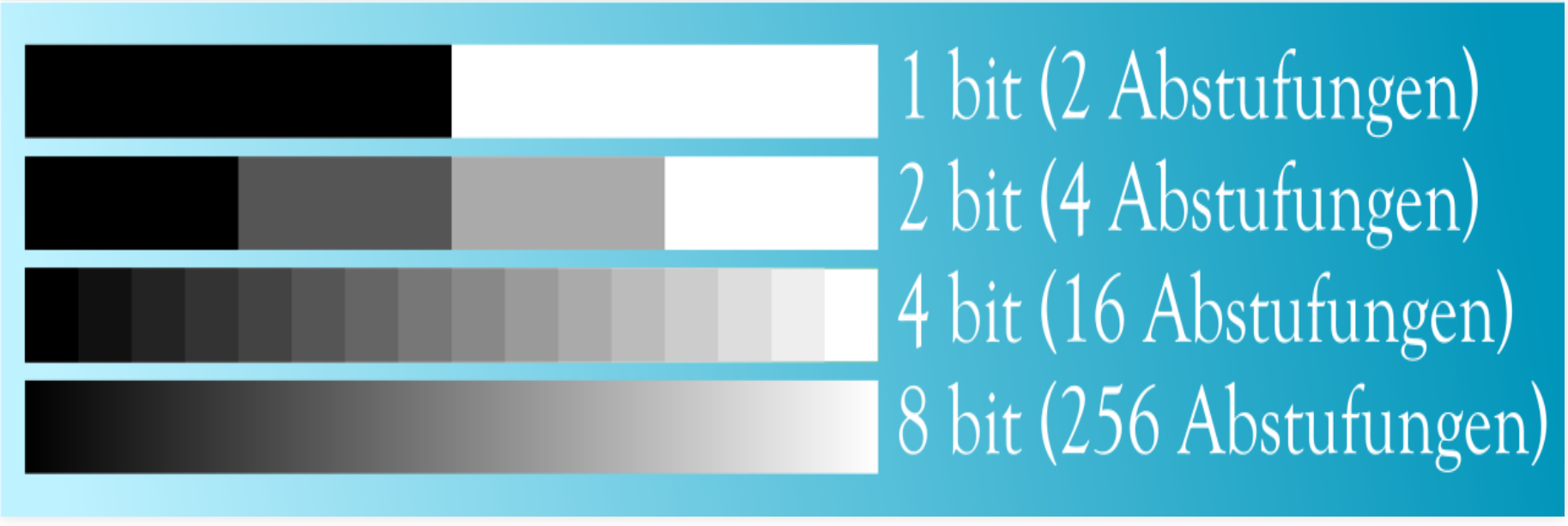
Parameter Bildgröße

- Bildgröße = Pixelanzahl: bspw. 8x8 Pixel / 64 Pixel
- Bildfläche = ergibt sich aus Bildgröße (in Pixel) und Pixelgröße
- Auflösung = Pixeldichte, also Anzahl Pixel pro Fläche (bspw. dpi oder ppi)
- Je geringer die Pixelgröße bei gleicher Bildfläche ist, desto höher ist die Pixeldichte

Parameter Farbtiefe

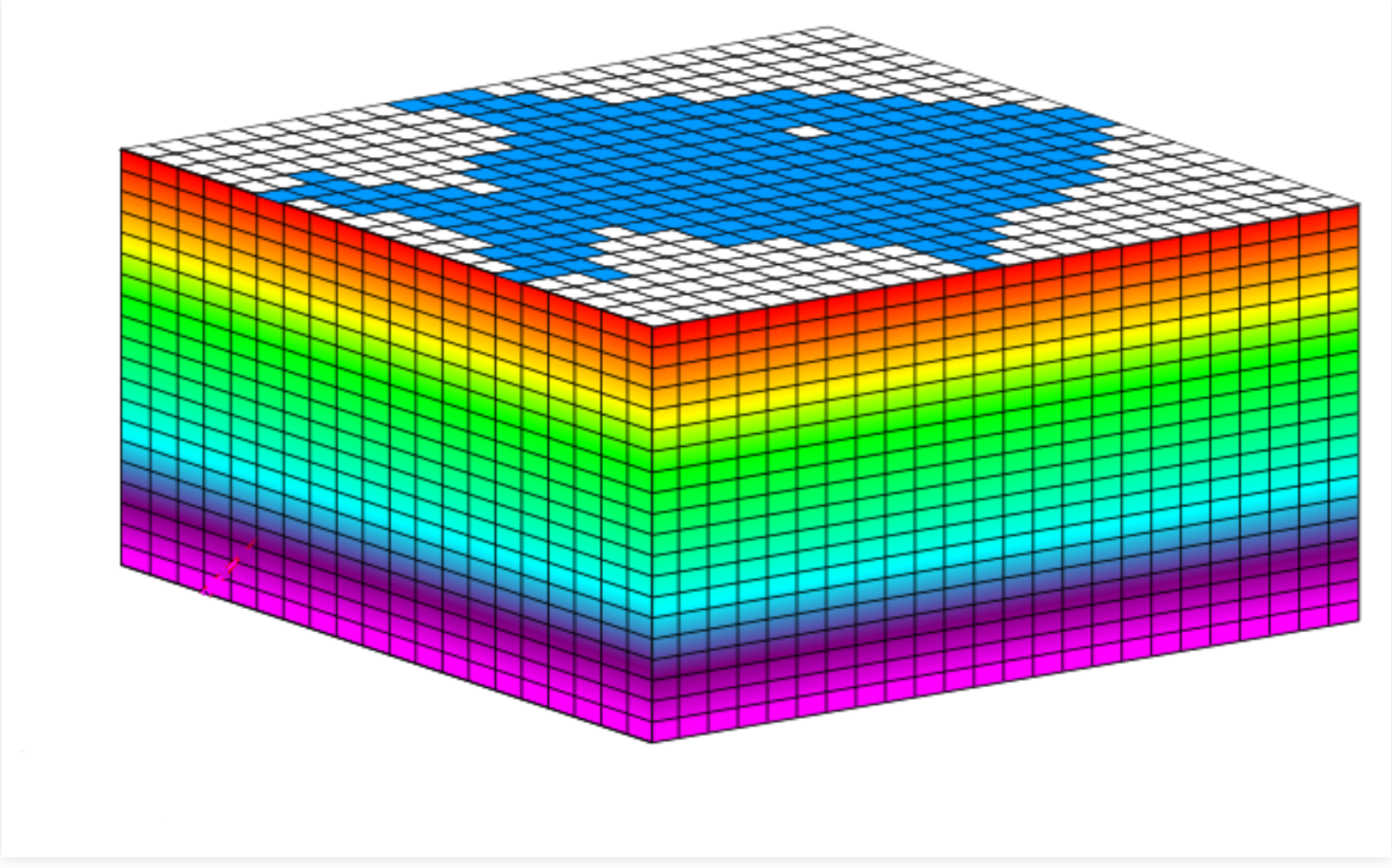
- monochrom
  - 1 Bit: 2 Werte: s/w
  - 2 Bit: 2^2 Werte: 4 Graustufen
  - 8 Bit: 2^8 Werte = 256 Graustufen
- farbig (3 Kanäle, RGB)
  - 3 x 8 Bit = (2^8)^3 Werte = ca. 16 Millionen Farben

Farbtiefe

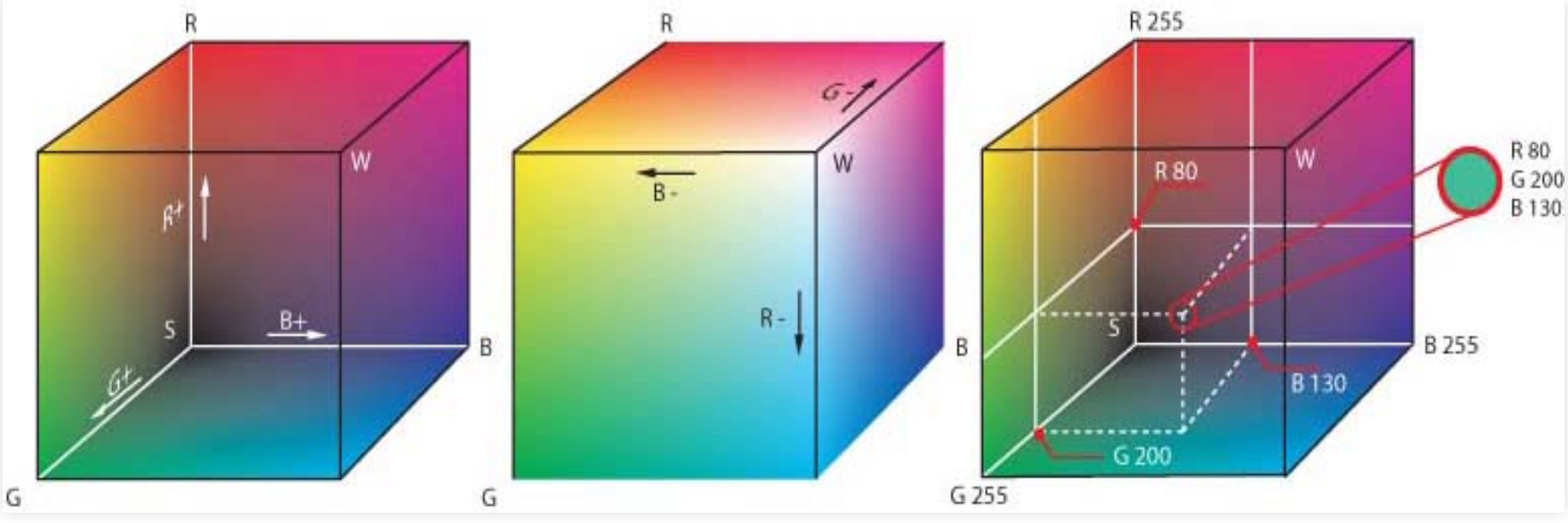


Bildquelle: Thomas R. Schwarz, Wikipedia, https://de.wikipedia.org/wiki/Farbtiefe_(Computergrafik)#/media/File:Farbtiefe.svg, gemeinfrei.

Warum Farbtiefe?

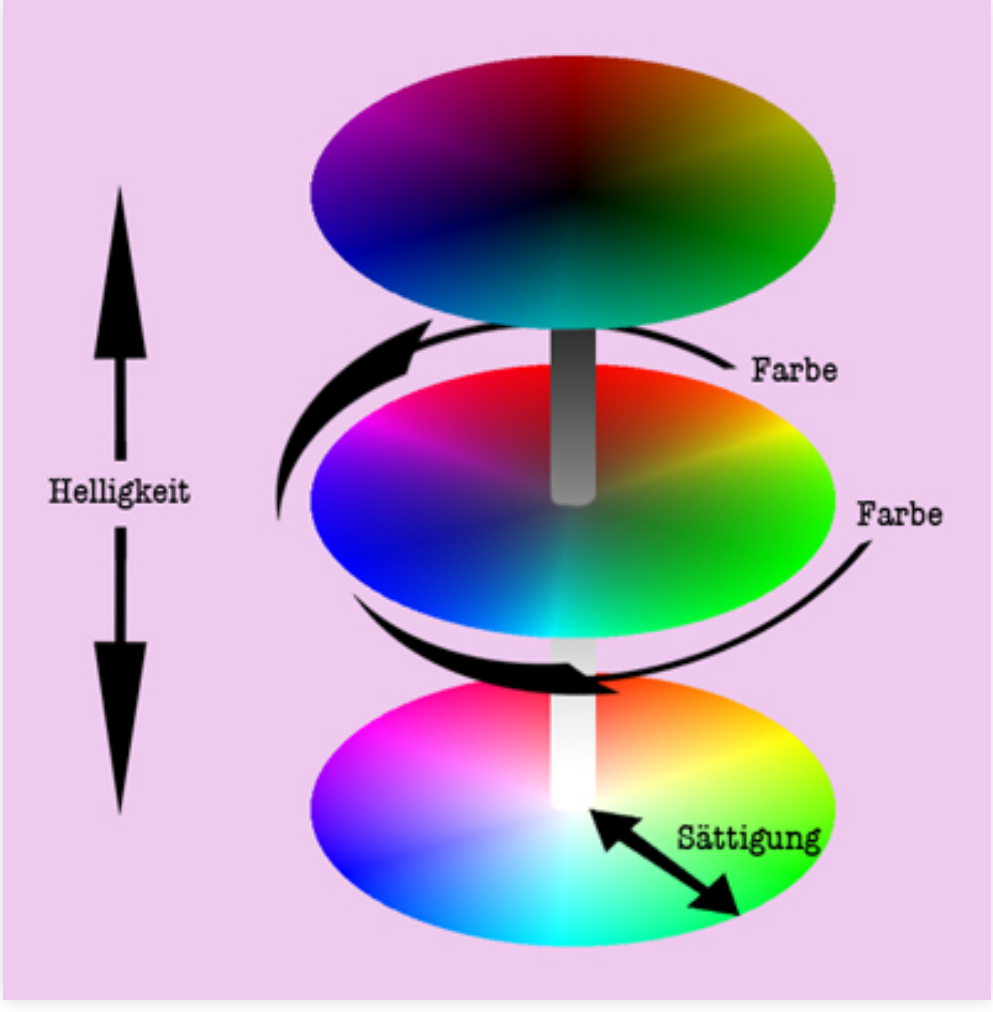


Farbräume: RGB



RGB = Red, Green, Blue. Bildquelle: https://commons.wikimedia.org/wiki/File:RGB_farbwuerfel.jpg, Lizenz CC BY-SA 3.0

Farbräume: HSL



HSL = Hue, Saturation, Lightness. Bildquelle: Benutzer Friedrich Graf, https://de.wikipedia.org/wiki/Farbraum#/media/Datei:Farbstruktur.jpg, Lizenz CC BY-SA 2.0

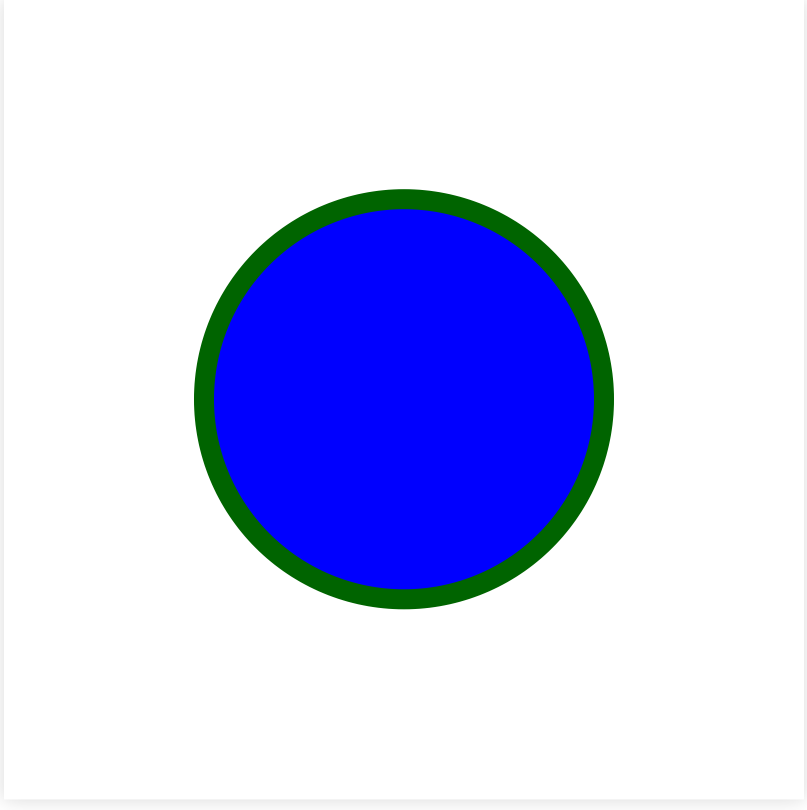
Vektorgrafiken (SVG)

- SVG = Scalable Vector Grafics
- Nicht pixelbasiert
- verwendet "grafische Primitive" in einem Koordinatensystem
- Primitive
  - Punkte, Linien, Ellipsen, Polygone, Rechteck
  - Farbe, Strichform, Füllung
- Koordinatensystem
  - meist zweidimensional (x und y-Achse)
  - z-Ordnung: Reihenfolge bei Überlappung

mycircle.svg

```
<!--?xml version="1.0"?-->
<svg width="400" height="400" xmlns="http://www.w3.org/2000/svg">
  <title>my circle</title>
  <g>
    <ellipse cx="200" cy="200" rx="100" ry="100" fill="darkblue" stroke=
  </g>
</svg>
```

mycircle.svg

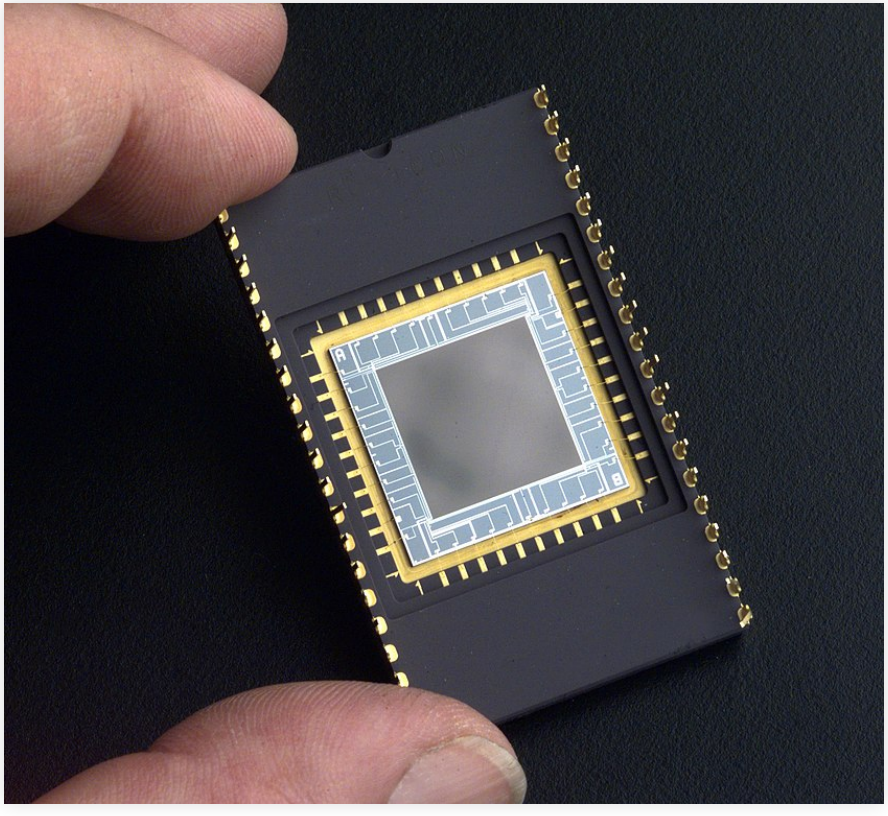


(B) Bilddigitalisierung

Bilddigitalisierung

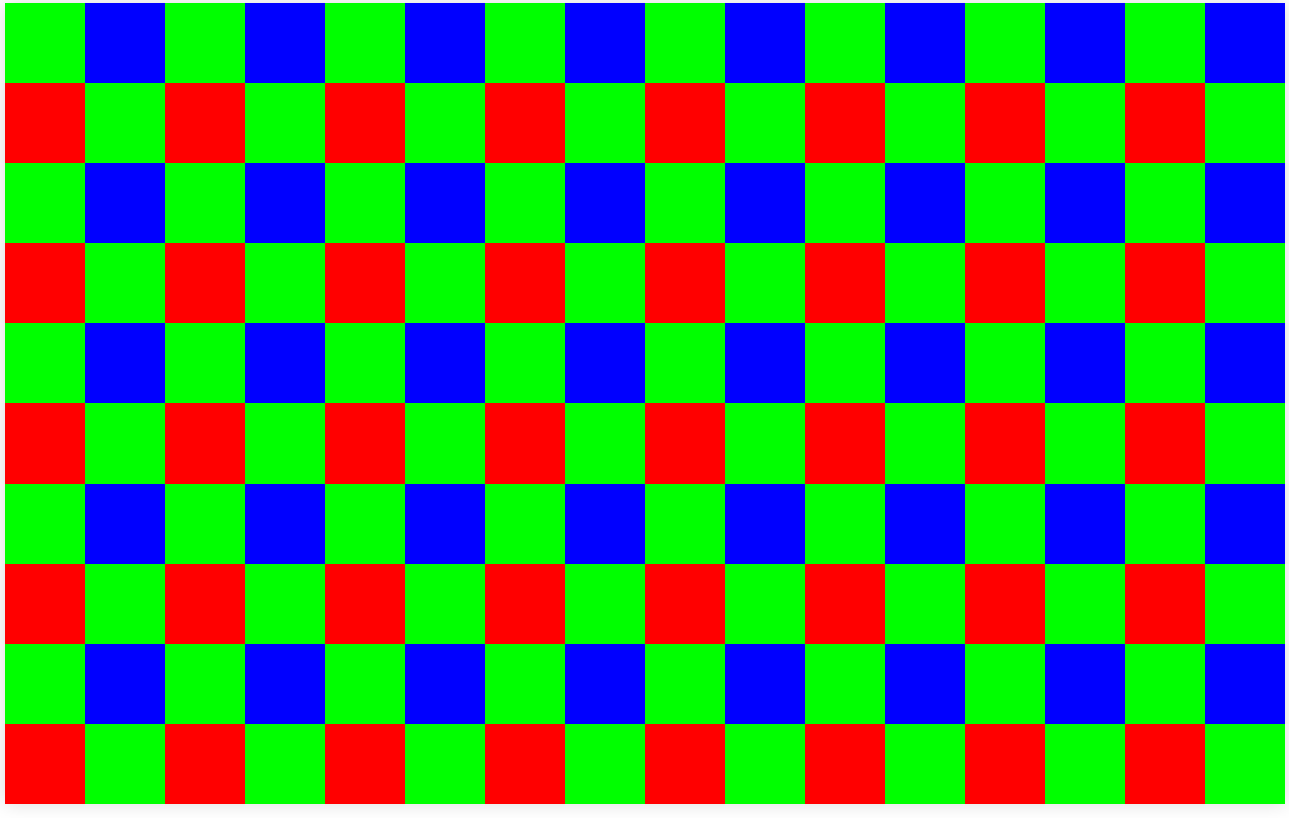
- Erstellen einer visuellen, digitalen Repräsentation
- Mit Digitalkamera oder Scanner
- Bildgebendes Prinzip: lichtempfindliche Dioden
- Analog-Digital-Wandler
  - Anzahl der Dioden = Bildgröße
  - Dioden pro Fläche = Auflösung
  - Anzahl der registrierten Farbwerte = Farbtiefe

Bildsensor einer Digitalkamera



Bildquelle: NASA, Wikipedia, https://de.wikipedia.org/wiki/Datei:Delta-Doped_Charged_Coupled_Devices_(CCD)_for_Ultra-Violet_and_Visible_Detection.jpg, gemeinfrei.

Bayer-Muster



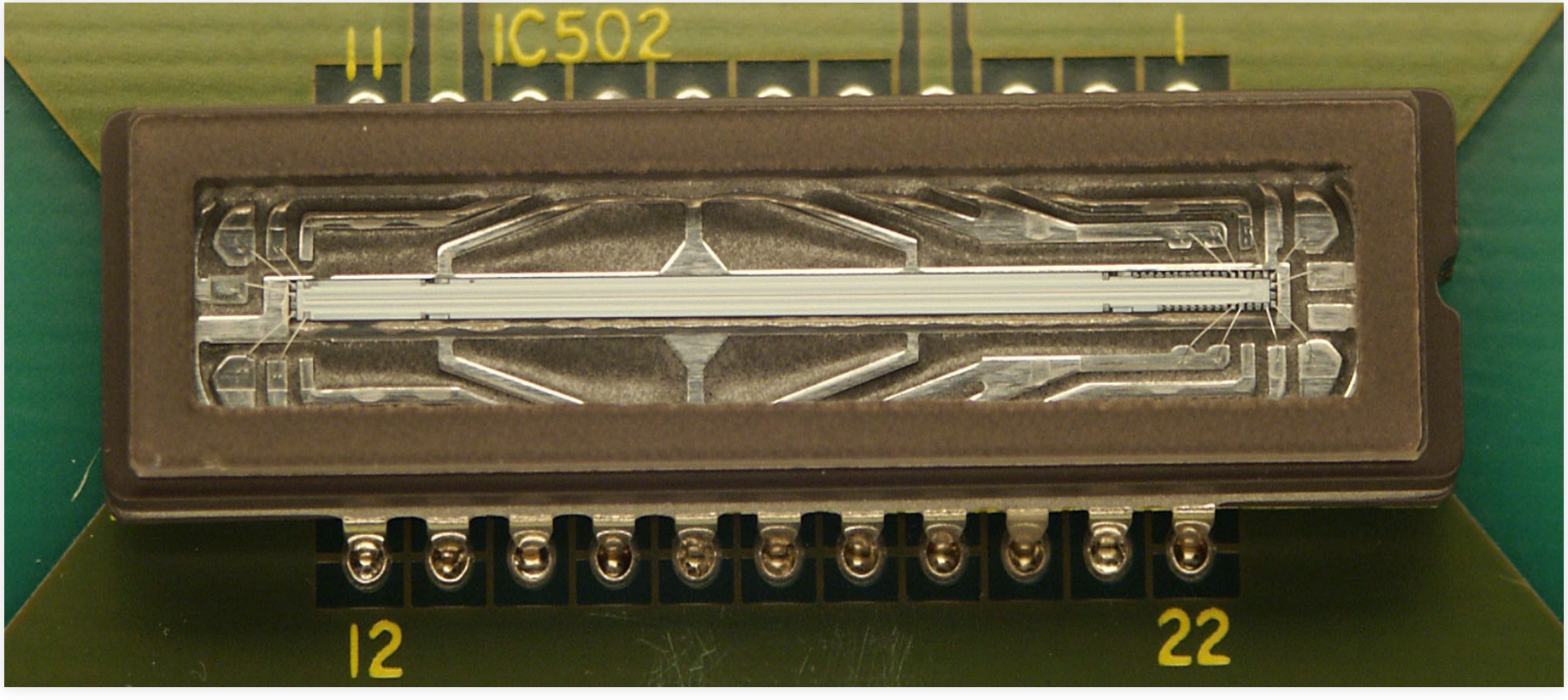
50% Grün, 25% Rot, 25% Blau; 1 Farbe pro Pixel; Interpolation.

Bildquelle: Nutzer Amada44, https://de.wikipedia.org/wiki/Bayer-Sensor#/media/File:Bayer_matrix.svg, Lizenz: gemeinfrei.

Scanner

- Lichtempfindliche Dioden (wie bei der Digitalkamera)
- Aber: in einer langen Reihe angeordnet
- Meist drei Reihen für die drei Grundfarben
- Flachbettscanner: Glasplatte und bewegliche Sensorenleiste
- Dokumentenscanner: Feste Sensorenleiste, Papier wird daran vorbeigezogen

Sensor eines Flachbettscanners



Buchwippe

Scanroboter

https://www.youtube.com/watch?v=cmhIJOqepVU

”Book Flipping Scanner”

https://www.youtube.com/watch?v=03ccxwNssmo

3. Digitalisierung von Text

Manuell oder automatisch?

- Manuelle Texterfassung
  - Transkription / "double keying"
  - Wird oft an externe Dienstleister vergeben
  - Genauigkeit: bis zu 99,997% Zeichenkorrektheit
- Automatische Texterkennung
  - Optical Character Recognition (OCR)
  - Endnutzer-Software oder forschungsnahe Software
  - Nutzt Mustererkennung, Machine Learning und/oder Deep Learning

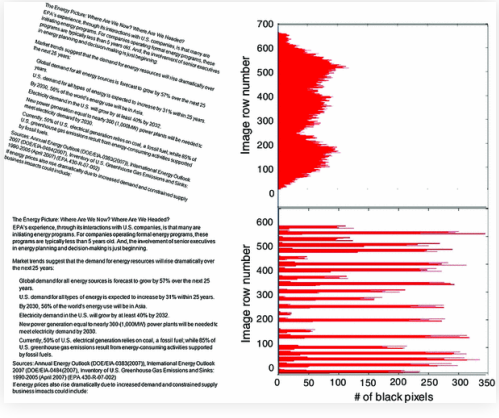
Die Schritte der OCR

1. Vorverabeitung der Vorlage
2. Segmentierung
3. Zeichenerkennung
4. Nachbereitung

1. Vorverabeitung der Vorlage

- Bildkorrekturen
  - Kontrast
  - Verzerrungen
  - Rotation
- Binarisierung (optional)
  - Transformation zu s/w-Bild

Bildrotation

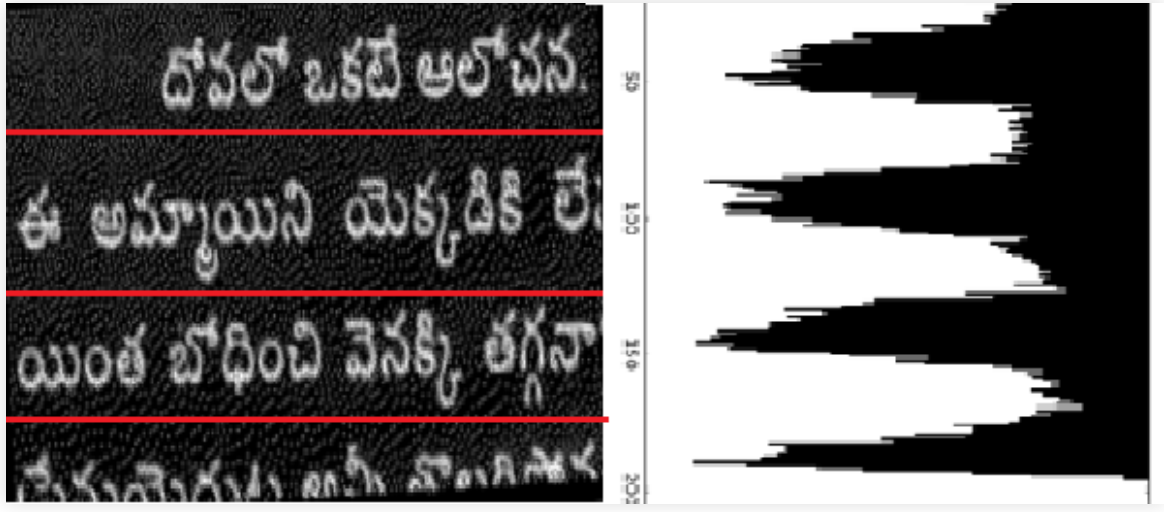


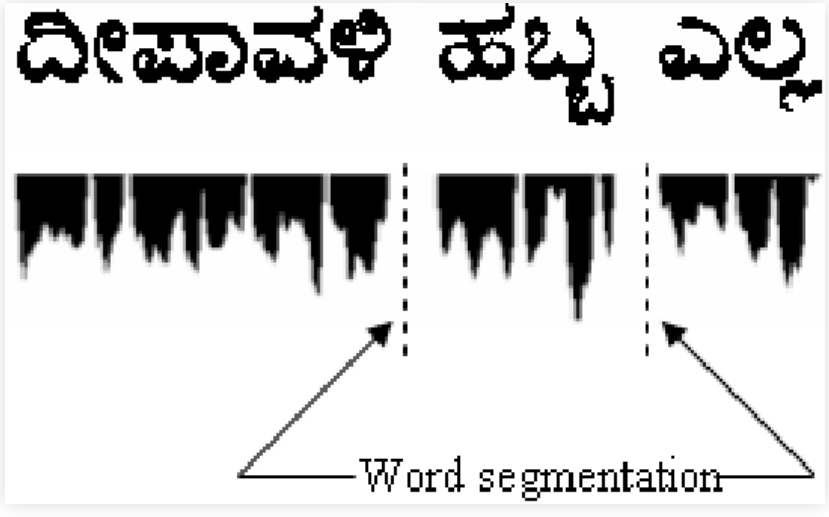
(Bildquelle: Towards Data Science)

2. Segmentierung

- Layout-Analyse
  - Textblöcke vs. Bildblöcke
  - Lauftitel, Überschrift, Absatz, Seitenzahl, etc.
- Segmentierung
  - Zeilenerkennung
  - ggfs. Wörter und Buchstaben (aber: Ligaturen)

Beispiel Zeilen-/Worttrennung





(Bildquelle: Towards Data Science)

3. Zeichenerkennung

- Musterabgleich oder Machine Learning
  - Musterabgleich: prototypische Buchstaben, Ähnlichkeit als Pixelüberlappung
  - Machine Learning / Deep Learning: feature-basiert (Kanten, Kurven, etc.)
- In beiden Fällen:
  - Anpassung an Schriftarten durch Training etc.
  - Ligaturen sind eine Herausforderung

Beispiel Ligaturen

(Bildquelle: fonts.com)

4. Nachbereitung

- Wörterbuchbasiert: Liste von Wörtern, die in einer Sprache vorkommen
- Sprachmodellbasiert: bspw. Wahrscheinlichkeit der Abfolge von Buchstaben in einer Sprache

Software

- Finereader (seit 1993; Abby; proprietär)
- OCRopus (seit 2007; Google / DFKI; Open Source)
- Tesseract (seit 1985; HP, Google; Open Source)
- OCR4all (seit 2018; Universität Würzburg; Open Source)

Faktoren für die OCR-Qualität

- Druckqualität (u.a. verblasste Buchstaben)
- Komplexität des Layouts (Marginalien, etc.)
- verwendete Schriftart (u.a. Fraktur / Antiqua)
- Anzahl der verwendete Sprache(n)
- moderne oder historische Sprachstufe (u.a. Zeicheninventar, Abkürzungen)
- manuelle Unterstreichungen, Annotationen, Marginalien
- Bildqualität des Scans (u.a. Auflösung, Ausrichtung)

OCR-Qualität

Zeichen-Genauigkeit	Verfahren	Zweck
99,997%	double keying mit Korrekturen	Editionen
99,95%	double keying	Negativsuche
99%	OCR bei sehr guter Vorlage	Text Mining
95%	OCR bei noch guter Vorlage	Positivsuche

Andere Bereiche der Digitalisierung

- Multispektralaufnahmen
- Videodigitalisierung
- Laserscanning (Geländemodelle)
- Audiodigitalisierung (Musik, Sprache)
- 3D-Digitalisierung (Skulpturen, Artefakte)
- Handschriftenerkennung (HTR)
- Optical Music Recognition :-)
- uvm.

4. Nach der Digitalisierung: Langzeitverfügbarkeit

Langzeitarchivierung (LZA)

Langzeit bedeutet für die Bestandserhaltung digitaler Ressourcen nicht die Abgabe einer Garantieerklärung über fünf oder fünfzig Jahre, sondern die verantwortliche Entwicklung von Strategien, die den beständigen, vom Informationsmarkt verursachten Wandel bewältigen können. [...]
Archivieren bedeutet [...] mehr als nur die dauerhafte Speicherung digitaler Informationen auf einem Datenträger. Vielmehr schließt es die Erhaltung der dauerhaften Verfügbarkeit und damit eine Nachnutzung und Interpretierbarkeit der digitalen Ressourcen mit ein.
(Heike Neuroth in nestor-Handbuch 2012)

Zwei Aspekte von LZA

- Archivierung: Erhaltung der Datensubstanz
  - sog. "bit-stream-preservation" (rein technische Ebene)
- Nachnutzbarkeit: Erhaltung der Benutzbarkeit
  - Verwendung von Standards; Dokumentation
  - Migration auf aktuelle Datenformate (offene Standards!)
  - Software-seitige Emulation von Systemumgebungen
  - Interoperabilität: technisch und intellektuell

Praktisches zur Langzeitverfügbarkeit

- Bild-Digitalisate:
  - unkomprimierten "Master" archivieren ("Baseline TIFF")
  - JPEG oder PNG für die Präsentation im Netz
- Volltext-Digitalisate:
  - in Standard-Format archivieren (XML-TEI)
  - ggfs. auch die Bild-Text-Link-Information, ALTO-Format (Analyzed Layout and Text Object)
- Alle Digitalisattypen: Metadaten
  - Administrativ, technisch, deskriptiv
  - in Standardformaten festhalten
  - Jedes Objekt mit Identifier versehen

Abschluss

Fragen?

Lektürehinweise

- Malte Rehbein, "Digitalisierung", in: *Digital Humanities: Eine Einführung*, hg. von Fotis Jannidis, Hubertus Kohle und Malte Rehbein. Stuttgart: Metzler, 2017, S. 179-196.

Weitere Empfehlungen

- "DFG-Praxisregeln Digitalisierung", Bonn: Deutsche Forschungsgemeinschaft, 2016. http://www.dfg.de/formulare/12_151/
- Melissa Terras, "Digitization and digital resources in the humanities", *Digital Humanities in Practice*, hg. von Claire Warwick, Melissa Terras, Julianne Nyhan. London: Facet, 2012, 47-70.

Darüber hinaus

- *nestor-Handbuch: Eine kleine Enzyklopädie der digitalen Langzeitarchivierung*. Version 2.3, hg. von Heike Neuroth, Achim Oßwald, Regine Scheffel, Stefan Strathmann, Karsten Huth. Göttingen: nestor, 2010. http://www.nestor.sub.uni-goettingen.de/handbuch/index.php (Einleitung sowie Kapitel 4.2 zu OAIS)

Nächste Sitzung

- Thema: Einstieg ins Programmieren
- "Software development is a learning experience with written code as side effect." (m2spring auf reddit.com)
- Vorbereitung: Fotis Jannidis, "Grundbegriffe des Programmierens" (Kapitel 6 im Lehrbuch)

Christof Schöch, 2019
http://www.christof-schoech.de
