

FreieSoftwareOG –

Vom Bild zum bearbeitbaren Text - OCR-Tools



OCR – Begriffserklärung

Texterkennung ist ein Begriff aus der Informationstechnik.

Es bezeichnet die automatisierte Text- bzw. Schrifterkennung innerhalb von Bildern.

Ursprünglich basierte die automatische Texterkennung auf optischer Zeichenerkennung (optical character recognition, OCR).

Diese Technik wird zunehmend durch neuronale Netze, die ganze Zeilen statt einzelner Zeichen verarbeiten, abgelöst.

OCR – Begriffserklärung

Texterkennung ist deshalb notwendig, weil optische Eingabegeräte (Scanner, Digitalkameras, Faxempfänger) als Ergebnis ausschließlich Rastergrafiken liefern können, d. h. in Zeilen und Spalten angeordnete Punkte unterschiedlicher Färbung (Pixel).

Texterkennung bezeichnet dabei die Aufgabe, die so dargestellten Buchstaben als solche zu erkennen, d. h. zu identifizieren und ihnen den Zahlenwert zuzuordnen, der ihnen nach üblicher Textcodierung zukommt (ASCII, Unicode).

OCR – Begriffserklärung

Automatische Texterkennung und OCR werden im deutschen Sprachraum oft als Synonym verwendet.

In technischer Hinsicht bezieht sich OCR jedoch nur auf den Teilbereich der Mustervergleiche von separierten Bildteilen als Kandidaten zur Erkennung von Einzelzeichen.

OCR – Begriffserklärung

Diesem OCR-Prozess geht eine globale Strukturerkennung voraus, in der zuerst Textblöcke von graphischen Elementen unterschieden, die Zeilenstrukturen erkannt und schließlich Einzelzeichen separiert werden.

Bei der Entscheidung, welches Zeichen vorliegt, kann über weitere Algorithmen ein sprachlicher Kontext berücksichtigt werden.

OCR – Begriffserklärung

Ursprünglich wurden zur automatischen Texterkennung eigens entworfene Schriftarten entwickelt, die zum Beispiel für das Bedrucken von Scheckformularen verwendet wurden.

Diese Schriftarten waren so gestaltet, dass die einzelnen Zeichen von einem OCR-Lesegerät schnell und ohne großen Rechenaufwand unterschieden werden konnten.

OCR – Begriffserklärung

So zeichnet sich die Schriftart OCR-A (DIN 66008, ISO 1073-1) durch einander besonders unähnliche Zeichen, besonders bei den Ziffern, aus.

Das ist OCR-A von 1968

OCR – Begriffserklärung

OCR-B (ISO 1073-2) ähnelt mehr einer serifenlosen, nicht-proportionalen Schriftart, während OCR-H (DIN 66225) handgeschriebenen Ziffern und Großbuchstaben nachempfunden wurde.

Das ist OCR-B von 1968

OCR – Begriffserklärung

Die gestiegene Leistungsfähigkeit moderner Computer und verbesserte Algorithmen erlauben inzwischen auch die Erkennung von „normalen“ Druckerschriftarten bis hin zu Handschriften (etwa bei der Briefverteilung);

wenn jedoch Lesbarkeit durch Menschen nicht vorrangig ist, werden drucktechnisch und erkenntungstechnisch einfacher handhabbare Strichcodes genutzt.

OCR – Tesseract

tesseract-ocr ist ein Kommandozeilenprogramm zur Texterkennung.

Ursprünglich von Hewlett-Packard zwischen 1984 und 1995 als kommerzielles Programm entwickelt, wurde der Code 2005 freigegeben.

OCR – Tesseract

Die Entwicklung wird von Google unterstützt, da eine Open-Source-Lösung zur Erstellung von E-Books benötigt wurde. Das Programm unterstützt etliche westeuropäische und asiatische Sprachen wie z.B. vietnamesisch.

OCR – Tesseract

Zunächst war tesseract-ocr ein reines Zeichenerkennungs-Programm, es lieferte keine Layout-Analyse, und gab nur unformatierten Text aus.

Ab Version 3.00 ist auch hOCR*, ab 3.03 zudem als PDF, ab 4.0.0 auch weitergehende Layoutanalysen (ALTO) möglich.

Die Texterkennung kann theoretisch auch „trainiert“ werden.

OCR – Tesseract

Es gibt einige Projekte, die auf tesseract-ocr als OCR-Engine zurückgreifen.

Auch aus XSane heraus kann das Programm verwendet werden.

OCR – Tesseract

Zunächst die Engine installieren

sudo apt-get install tesseract-ocr

...dann

apt-cache search tesseract-ocr

...und dann

apt-get install tesseract-ocr-deu

```
15:11
> apt-cache search tesseract-ocr
gimagereader - Grafische GTK+-Oberfläche für tesseract-ocr
gimagereader-common - gImageReader common files
gimagereader-qt5 - Graphical Qt 5 front-end to tesseract-ocr
python3-pyocr - Python wrapper for OCR engines (Python 3)
python3-tesseract - Python wrapper for the tesseract-ocr API (Python3 version)
tesseract-ocr - Kommandozeilenwerkzeug für (optische) Zeichenerkennung (OCR)
tesseract-ocr-all - Optische Zeichenerkennung Tesseract mit allen Sprach- und Skriptpaketen
tesseract-ocr-afr - tesseract-ocr language files for Afrikaans
tesseract-ocr-amh - tesseract-ocr language files for Amharic
tesseract-ocr-ara - tesseract-ocr language files for Arabic
tesseract-ocr-asm - tesseract-ocr language files for Assamese
tesseract-ocr-aze - tesseract-ocr language files for Azerbaijani
tesseract-ocr-aze-cyrl - tesseract-ocr language files for Azerbaijani (Cyrillic)
tesseract-ocr-bel - tesseract-ocr language files for Belarusian
tesseract-ocr-ben - tesseract-ocr language files for Bengali
tesseract-ocr-bod - tesseract-ocr language files for Tibetan Standard
tesseract-ocr-bos - tesseract-ocr language files for Bosnian
tesseract-ocr-bre - tesseract-ocr language files for Breton
tesseract-ocr-bul - tesseract-ocr language files for Bulgarian
tesseract-ocr-cat - tesseract-ocr language files for Catalan
tesseract-ocr-ceb - tesseract-ocr language files for Cebuano
tesseract-ocr-ces - tesseract-ocr language files for Czech
tesseract-ocr-chi-sim - tesseract-ocr language files for Chinese - Simplified
tesseract-ocr-chi-sim-vert - tesseract-ocr language files for Chinese - Simplified (vertical)
tesseract-ocr-chi-tra - tesseract-ocr language files for Chinese - Traditional
tesseract-ocr-chi-tra-vert - tesseract-ocr language files for Chinese - Traditional (vertical)
tesseract-ocr-chr - tesseract-ocr language files for Cherokee
tesseract-ocr-cos - tesseract-ocr language files for Corsican
tesseract-ocr-cym - tesseract-ocr language files for Welsh
tesseract-ocr-dan - tesseract-ocr language files for Danish
tesseract-ocr-deu - tesseract-ocr language files for German
tesseract-ocr-div - tesseract-ocr language files for Divehi
tesseract-ocr-dzo - tesseract-ocr language files for Dzongkha
tesseract-ocr-ell - tesseract-ocr language files for Greek
```

OCR – OCRmyPDF

- Befehlszeilen-Werkzeug auf Basis von Tesseract
- Erstellt eine durchsuchbare Textebene des PDFs
- Verwendet den PDF/A-2b-Standard für das Ergebnis
- Verfügbar für GNU/Linux, Mac und Windows (WSL)

OCR – glimageReader

- Ein einfaches Gtk/Qt-Frontend für Tesseract
- Verarbeitet neben Bildern auch PDFs und direkte SANE-Scans
- Verfügbar für GNU/Linux und Windows

OCR – gImageReader

scan2.png - gImageReader

Datei Ansicht Hilfe

Öffnen Erwerben

Deutsch (de_DE) Alles erkennen

Bilder Erwerben

Gerät: BearPaw 1200 CU

Ausgabebild: /home/heinrich/scan3.png

Auflösung: 300 dpi

0 0 100 %

Einfügen An Cursorstelle

Aufgabe 8

a) Lies die nachfolgenden Informationen und fertige eine Zeitleiste an.
In welchen Abschnitten verlief die Einwanderung?

b) Welche Gründe mögen für die zunehmende Ausbreitung der Asiatischen Tigermücke in den Norden verantwortlich sein? Kennzeichne die Fleisewege und die Transportmittel in der Weltkarte.

c) Informiere dich: Welche Kriterien müssen erfüllt sein, damit man von Neozoen, also etablierten Arten spricht?

Warum spielt wohl das Jahr 1492 eine wichtige Rolle? 1967 wurden große Teile Asiens und die Inseln im Indischen und im Pazifischen Ozean als typisches Verbreitungsgebiet für die Asiatische Tigermücke angegeben. In Europa fand man das Tier zuerst 1979 in Albanien. Vermutlich kam sie dorthin mit Warenlieferungen aus China. 1990/91 wurde sie, so nehmen Experten an, mit gebrauchten Autoreifen aus dem Süden der USA (Georgia) nach Italien gebracht. Seitdem hat sie sich in ganz Italien und auch auf Sizilien und Sardinien verbreitet. Seit 1999 findet man sie in Frankreich, vor allem im Süden. Dann erfolgte eine rasche Verbreitung in großen Teilen Europas: In Belgien wurde sie im Jahre 2000 zum ersten Mal nachgewiesen, 2001 in Montenegro, 2003 in Kroatien, 2004 in Spanien und in Kroatien, 2005 in den Niederlanden, in Slowenien und auf Korsika, 2006 in Bosnien und Herzegowina. Im Herbst 2007 wurden sie in Deutschland entdeckt.

Auch weltweit bemerkt man einen Vormarsch der Asiatischen Tigermücke. Ende 1992 war die Art in mehr als 20 Staaten der USA heimisch, 1998 fand man sie in Brasilien, dann in Mexiko, 1993 in der Dominikanischen Republik, 1995 in Bolivien, Kuba, Honduras und Guatemala, 1996 in El Salvador, 1997 auf den Kaimaninseln, 1998 in Argentinien, 1999 in Paraguay, 2001 in Kolumbien, 2002 in Peru und 2003 in Uruguay und Nicaragua. In Afrika soll sie sich seit mindestens 1991 heimisch, 1999/2000 breitete sie sich nach Kamerun und Äquatorialguinea und 2006 nach Gabun aus. In Naher Osten fand man die Asiatische Tigermücke 2003 im Libanon und in Israel und 2005 in Syrien. In Australien und Neuseeland ist die Art nicht heimisch. Sie wurde zwar mehrfach eingeschleppt, konnte sich dort aber nicht etablieren. Ein Grund dafür ist wohl die konstante Überwachtung der Küsten und Flugplätze in diesen Ländern.

Thema 3 – Reisewege und -mittel

Wieviele gibt es mehr als 2000 Sechsmalstagen. Sie leben meist in der Nähe von Wasserstellen und, abgesehen von den Einflüssen der Polargebiete und in Wäldern, kommen sie nahezu überall vor. Die verschiedenen Arten haben aber jeweils spezifische Verbreitungsgrenzen. Obwohl die Asiatische Tigermücke ursprünglich in den Tropen und Subtropen lebt, hat sie sich erfolgreich in kühleren Regionen angepasst. Hier überwinterst sie dann meist im Ei-Stadium und ist nicht, wie in ihrer Heimat, gänzlich aktiv. Mitteleuropäer haben sich die Tiere so gut angepasst, dass die Eier sogar Minustemperaturen und Schnee überleben.

Aufgabe 7

Schau dir die Weltkarte an und fertige eine Liste an: In welchen Ländern kommt die Asiatische Tigermücke vor? (Markierung?) In welchen Ländern ist sie in den letzten 30 Jahren eingewandert? (Markierung?) Nimm auf einen Atlas zu Hilfe, um die Ländernamen zu finden.

Abb. 2: Verbreitungsbereiche der Asiatischen Tigermücke

Aufgabe 6

a) Lies die nachfolgenden Informationen und fertige eine Zeitleiste an.
In welchen Abschnitten verlief die Einwanderung?

b) Welche Gründe mögen für die zunehmende Ausbreitung der Asiatischen Tigermücke in den Norden verantwortlich sein? Kennzeichne die Reisewege und die Transportmittel in der Weltkarte.

c) Informiere dich: Welche Kriterien müssen erfüllt sein, damit man von Neozoen, also etablierten Arten spricht? Warum spielt wohl das Jahr 1492 eine wichtige Rolle?

1967 wurden große Teile Asiens und die Inseln im Indischen und im Pazifischen Ozean als typisches Verbreitungsgebiet für die Asiatische Tigermücke angegeben. In Europa fand man das Tier zuerst 1979 in Albanien. Vermutlich kam sie dorthin mit Warenlieferungen aus China. 1990/91 wurde sie, so nehmen Experten an, mit gebrauchten Autoreifen aus dem Süden der USA (Georgia) nach Italien gebracht. Seitdem hat sie sich in ganz Italien und auch auf Sizilien und Sardinien verbreitet. Seit 1999 findet man sie in Frankreich, vor allem im Süden. Dann erfolgte eine rasche Verbreitung in großen Teilen Europas: In Belgien wurde sie im Jahre 2000 zum ersten Mal nachgewiesen, 2001 in Montenegro, 2003 in Kroatien, 2004 in Spanien und in Kroatien, 2005 in den Niederlanden, in Slowenien und auf Korsika, 2006 in Bosnien und Herzegowina. Im Herbst 2007 wurden sie in Deutschland entdeckt.

Auch weltweit bemerkt man einen Vormarsch der Asiatischen Tigermücke. Ende 1992 war die Art in mehr als 20 Staaten der USA heimisch, 1998 fand man sie in Brasilien, dann in Mexiko, 1993 in der Dominikanischen Republik, 1995 in Bolivien, Kuba, Honduras und Guatemala, 1996 in El Salvador, 1997 auf den Kaimaninseln, 1998 in Argentinien, 1999 in Paraguay, 2001 in Kolumbien, 2002 in Peru und 2003 in Uruguay und Nicaragua. In Afrika soll sie sich seit mindestens 1991 heimisch, 1999/2000 breitete sie sich nach Kamerun und Äquatorialguinea und 2006 nach Gabun aus. In Naher Osten fand man die Asiatische Tigermücke 2003 im Libanon und in Israel und 2005 in Syrien. In Australien und Neuseeland ist die Art nicht heimisch. Sie wurde zwar mehrfach eingeschleppt, konnte sich dort aber nicht etablieren. Ein Grund dafür ist wohl die konstante Überwachtung der Küsten und Flugplätze in diesen Ländern.

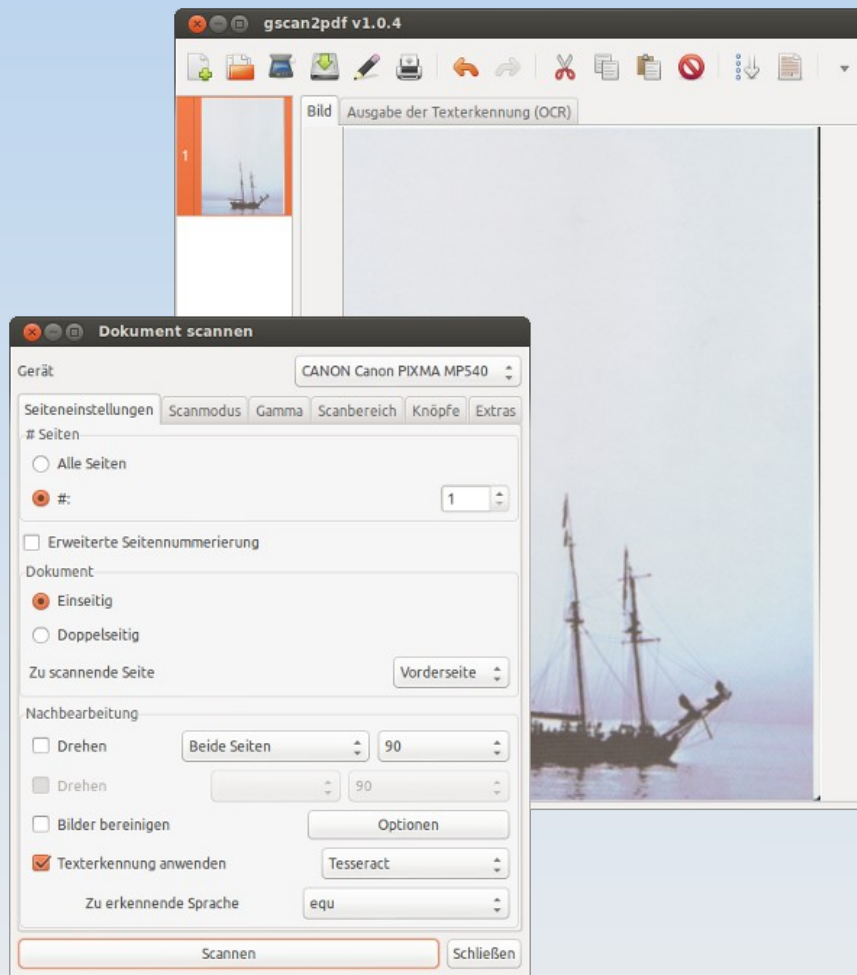
Um eine spezifische Region zu erkennen kann ein Rechteck darüber gezogen werden.

FreieSoftware0G.org

OCR – gscan2pdf

- Ein grafisches Frontend zur Erstellung von PDFs und DjVus*
(OCR via Tesseract)
- Verfügbar für GNU/Linux, Mac und Windows (WSL)

OCR – gscan2pdf



OCR – NormCap

- Screenshot-Werkzeug mit Fokus auf OCR via Tesseract
- Verfügbar für GNU/Linux, Mac und Windows
- Hands-On nach der Präsentation

OCR – NormCap

Einführung in NormCap

Grundlegende Bedienung

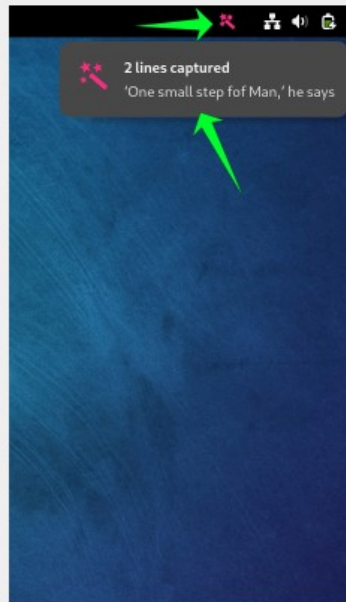
1. Bereich auswählen

Warte, bis ein rosa Rand um den Bildschirm erscheint, und wähle dann den gewünschten Bereich aus.



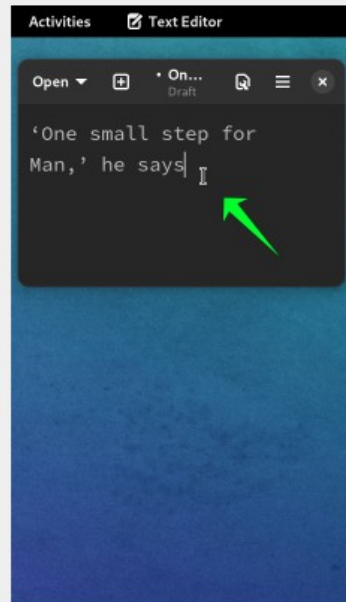
2. Erkennung abwarten

Die Verarbeitung benötigt Zeit. Warte auf eine Benachrichtigung oder eine Farbänderung des Symbols im Infobereich.



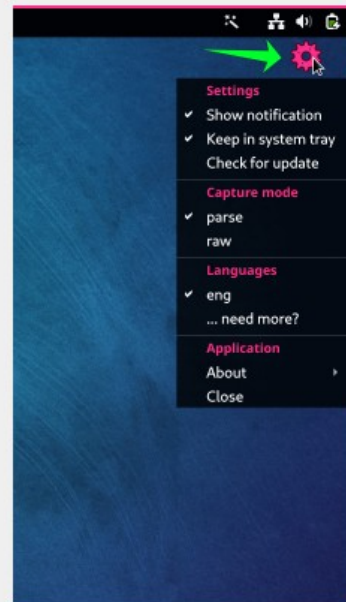
3. Aus Zwischenablage einfügen

Der erkannte Text wird in die Zwischenablage Deines Systems kopiert. Füge es in eine beliebige Anwendung ein (Strg + v).



Einstellungen und mehr

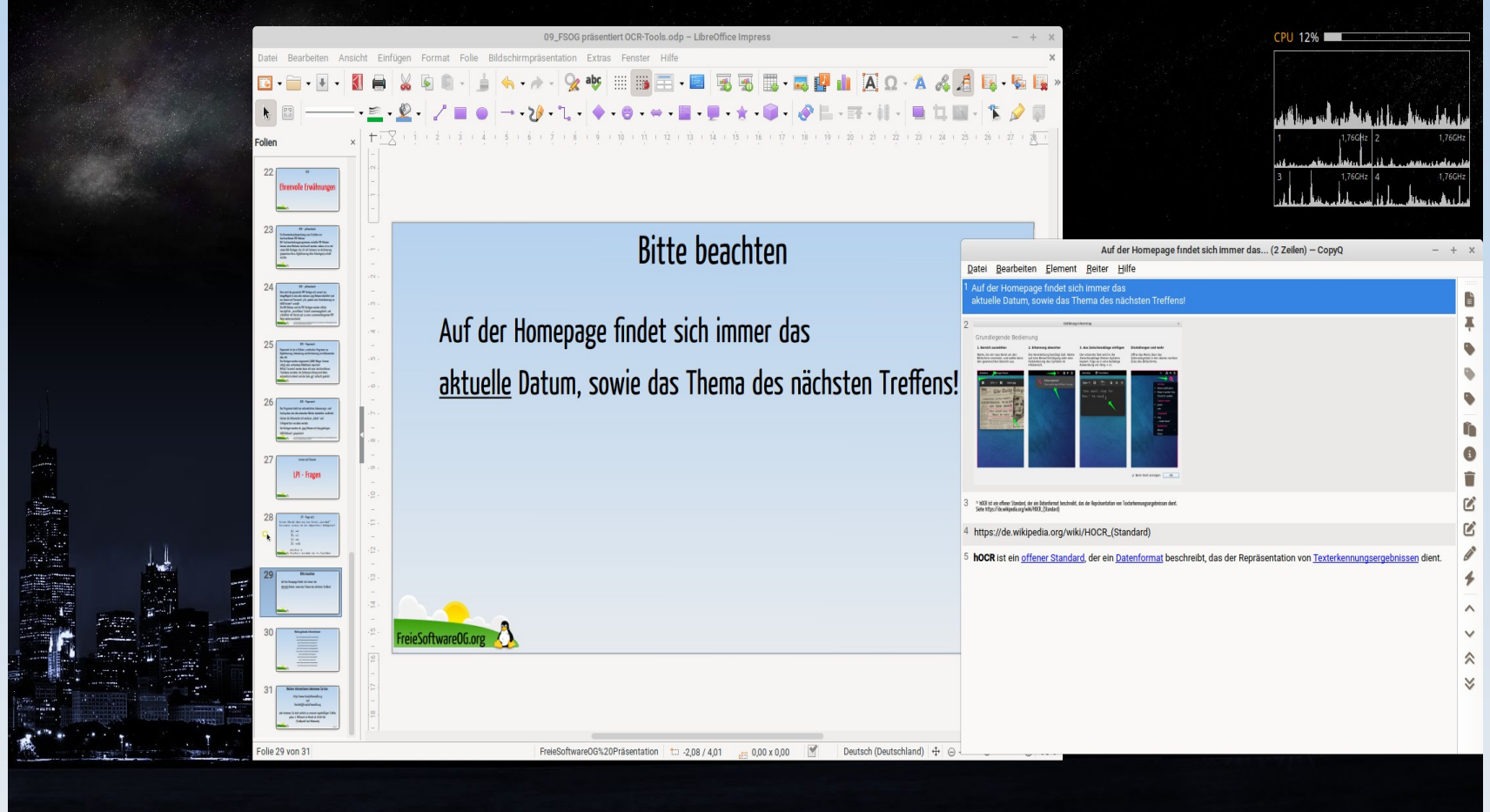
Öffne das Menü über das Zahnradsymbol in der oberen rechten Ecke des Bildschirms.



☒ Beim Start anzeigen

Ok

OCR – NormCap



The screenshot shows a presentation window titled "09_FSDG präsentiert OCR-Tools.odp - LibreOffice Impress". The presentation slide is titled "Bitte beachten" and contains the text: "Auf der Homepage findet sich immer das aktuelle Datum, sowie das Thema des nächsten Treffens!". The slide also features a logo for "FreieSoftwareOG.org" at the bottom. To the right of the presentation window, there is a terminal window titled "Auf der Homepage findet sich immer das... (2 Zeilen) - CopyQ". The terminal window displays the following text: "1 Auf der Homepage findet sich immer das aktuelle Datum, sowie das Thema des nächsten Treffens!". Below this, there are three small images showing a mobile app interface. The terminal window also shows a list of links and a description of the "hOCR" format. In the top right corner, there is a CPU usage monitor showing "CPU 12%" and a waveform graph. The background of the presentation window shows a night cityscape with the Willis Tower.

09_FSDG präsentiert OCR-Tools.odp - LibreOffice Impress

Bitte beachten

Auf der Homepage findet sich immer das aktuelle Datum, sowie das Thema des nächsten Treffens!

FreieSoftwareOG.org

Auf der Homepage findet sich immer das... (2 Zeilen) - CopyQ

1 Auf der Homepage findet sich immer das aktuelle Datum, sowie das Thema des nächsten Treffens!

2

Grundlegende Bedeutung

3 * hOCR ist ein offener Standard, der als Repräsentation von Texterkennungsergebnissen dient. Siehe [https://de.wikipedia.org/wiki/hOCR_\(Standard\)](https://de.wikipedia.org/wiki/hOCR_(Standard))

4 [https://de.wikipedia.org/wiki/hOCR_\(Standard\)](https://de.wikipedia.org/wiki/hOCR_(Standard))

5 hOCR ist ein offener Standard, der ein Datenformat beschreibt, das die Repräsentation von Texterkennungsergebnissen dient.

CPU 12%

OCR

Ehrenvolle Erwähnungen



OCR – pdfsandwich

Ein Kommandozeilenwerkzeug zum Erstellen von durchsuchbaren PDF-Dateien.

Mit Textverarbeitungsprogrammen erstellte PDF-Dateien können ohne Weiteres durchsucht werden, anders ist es mit reinen Bild-Vorlagen, die z.B. mit Scannern zur Archivierung (papierloses Büro, Digitalisierung alter Unterlagen) erstellt wurden.

OCR – pdfsandwich

Dazu wird die gescannte PDF-Vorlage mit convert aus ImageMagick in eine oder mehrere .png-Dateien überführt und aus diesen mit Tesseract 3.0x, jeweils eine Texterkennung im hOCR-Format* erstellt.

Die OCR-Dateien und die PDF-Vorlagen werden mittels hocr2pdf als „unsichtbare“ Schicht zusammengeführt, und schließlich mit Ghostscript zu einer zusammenhängenden PDF-Datei weiterverarbeitet.

OCR – Paperwork

Paperwork ist ein in Python 3 verfasstes Programm zur Digitalisierung, Indexierung und Archivierung von Dokumenten aller Art.

Die Vorlagen werden eingescannt (SANE-fähiger Scanner nötig) oder vorhandene Bilddateien importiert.

Mittels Tesseract werden diese mit einer durchsuchbaren Textebene versehen, die Seitenausrichtung wird dabei automatisch erkannt und die Seite ggf. aufrecht gedreht.

OCR – Paperwork

Das Programm bietet ein automatisches Indexierungs- und Suchsystem, das alle erkannten Wörter beinhaltet, zusätzlich können die Dokumente mit weiteren „Labels“ und Schlagwörtern versehen werden.

Die Vorlagen werden als .jpeg-Dateien mit dazugehörigen hOCR-Dateien* gespeichert.

Lernen und Staunen

LPI - Fragen



LPI - Frage #22

Welche Option kann mit dem Befehl „useradd“ verwendet werden, um die Login-Shell anzugeben?

- A) -s
- B) -l
- C) -u
- D) -sh

Antwort: A

Beispiel: `useradd edi -s /bin/bash`

Bitte beachten

Auf der Homepage findet sich immer das
aktuelle Datum, sowie das Thema des nächsten Treffens!

Weitergehende Informationen

<https://de.wikipedia.org/wiki/Texterkennung>

<https://wiki.ubuntuusers.de/tesseract-ocr/>

<https://wiki.ubuntuusers.de/OCRmyPDF/>

<https://github.com/ocrmypdf/OCRmyPDF>

<https://wiki.ubuntuusers.de/gImageReader/>

<https://github.com/manisandro/gImageReader>

<https://gscan2pdf.sourceforge.net/>

<https://wiki.ubuntuusers.de/gscan2pdf/>

<https://dynobo.github.io/normcap/>

<https://wiki.ubuntuusers.de/pdfsandwich/>

<https://wiki.ubuntuusers.de/Paperwork/>

Weitere Informationen bekommen Sie hier:

<http://www.FreieSoftware0G.org>

und

Kontakt@FreieSoftware0G.org

oder kommen Sie doch einfach zu unserem regelmäßigen Treffen,
jeden 1. Mittwoch im Monat ab 20:00 Uhr.
(Treffpunkt laut Webseite)

