



OCR for Most of the World's Languages

Sep 3, 2015

Ashok Popat

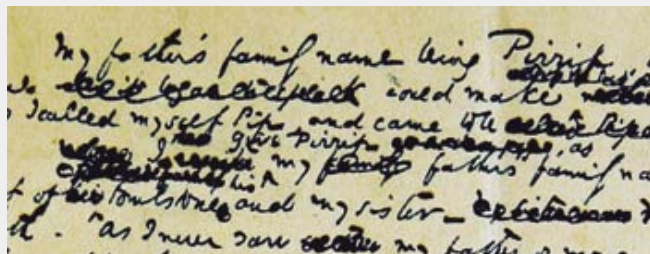
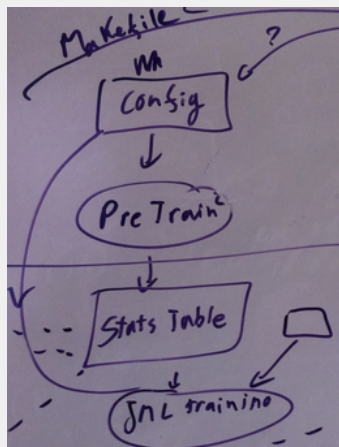
Research Scientist, Google Inc.

Outline

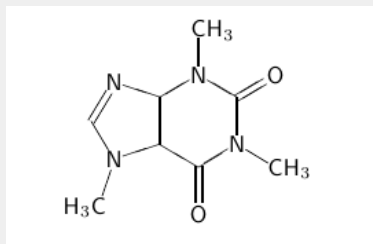
1. Optical Character Recognition
2. Approach
3. Reflections and comments

Optical Character Recognition

THE POSITION OF THE CITY IN THE UNITED STATES



$$s[n] \stackrel{\text{def}}{=} T \int_{-1/T}^{1/T} S_{1/T}(f) \cdot e^{i2\pi f n T} df = T \underbrace{\int_{-\infty}^{\infty} S(f) \cdot e^{i2\pi f n T} df}_{\stackrel{\text{def}}{=} s(nT)}$$



Optical Character Recognition

THE POSITION OF THE CITY IN THE UNITED STATES

University of Wisconsin

Essay on the Immediate Data of Consciousness,

應收款項與應付款項比率

وهنت «^(۳) (Grenfell and Hunt) — هي المقصودة بالاشارة التي وردت

มีตัวเลขนักท่องเที่ยวประมาณ 220,000 คน

Examples from Google Books

Multiple scripts / languages on a page:

(601, 602.) CIVITAS. LIB. X. 181

χρωμάτων δὲ καὶ σχημάτων θεωροῦσι; Πάνυ μὲν οὖν. Οὕτω δὲ, οἶμαι, καὶ τὸν ποιητικὸν φήσομεν χροῶματα ἅττα ἐκάστων τῶν τεχνῶν τοῖς ὀνόμασι καὶ ῥήμασιν ἐπιχρωματίζειν, αὐτὸν οὐκ ἐπαίοντα ἀλλ' ἢ μιμῆσθαι, ὥστε ἐτέροις τοιοῦτοις ἐκ τῶν λόγων θεωροῦσι δοκεῖν, εἴαν τε περὶ σκυτοτομίας τις λέγῃ ἐν μέτρῳ καὶ ῥυθμῷ καὶ ἁρμονίᾳ, πάνυ εὖ δοκεῖν λέγεσθαι, εἴαν τε περὶ στρατηγίας εἴαν τε περὶ ἄλλου ὁπουοῦν· οὕτω φύσει αὐτὰ ταῦτα μεγάλῃ τινὰ κλησιν ἔχειν. ἐπεὶ γυμνοθέντα γε τῶν τῆς μουσικῆς χρωμάτων τὰ τῶν ποιητῶν, αὐτὰ ἐφ' αὐτῶν λεγόμενα, οἶμά σε εἰδέναι οἷα φαίνεται. τεθέασαι γάρ που. Ἐγὼ γ', ἔφη. Οὐκ οὔν, ἦν δ' ἐγὼ, ἔοικε τοῖς τῶν ὠραίων προσώποις, καλῶν δὲ μὴ, οἷα γίγνεται ἰδεῖν, ὅταν αὐτὰ τὸ ἄνθος προλίπη; Παντάπασιν, ἦ δ' ὅς. Ἴθι δὲ, τότε ἄθρει· ὁ τοῦ εἰδῶλου ποιητῆς, ὁ μιμητῆς, φάμεν, τοῦ μὲν ὄντος οὐδὲν ἐπαίει, τοῦ δὲ φαινομένου. οὐχ οὕτω; Ναί. Μὴ τοίνυν ἡμίσεως αὐτὸ καταλείπομεν ῥηθὲν, ἀλλ' ἱκανῶς ἴδωμεν. Λέγε, ἔφη. Ζωγράφος,

[socr.] Haud secus, opinor, poetam dicemus colores quosdam artium singularum nominibus verbisque exprimere, cum ipse nihil sciat nisi imitari, adeo ut aliis similibus secundum verba spectantibus bene dicere videatur, sive de sutrina dicat versibus et numero et harmonia, sive de re militari, seu de quibuslibet aliis: tantam natura in his ipsis inesse delectationem. nam musicæ coloribus nudata poemata per se sola spectata scire te arbitror qualia videantur. vidisti enim, nisi fallor. [cl.] Vidi, inquit. [socr.] Nonne perinde se habent, inquam, ac facies eorum, qui in flore ætatis constituti neque pulchri sunt, quales visu fiunt, quando flos eis decedit? [cl.] Plane, inquit. [socr.] Nunc, age, hoc considera: imitator, hoc est, simulacri auctor, veri cognitionem nullam habet, sed visi. nonne? [cl.] Ita est. [socr.] Ne igitur semiperfectum id relinquamus, sed plene perspiciamus. [cl.] Dicis, inquit. [socr.] Pictor, dicimus, habenas pinget et frenum? [cl.] Ita. [socr.] Faciet autem coriarius et faber ærarius? [cl.] Faciet. [socr.] Num igitur


Examples from Google Books (cont.)

Per-word script and language variation:

Dieser Familie nähern sich die Slawonischen Worte *лана*, *лопастъ* und vorzüglich *лопата* (*lapa*, *lopast* und *lopata*, Pfote, Flügel einer Mütze, Schaufel), mit welcher ein Blatt durch seine Fläche eine grosse Aehnlichkeit hat. Daher nennen auch wir die grossen Blätter einiger Gewächse *лапушникъ* (*lapuschnik*, Klette). Wie sehr sich aber auch die letztern Worte

Examples from Google Books (cont.)

§. 6. Vesezeichen:

1. virâma, „Rußezeichen“ , ein unten rechts an den Konsonanten angefügter Strich, nimmt dem Konsonanten den a-Laut, mit welchem er nach § 1 und § 4 an und für sich zu sprechen ist. Der Virâma findet sich namentlich am Ende eines Wortes und manchmal anstatt einer Ligatur; z. B.

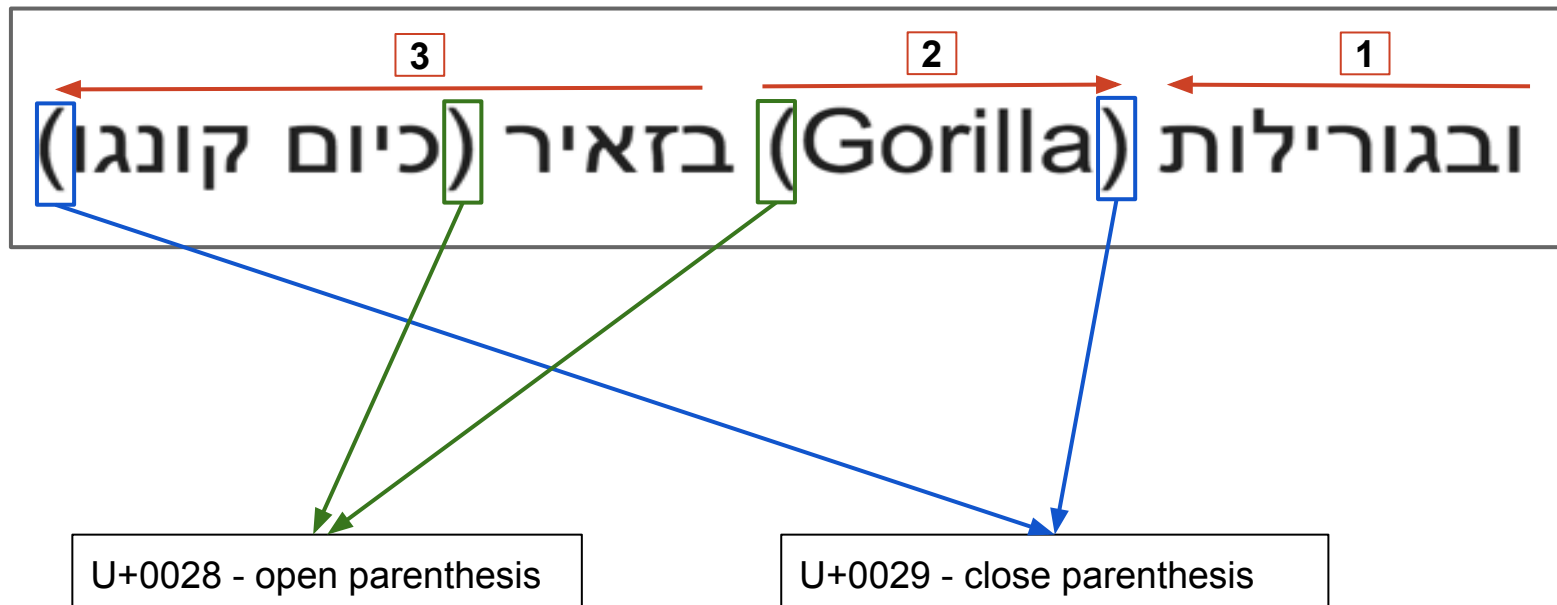
यत् yat, आसीत् âsit, आगमद् देवी âgamad dêvi;
क्त्र ktra kann auch geschrieben werden क्त्र.

2. avagraha, „Apostroph“ ऽ, steht am Anfange eines Wortes zur Bezeichnung eines vorn abgefallenen a: z. B. दुःखितो
ऽभवत् duḥkhitô 'bhavat.

What's a "character"?

Result	Unicode	Transliteration
र	0930	ra
र,	0930 094d	r
रु	0930 094d 0926	rda
रु,	0930 094d 0926 094d	rd
रुव	0930 094d 0926 094d 0935	rdva
रुवि	0930 094d 0926 094d 0935 093f	rdvi
रुविक	0930 094d 0926 094d 0935 093f 0915	rdvika

Bidirectional issues



Connected scripts

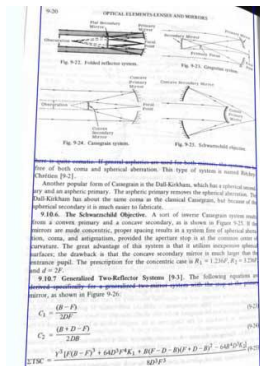
Naskh style: صفحات

Nastaliq style:

صفحات

Part 2: Approach

Optical character recognition as text-line decoding



Input Document

element. A Mangin mirror consists of a thick negative meniscus lens.

Preprocessing /
Layout Analysis

Text Line
Recognition

Digital (Unicode) Text

Style, Size, Position,
Font, Weight

Script, Language

Goal: universal, accurate OCR

- Universal
 - Omni-script
 - Omni-language
 - Omni-setting
- Accuracy and Speed
 - Best-in-world, approaching human accuracy
 - Speed comparable to commercial engines

Inspiration: Markov-model-based approaches

- Document image decoding [Kopec and Chou, 1994]
 - Explicit model of typesetting process: seek to invert
 - Influenced by speech recognition methods
 - Extremely high accuracy when models match the data

- BBN Byblos system [Schwartz et al., 1996]
 - Treat text line like a speech waveform
 - Built on existing speech recognition system
 - First successful Arabic OCR

Underlying both: noisy channel model

- Communication theory perspective
 - Source produces a message m according to $P(m)$
 - Channel (noisily) renders observed image x according to $P(x|m)$
 - OCR task: given x , produce an estimate of m
 - Goal: choose m' to minimize error rate:
$$m' = \arg \max_m P(m|x)$$
- Challenges
 - Nobody tells us what $P(m|x)$ is (modeling task)
 - Even if we knew $P(m|x)$, how to compute $\arg \max_m$?

Component models

- Language models
 - Character- and Word N-grams with appropriate smoothing (ProdLM)

- Likelihood component
 - **Speech, BBN OCR:** GMMs, DNNs for HMM state-conditional densities, optimized for held-out likelihood
 - **DID:** Learned probabilistic character templates (foreground, background, “don’t-care”)
 - **Ours:** Sliding window / deep network / HMMs

Generalization of the noisy channel model

- Speech approach

$$\begin{aligned} m' &= \arg \max_m P(m)^\alpha P(x|m) \\ &= \arg \max_m \alpha \log P(m) + \log P(x|m) \end{aligned}$$

- Generalize to multiple feature functions

$$m' = \arg \max_m \sum_i \lambda_i h_i(x, m)$$

- Learn $\{\lambda\}$ via minimum error-rate training

Principles

- Minimize language-specific engineering
- Prefer integrated, wholistic decisions to pipelined steps
- Take advantage of data (labeled, unlabeled)
- Take advantage of advances in other areas (MT, Speech, NLP, CV,...)

Accuracy and Speed over time

- More and more accurate
- Faster and faster

Technical advances in the past few years

- Optical model
 - GMM -> DNN
 - DNN -> LSTM
 - Sequential discriminative training of DNN/LSTM
- Language model
 - N-gram -> RNN-LM
- Decoding
 - Pruning algorithms designed for OCR
 - Automatic decoding parameter optimization

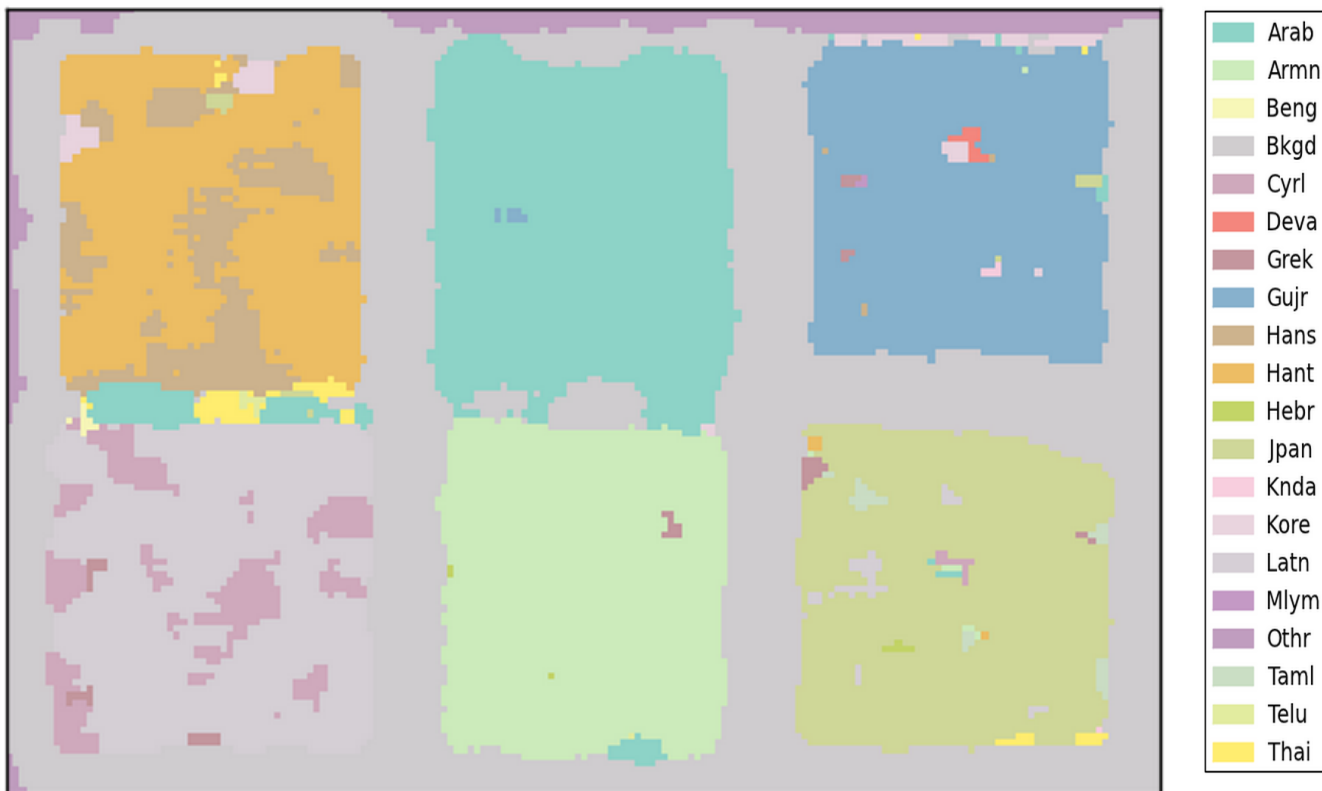
Script and Language Identification

- Some parameters usefully considered piecewise stationary latent processes
 - Font
 - Style (bold, italic,...)
 - Point size
 - Script
 - Language
 - Topic
- Most of these have low information rate → *exploit!*

Script ID approach 1: re-use OCR engine

- Script class seen as evolving as a hidden Markov process
- Pretend all letters of a given script are different glyph instances of the same “letter” (script class label)
- Do OCR with a very small vocabulary
- Reasonably accurate, significant hit on processing time
- Details: Genzel et al., “HMM-based script identification for OCR,” 2013

Alternative approach (Li et al., 2015)



Countries we don't cover



Part 3: Reflections and Comments

Unicode: a Godsend for OCR

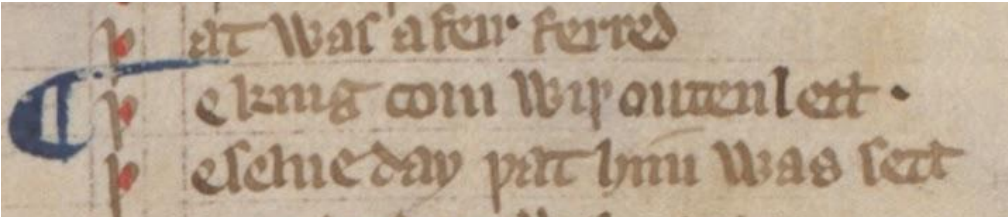
- Defining the goal requires specifying representation space
- Duality
 - Synthetic data
 - Document Image Decoding
 - Noisy Channel Formulation
- Internationalization libraries and resources, BiDi
- Corollary: OCR could not have been solved when it was most worked on

Changing styles, orthographies

മനുഷ്യരൽലാവരും തുല്യാവകാശങ്ങളോടും അന്തസ്സോടും സാമന്തൃത്തോടുംകൂടി ജനിച്ചവരാണ്. അന്യോന്യം ഭ്രാന്തഭാവത്തോടെ പെരുമാറുവാനാണ് മനുഷ്യനു വിവേകബുദ്ധിയും മനസ്സാക്ഷിയും സിദ്ധമായിരിക്കുന്നത്.

മനുഷ്യരൽലാവരും തുല്യാവകാശങ്ങളോടും അന്തസ്സോടും സാമന്തൃത്തോടുംകൂടി ജനിച്ചവരാണ്. അന്യോന്യം ഭ്രാന്തഭാവത്തോടെ പെരുമാറുവാനാണ് മനുഷ്യനു വിവേകബുദ്ധിയും മനസ്സാക്ഷിയും സിദ്ധമായിരിക്കുന്നത്.

Wir hoffen mit unserer "Gutenberg Presse" zur Wiederbelebung der Fraktur-
schriften - ohne jedweden politischen Nebengedanken - beizutragen. Leider verbieten
uns die hohen Produktionskosten eine Deutsche Version dieses Be-
nutzerhandbüchleins herauszugeben, Sie werden aber den Deutschen Text auf den
Programmdisketten finden. Bitte lesen Sie die „liesmich“ Datei für weitere
Informationen. Wir freuen uns auch über Ihre Kommentare und Anregungen.
Kontraktinformationen sind am Ende dieses Büchleins angegeben.



Then and now



Academia and Industry

- Strengths
- Evolving roles
- Cooperation

Can OCR finally be a “solved problem?”

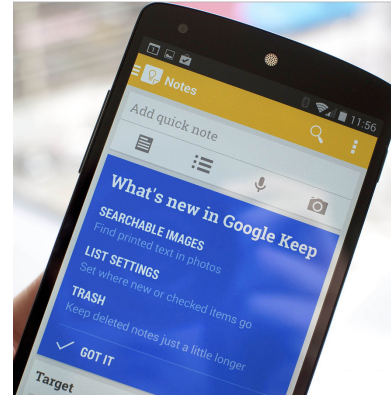
- Available to anyone, anywhere, ideally free-of-charge
- Network / cloud not required, keep your documents
- All languages, scripts, typefaces
- Quasi-linguistic: math, diagrams
- Regional libraries, cultural preservation efforts
- Newspapers, manuscripts, magazines, books



OCR in Google Products



Google Keep - notes and lists



Google Translate

