Google

5. Page Layout Analysis

Finding text regions on pages from books, magazines, and newspapers.

Ray Smith, Google Inc.

Background

- Historically Tesseract had no page layout analysis, but did have text-line finding, assuming a single column of text.
- Cube relies on Tesseract's page-layout/line finding.
- Tesseract's existing text-line finding is also weak wrt diacritics, especially for Arabic and Thai.
- Past methods tend to be:
  - (Bottom-up) Insufficiently aware of page-layout rules or
  - (Top-down) Insufficiently general.

Past Methods: Bottom-Up

- Analyze groups of pixels or connected components to classify into text/image/graphic/blank/line.
- Spread/smear/anneal groups of pixels by some neighborhood voting scheme, morphology or voronoi/graph algorithms.
- Find connected components of labels to group pixels into typed regions.
- Box-up regions into rectangles where possible.
- Morphological approach is very similar.
- Hard to include knowledge like "Columns should usually be the same size."

Past Methods: Top Down

- Often starts with a (possibly pre-trained) model of layout, eg 2-column journal page.
- Attempts to cut the image into the required parts, either with recursive vertical/horizontal cuts, or finding rectangles of whitespace.
- Methods usually fail on non-rectangular regions.
- Methods can often only deal with pages that fit the model.

New Method: Hybrid

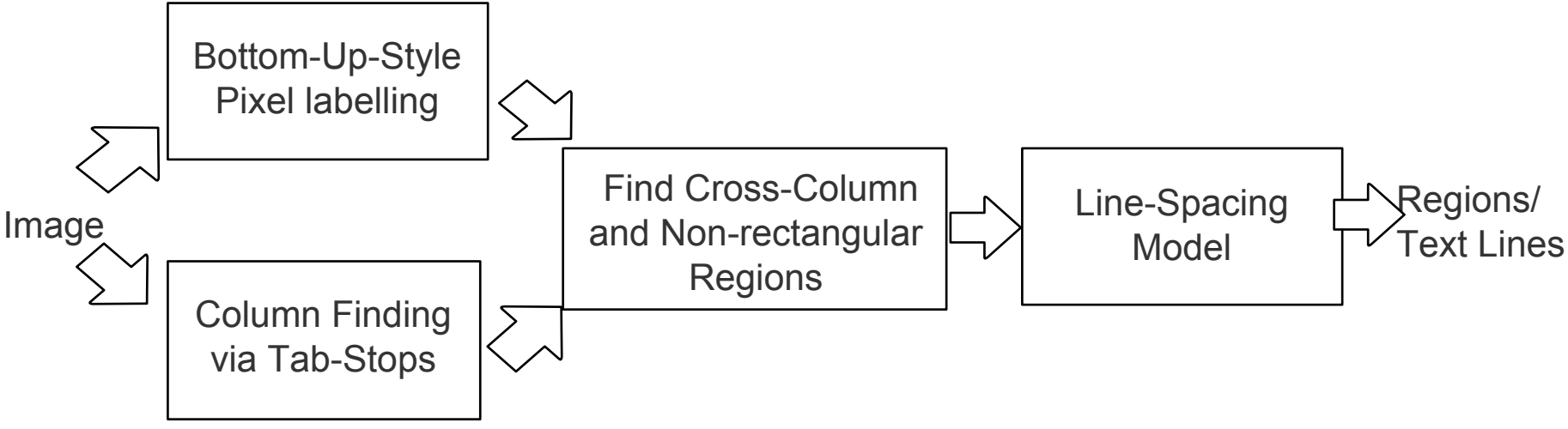


Image-Level Page Layout Analysis

Input Image

TECHNOLOGY

The Gist of Translation

How long will it be before machines make the Web multilingual?

By JENNIFER L. SCHENKER PARIS

IMAGINE IF CYBERSHOPPERS IN PARIS, Beijing and Rio could go to the Gap's website and find a description of the clothes—and an order form—in their native language. That day isn't far off. Machine-language translation—a process that uses computer software to translate text from one language into another—is maturing precisely at the time when the need for it, particularly on international websites, is escalating. Analysts figure that by 2003 about three-quarters of the 1 billion people communicating and doing business online will live outside the U.S., and most of them won't speak English.

Late last month Autodesk, a U.S. software company, began offering machine-generated translation to its European customers at 30% less than it costs to have humans do the job. In its first application of the software, Autodesk is providing French and Spanish versions of 5,000 online customer-support documents—the ones that tell you what to do when you run into technical problems. German and Italian will be added in about a month, with Asian languages to follow, says Mirko Plitt, a process analyst at Autodesk's global center for localizing content in Neuchâtel, Switzerland.

The company supplying Autodesk's translation software is Systran, which is based in Soisy-sous-Montmorency, France (www.systran-axff.com). It makes software that can translate more than 20 language pairs and provides the technology behind the free online translation services offered by AltaVista's BabelFish, Google, Lycos and AOL, as well as a wireless portal run by Oracle.

The quality of these free translation services tends to be uneven because the subject matter is so diverse and the process relatively imprecise. (It is called gisting, because it yields translations that provide the gist of a document.) What Systran offers Autodesk—and other multinational companies—is improved translation software customized for the context of the website. "Multinational organizations trying to provide quality service to all their customers, business partners and employees will want to have this capability," says Steve McClure, a research vice president at the technology consultancy IDC in Framingham, Massachusetts. Given Systran's head start, IDC says it holds the No.1 position worldwide for machine-language translation, ahead of IBM and Belgium's Mendez.

Machine translation has long been an elusive goal of computer experts. Like its human counterpart, the software has to take into account the grammatical structure of each language; it can't just give a substitution for every word unless it understands all of the words in a given sentence and how one influences the other. Early attempts at decoding natural languages through mathematical techniques using closet-sized computers produced poor results. Then Peter Toma, a linguistics researcher who began his work in 1957 at the California Institute of Technology, introduced linguistic rules into the process. He founded Systran in 1968. Among its first clients was the U.S. Air Force, which asked the company to develop a Russian-into-English machine-translation system. Later work for NASA caught the attention of the European Commission, which asked Systran to develop systems to translate European languages. In 1986 the company was purchased by Gachot, a French maker of industrial valves and fittings. Systran, which trades on the Nouveau Marché, last year had profits of $876,000 on revenues of $8.3 million.

For decades Systran's machine-translation technology was available only to U.S. and European government agencies and other public institutions via mainframe computers. Today its software runs on many different hardware platforms—including the type of Linux and Unix servers used by multinational corporations. At Autodesk the Systran software will not replace the people manning the multilingual help desk in Neuchâtel, says Plitt, but it's likely to make their jobs easier.

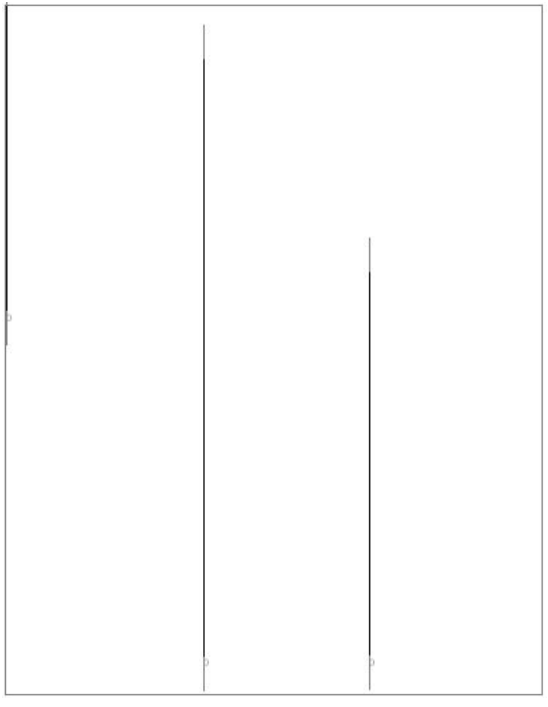
Systran has other multinational clients, including Ford and Price Waterhouse Coopers, that use the technology internally. But to create a mass market, Systran will need a lot more customers like Autodesk, which Systran CEO Dimitris Sahotalakis calls a "showcase client," putting the software on their external websites. If analysts are right, translating customer-support information is a significant first step, after which e-commerce applications should follow. Beyond that, the World Wide Web will be not just multinational but multilingual as well. ■

CAN WE TALK? Autodesk's Mirko Plitt says German, Italian and Asian languages are next

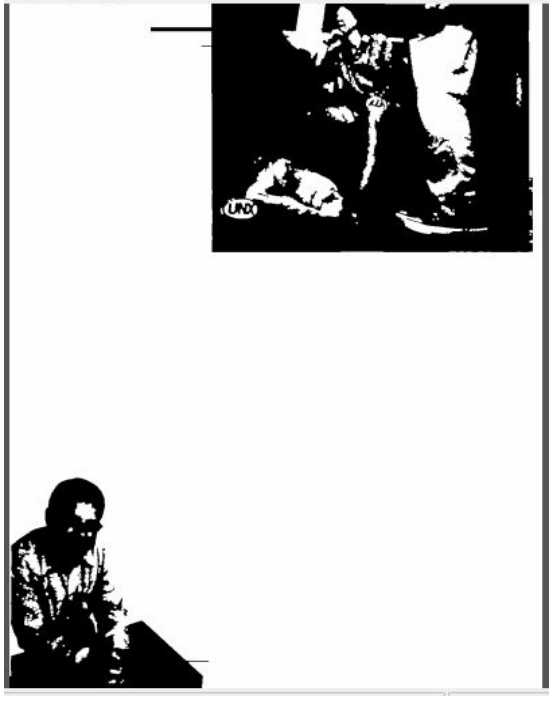
HELP DESK: Employees at Autodesk in Switzerland may find their jobs easier with Systran

TIME, JULY 10, 2000

Detected Lines



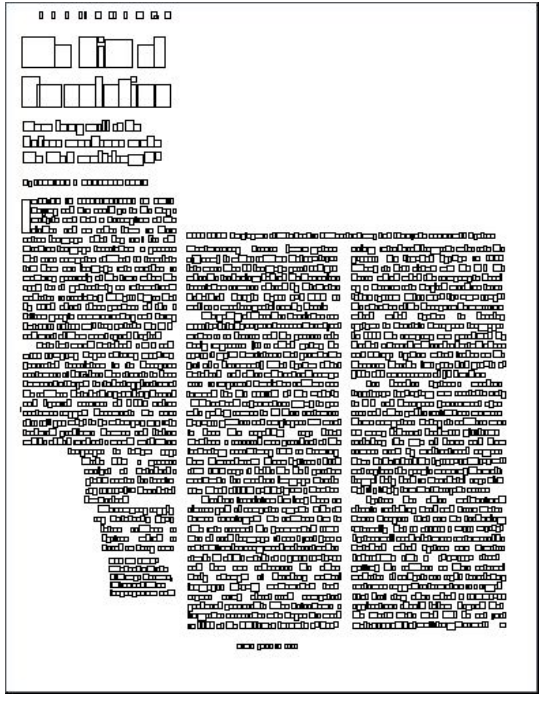
Detected Images



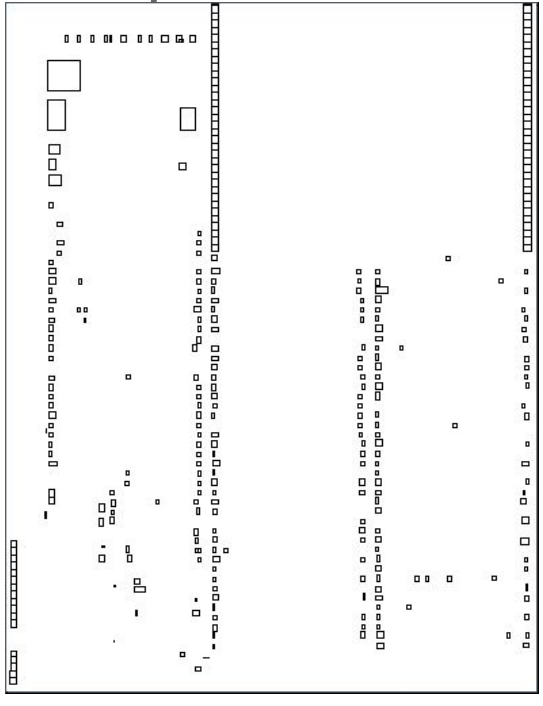
Orginal image from the Prima Layout anaylsis dataset: http://dataset.primaresearch.org

Connected Component Analysis

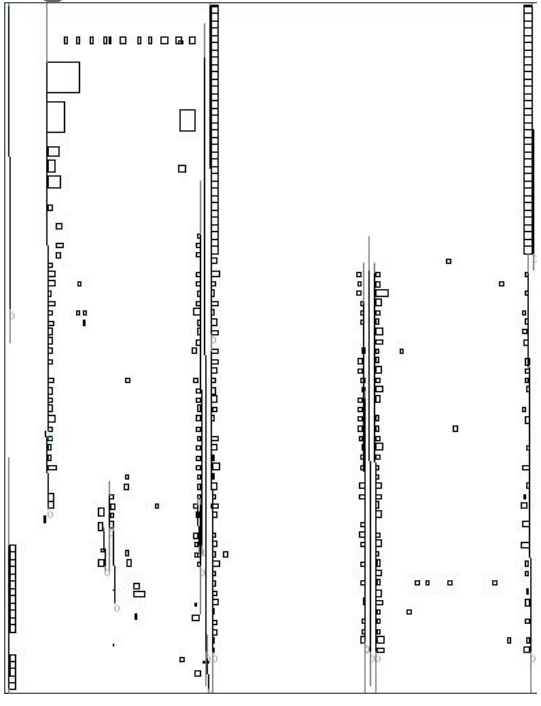
Text Components



Candidate Tab-Stop Components



Detected Tab Segments



Writing direction detection: Got Japanese?

Detect Local writing direction and do tab finding again...

大半、自由意思で居住

外務省、在日朝鮮人で発表

戦時徴用は245人

在日朝鮮人の北朝鮮帰還をめぐって韓国側などで「在日朝鮮人の大半は戦時中に日本政府が強制労働をさせるためにつれてきたもの

で、いまでは不要になったため送還するのだ」との趣旨の中傷を行なっているのに対し、外務省はこのほど「在日朝鮮人の引揚に関するいきさつ」について発表した。これによれば在日朝鮮人の総数は約六十一万人だが、このうち戦時中に徴用労務者として日本に来た者は二百四十五人にすぎないとされている。主な内容は次の通り。

一、戦前（昭和十四年）に日本内地に住んでいた朝鮮人は約百万人で、終戦直前（昭和二十年）には約二百万人となった。増加した百万人のうち、七十万人は自分から進んで内地に職を求めてきた個別渡航者と、その間の出生によるものである。残りの三十万人は大部分、土木事業の募集に応じてきた者で、戦時中の国民徴用令による徴用労務者はごく少数である。また、国民徴用令は日本内地では昭和十四年七月に実施されたが、朝鮮への適用はさしひかえ昭和十九年九月に実施されてお

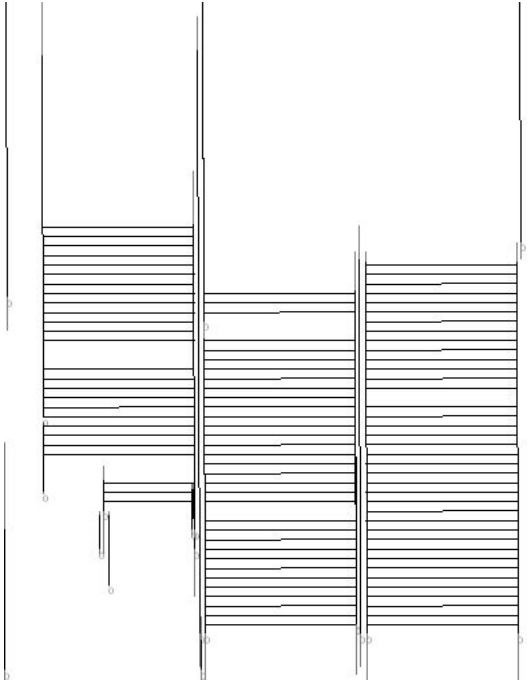
り、朝鮮人徴用労務者が導入されたのは、翌年三月の下関―釜山間の連絡が止まるまでのわずか七カ月間であった。

一、終戦後、昭和二十年八月から翌年三月まで、希望者が政府の援助、個別引揚げで合計百四十万人が帰還したほか、北朝鮮へは昭和二十一年三月、連合国の指令に基く北朝鮮引揚計画で三百五十人が帰還するなど、終戦時までに在日していた者のうち七五%が帰還している。戦時中に来日した労務者、復員軍人、軍属などは日本内地になじみが薄いため終戦後、帰国した者はごく少数である。現在、登録されている在日朝鮮人は総計六十一万人で、関係各省で来日の事情を調査した結果、戦時中に徴用労務者としてきた者は二百四十五人にすぎず、現在、日本に居住している者は犯罪者を除き、自由意思によって在留した者である。

By English: Asahi Shimbun 日本語: 朝日新聞 [Public domain], via Wikimedia Commons
https://commons.wikimedia.org/wiki/File%3AAsahi_Shimbun_newspaper_clipping_(13_July_1959_issue).jpg

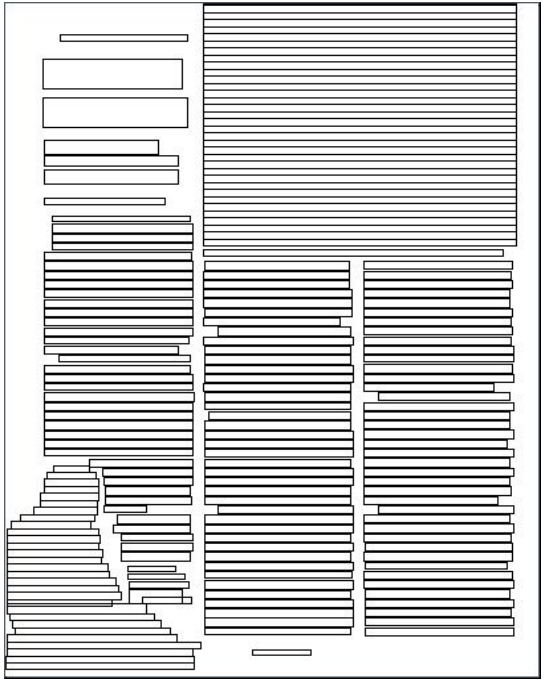
Column Finding

Connected Tabs



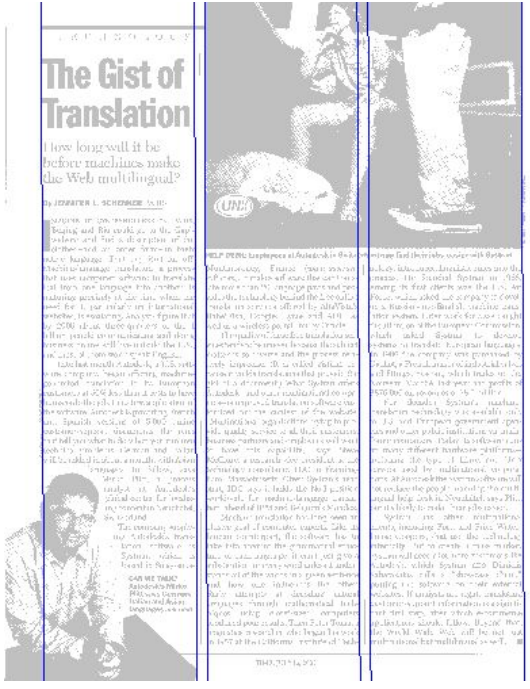
Validated Tab Segments

Candidate Column Partitions

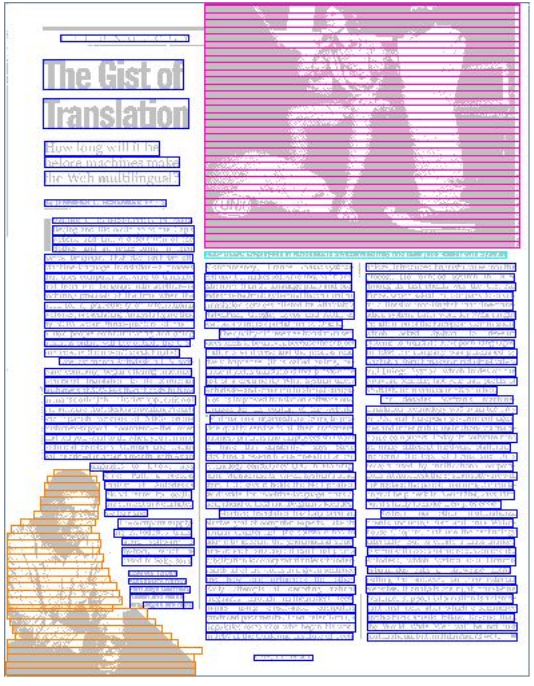


Block Finding

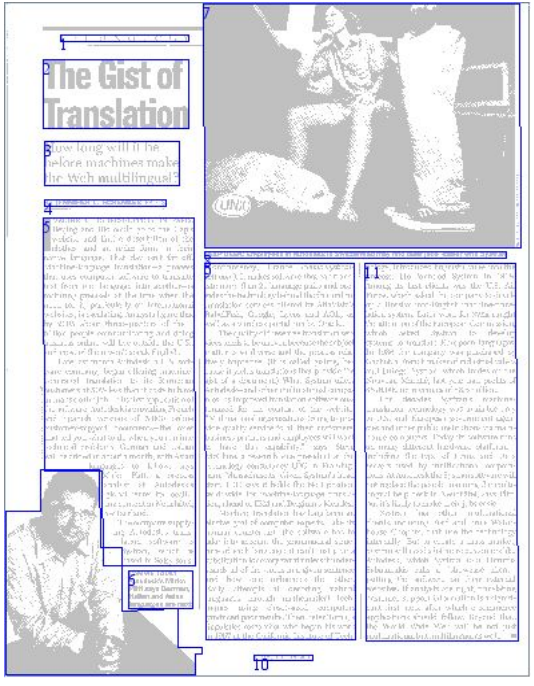
Detected Columns



Typed Column Partitions



Detected Blocks



Thanks for Listening!

Questions?