7. 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.

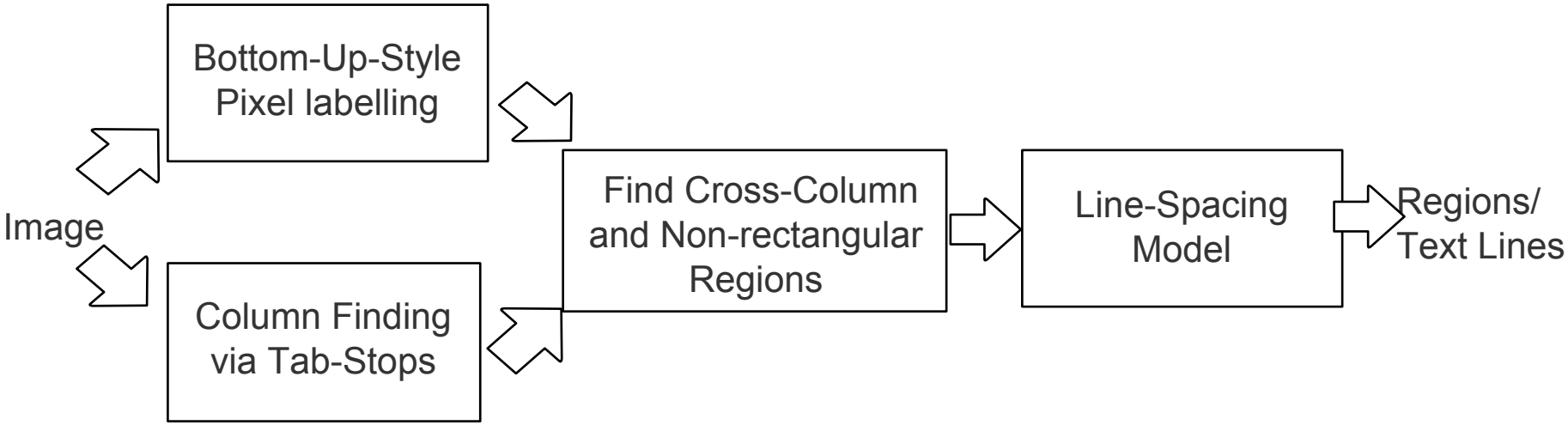
Past Methods: Bottom-Up

- 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



Google

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 2005 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 50% 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-soft.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.

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

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.

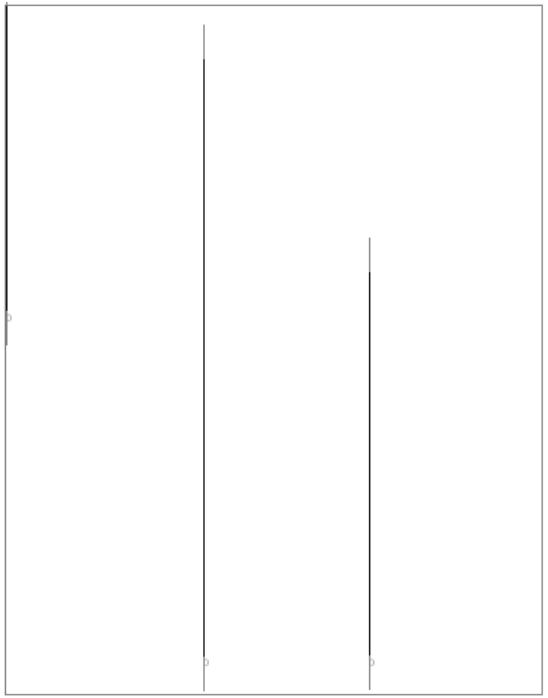
Machine translation has long been an elusive goal of computer experts. Like its human counterpart, the software has to take into account the grammatical structure of each language; it can't just give a substitution for every word unless it understands all of the words in a given sentence and how one 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.

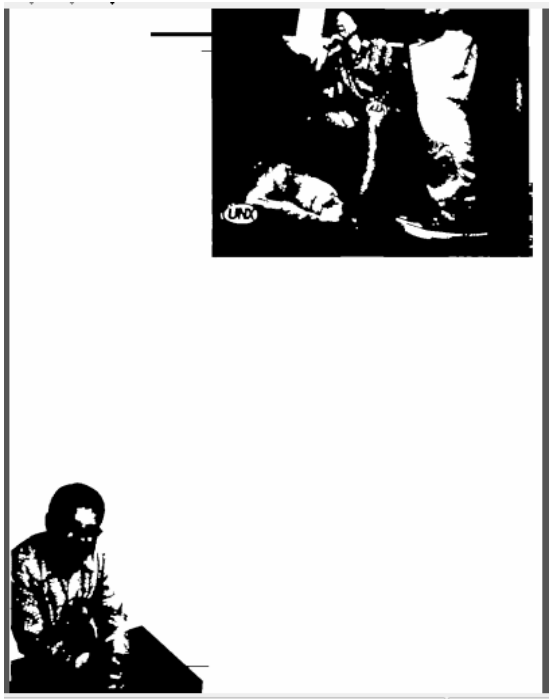
Systran has other multinational clients, including Ford and Price Waterhouse Coopers, that use the technology internally. But to create a mass market, Systran will need a lot more customers like Autodesk, which Systran CEO Dimitris Sabatakakis calls a "showcase client," putting the software on their 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. ■

TIME, JULY 30, 2001

Detected Lines



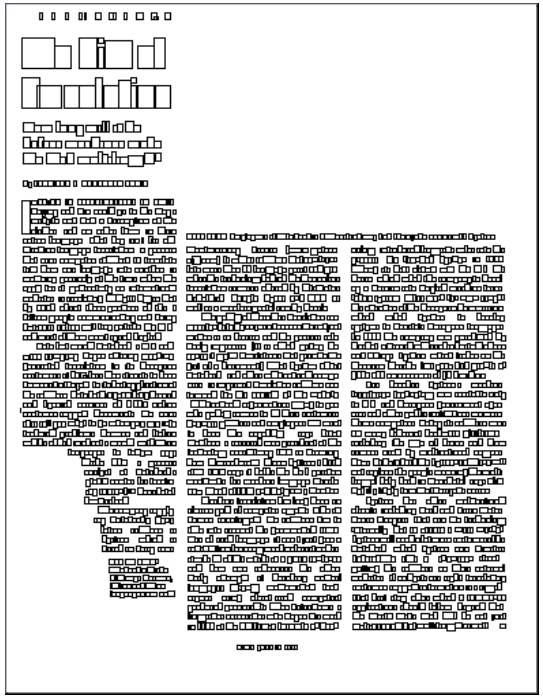
Detected Images



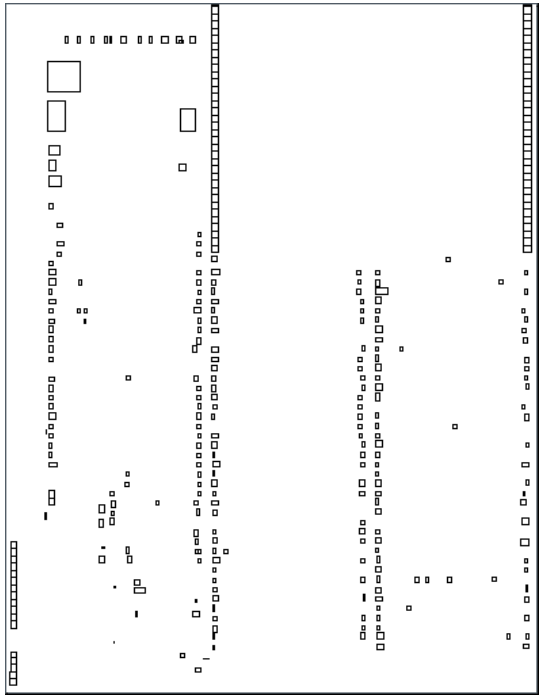
Google

Connected Component Analysis

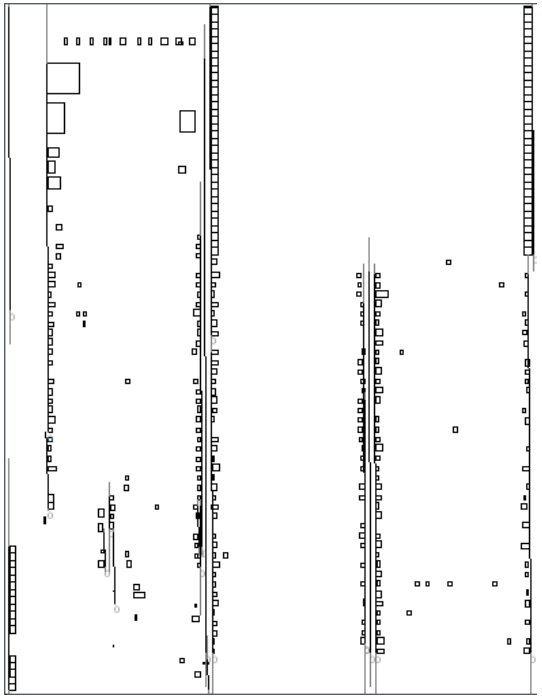
Text Components



Candidate Tab-Stop Components

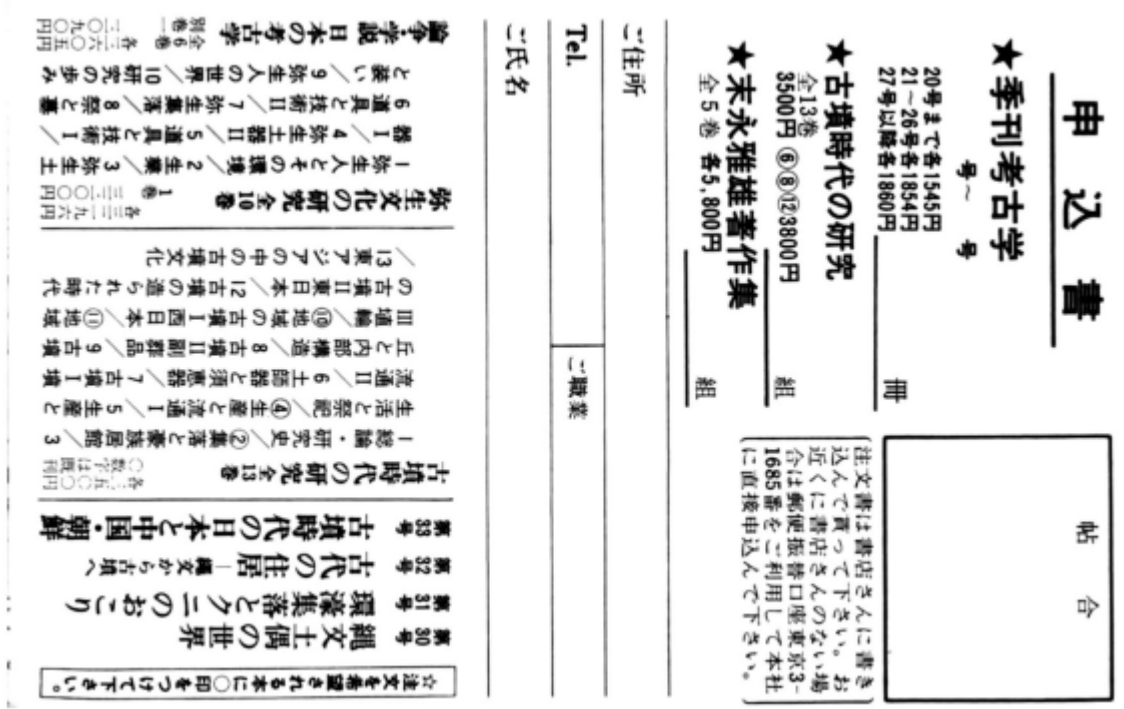


Detected Tab Segments



Writing direction detection: Got Japanese?

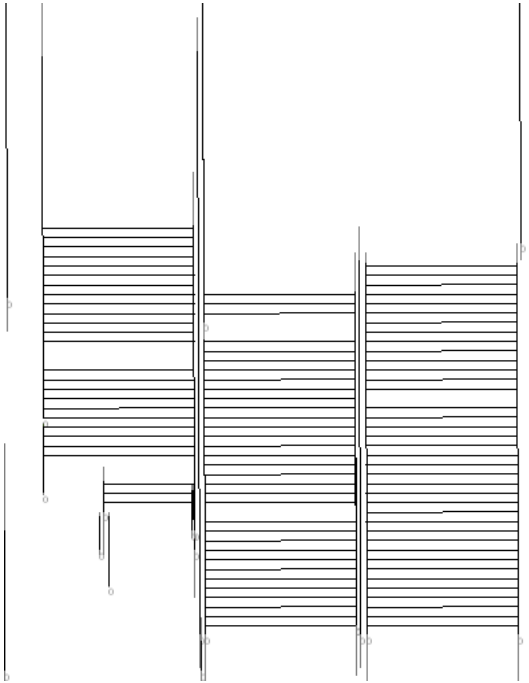
Detect Local writing direction and do tab finding again...



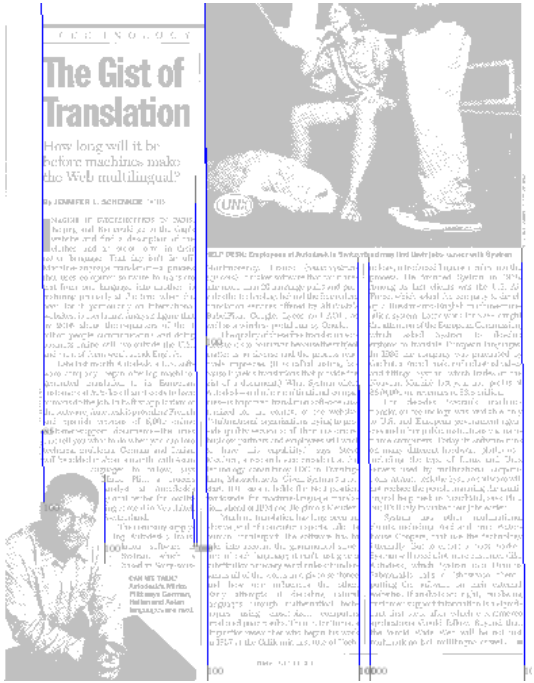
Google

Column Finding

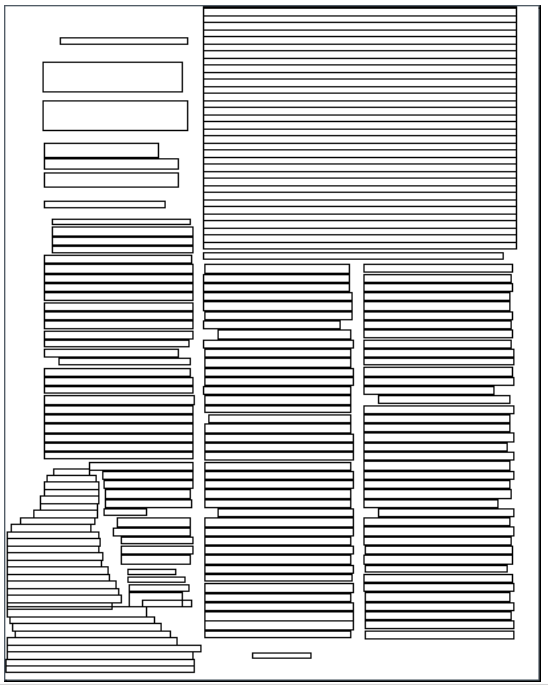
Connected Tabs



Validated Tab Segments

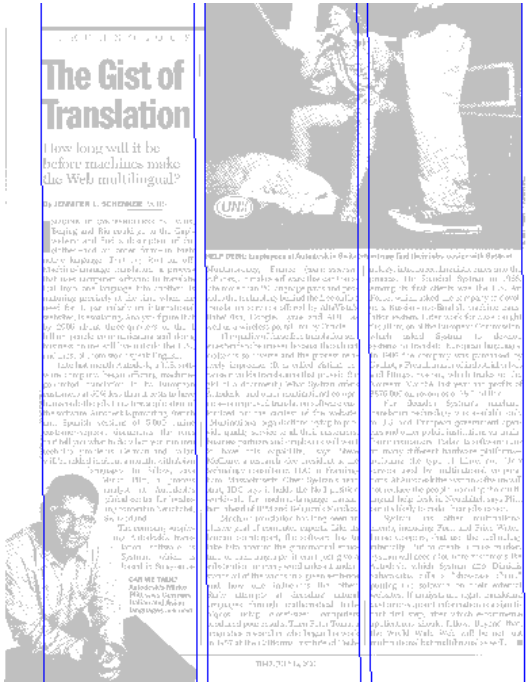


Candidate Column Partitions

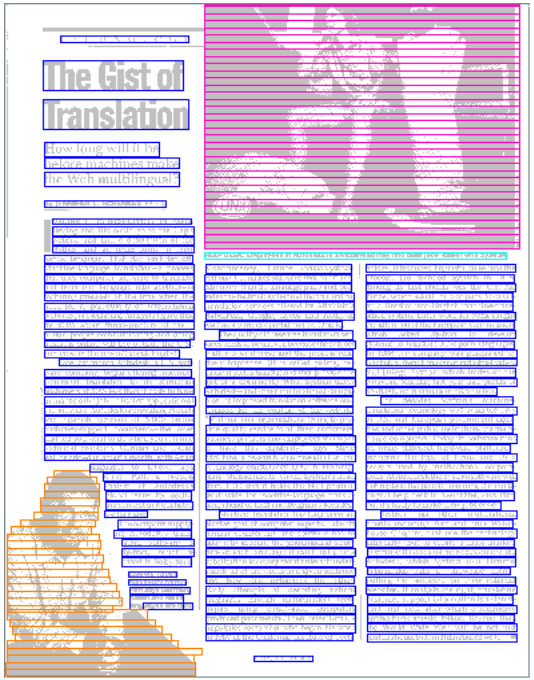


Block Finding

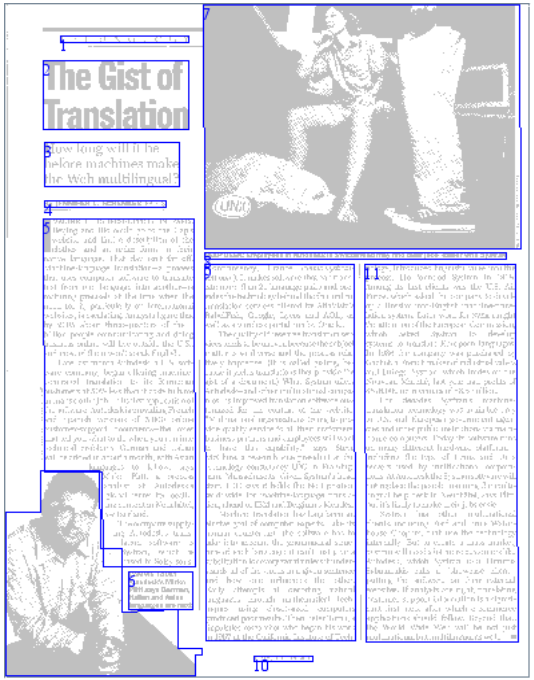
Detected Columns



Typed Column Partitions



Detected Blocks



Live demo of Page Layout Analysis

Simple magazine page with vertical text:

```
api/tesseract unlv/mag.3B/1/8001_044.3B.tif test1 inter segdemo strokewidth
```

Non-rectangular images:

```
api/tesseract unlv/mag.3B/0/8050_078.3B.tif test1 inter strokewidth
```

Multi-orientation Japanese:

```
api/tesseract -l jpn unlv/jpn.tif test1 inter segdemo strokewidth
```

Text on image:

```
api/tesseract unlv/mp00052bw.tif test1 inter strokewidth
```

Thanks for Listening!

Questions?