Why is graphics specialized?

We have general purpose retrieval software for text, and we’ve had it since the 1960s.

We have specialized graphics software for charts, maps, faces, CAD, abstract graphs, and so on.

Image retrieval in general still relies on words that can be found in the vicinity of the images.
Why is this baffling?

Words are not particularly precise. We have problems of synonymy and ambiguity. Depending on the language, we have problems of word order, declensional endings, and specialized phrases.

We can train pigeons to do some kinds of image retrieval; we can’t imagine training pigeons to do text retrieval.
You see these as the same

There is no similarity of color histogram, texture, or other simple image features. Nor, for that matter, is there any overlap with

cat
And for that matter


Right: Joan Miro, *Dutch Interior I*; Museum of Modern Art (NY).
Words, not the alphabet

This can’t be about the alphabet, since the major search engines do Chinese retrieval every microsecond.

There are about the right number of words. Letters are too common, and multi-word phrases are too rare.

Herman Simon suggested that 50,000 patterns was about the right number for an intellectually challenging but interesting activity. Shakespeare’s vocabulary is about 30,000 words; Jane Austen’s about 15,000.

There are too many different pictures.
It’s not the only difficult case

There are other examples of knowledge areas that we typically search with nearby words, despite better precision than we have with language:

*Mathematical theorems*

*Musical scores (aside from exact-match)*

and most puzzling to me

*Software*
We need a vocabulary or a model

If we could reduce these difficult problems in size by having a reasonable set of units, we could search those. Or, if we could fit them into a model, we could search the parameters – this is actually what is done in many domains. Chess is an example: we can scan pages from old books of games, recognize the symbols, and put the games into a formal notation and a large database of known chess games.

For imagery in general – whether vector or raster – we do not have an overall discussion model.
So images are too complex

If images are too complicated, then it helps to segment the space of graphical images. So we have specialized tools for the various kinds of imagery.

In recent decades we’ve seen ideas such as SIFT features or Gabor filters to try to find more generalized ways of describing images. Most of these are efforts to deal with photographs of naturalistic images, and great progress has been made; see Noah Snavely’s work on “Rome built in a day” which recognizes multiple amateur photographs of the same building.
Can we try to find elements?

Looking for repeated elements in a diagram isn’t that useful:

An easy image, but the similar items are not what is important.
What might work better?

After a suggestion from Venu Govindaraju, I tried looking at images drawn on a tablet. I asked a student, Annamarie Klose, to draw some pictures freehand. We captured the stroke timing as well as the final output (timings are slow because tablets are hard to draw upon). So for example:

- "apple" was drawn as 32 sec; stylus not lifted
- "apple" was drawn as 26 sec; 4 strokes
- "apple" was drawn as 3 sec, one stroke
It’s not always so neat

You can draw from the outside in..

was drawn as shown (top left first, bottom right last)
Still more complex

Steady polishing of the drawing to suggest texture
Little segmentation problem

System knows when the stylus was lifted or there was a pause; it then takes the segments and fits circles. Among 200 segments:

Circles on left, segments which are not circles on right. The degree measure is the gap in the circle.
But one can find elements

What about looking for circles in drawings?

The apple is pretty close to round, the hat really needs to be fit as a pair of ellipses.
A more challenging example

Here’s a search for circles in the grape drawing (left shows the circles on top of the original; in the right drawing only strokes not under the circles are shown).
Straight lines also work

Here are two more drawings with a match on similar features made of lines:
What do I hope to do?

The goal is to be able to search and to summarize.

Summarization, in this context, means making a simpler drawing with the same “meaning”.

Why? Let’s consider one specific application. Another student, Tony Gruenewald, works at LearningAlly (perhaps still better known as Recording for the Blind). They would like to be able to prepare tactile representations of the diagrams in the books and articles they read (see ACM SIGASSIST conferences for lots more on this problem).
Are circles and lines enough?

Realistically, no. People do not search at that level of detail, or even at the level of SIFT features. Nor would it make sense to say “I will abbreviate this picture by leaving out all horizontal lines” or anything like that.

So we are going to need something more ambitious and more semantically meaningful.

The goal is to take an image and produce a simpler one with a shorter description, that will be easier both to present and to compare with other images, without having to assume that the image comes from a particular population of well-modeled imagery.
Relevance to ICDAR

So we can zone a page, separate out the graphics, and then separate out the words inside the graphics.
(continued)

Top: whole excerpted image

Middle: the letters

Bottom: the graphics

Then you feed the words to an OCR program and eventually try to print them in Braille.
Another example of separation
(Continued)
A more complex example

where the buoys had been reported (3). They proved to be only logs of driftwood.
Por los puntos negros desde A hasta L anduvo la Fragata
Por los Colorados la Chalupá
Por los Verdes la Canoa

Leguas Castellanas

[Mapa de una región con leyendas en español]
Por los puntos negros desde A hasta L anduvo la Fragata
Por los Colorados la Chalupa
Por los Verdes la Canoa
Limitations of tactile media

The problem with this imagined pipeline – document to a separation into text (to be read aloud) and graphics (to be presented as a tactile image with Braille) – is resolution.

The standard for tactile presentation is that the minimum feature size is 0.1” while for a printed document a minimum feature size is more like 0.01” (good printing is at least 0.001”, but you’d a magnifying glass to see that properly).

So things must be really simplified.
A tactile map

This map is roughly 22x11 (standard sheet size is 11x11, and this map is double page).
How these are made today

The map is hand-engraved into an aluminum sheet.
Comparing maps

Cornwall, better resolution above, smoothed below
Simplifying maps

We have a model that lets us simplify maps. For example, choose a country, use fewer points on the outline, and give locations for four large cities.

But this is not so simple for other kinds of drawings.
What could you do with these?

Remember that in the tactile space we don’t have colors and textures (there is a small escape since there is a possibility of two different heights at each point).
Why is searching easier?

Nowadays, we can collect user observations: “people who looked at this image also looked at that one, or people who looked at text containing the words cat or dog then viewed the following images…”

Summarizing is harder.

We don’t usually have a way to tell which features of an image somebody looked at – even eye tracking will not help when the problem is knowing what property at a particular spot (color, density, connectivity, etc) is most important.
The value of time data

Sketches may give us a clue about what might be worth excerpting from an image. The right-hand picture is the first 10% of the strokes from the left-hand image.
Another example

This time, the right-hand picture is the first 16% of the strokes from the left-hand image.
Recent dataset from SIGGRAPH

There are 20,000 sketches from “How do humans sketch objects” by Eitz, Hays and Alexa (2012) online; they were studying recognition rather than decomposition, and some of their sketches were made less carefully (which may not matter). The data include sequence if not timing. Again, the sketches typically start with the outside. The strokes in the left part are the first 10% of the strokes in the drawing.
Strategy for future

Suppose we try...

By having people sketch, define a set of examples of abbreviated images.

Then train an algorithm to abbreviate images.

Given simpler images, isolate segments that are more general.
What about the other hard cases?

I’m not a mathematician, nor a musician, so I’m not sure what to do about either theorems or scores. Theorems seem harder – you can’t just say “leave out the subscripts”. Music search is an active area; exact-match works (copyright infringement) and more technology is on the way.

Software has me baffled. There are no ambiguities, and the set of syntactic operators is usually fairly small, and the problem is just that the variable names are irrelevant. I can envisage searching for a sequence of syntactic structures, or using them as index terms, but no products yet. There is promising research at this conference; see poster by Tuarob, Bhatia, Mitra and Giles (Automatic Detection of Pseudo-codes in Scholarly Documents Using Machine Learning).
How are things changing?

In 1985 I gave a similar talk, at a SIGIR conference. At the time, drawing maps was new, so I talked a lot about maps. But the general problem is still the same. To quote from that paper:

“Today successful graphics routines contain a great deal of local domain knowledge. There is no analog of the simple keyword systems that handle textual documents in any subject area. Just as computational linguists have found that subject matter expertise is necessary to do really sophisticated processing of English, it seems also necessary to sophisticated processing of pictures; the difference is that we don't know how to do unsophisticated processing of graphics.”
Conclusion

I’m not denying progress. One step forward was SIFT features, and another was recognizing the power of visual dictionaries (exact matching).

But we still need better and larger scale methods for handling many kinds of visual information. And I look forward to methods that, without using words, will search across CAD databases, photographs, and paintings. There is research in this conference about higher level features – we need more of that.