Limitations of Content-based Image Retrieval

Slide set for a plenary talk given on Tuesday, December 9, 2008 at the International Pattern Recognition Conference at Tampa, Florida.

The set includes a few additional slides that had been omitted from the original ICPR presentation because of time limits. The title box of such slides has a gray background.

Limitations of Content-based Image Retrieval

Theo Pavlidis

Stony Brook University
(formerly with Symbol Technologies)

www.theopavlidis.com

Note

Most of this presentation is based on a paper with the same title, posted on my web site in May, 2008:
technology/CBIR/PaperB/vers3.htm

References to that paper (citations, sections, appendices) in the slides are always inside **[ ]**.

A summary of that paper can be found in
technology/CBIR/summaryB.htm

References in **[ ]** are to
technology/CBIR/PaperB/icpr08.htm

Limitations of CBIR - Outline

- What is the true current state of the art?
- Methodological Problems of General CBIR
- Why is CBIR so Hard?
- What can we learn from the past?
- What can be done?

The True Current State of the Art

- Title of Editorial in Special Issue of IEEE Proceedings (April 2008): ***"The Holy Grail of Multimedia Information Retrieval: So Close or Yet So Far Away?"***
- **Close** if we take published results at face value.
- **Far Away** if we evaluate results from online test sites or look closely at the published results.
- The answer also depends on what do we mean by CBIR.

Two Kinds of CBIR

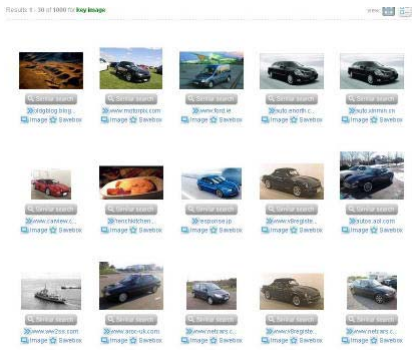
- **General:** We try to match a query image to an arbitrary collection of images, such as those found on the web. The goal of the query is to obtain images with the same object as the query. Such CBIR imitates web search engines for images rather than for text.
  - Given an image with a horse, find all images showing a horse (at least as their main subject).
- **Application Specific:** We try to match a query image to a collection of images of a specific type. For example, fingerprints, X-ray images of a specific organ, images of skin lesions, etc.

Results from Online Tests of General CBIR

- Two types of tests are publicly available: image retrieval and auto-tagging.
- Fewer sites are actually available than advertised.
  - For example, *Cortina* (cited in a Nov. 2008 PAMI paper) is not operational except for already tagged images.
- Results are generally poor. Only one site (GazoPa) produced a good match and that was only once. See [Appendix A] and [Recent Tests]. In some sites the system failed to produce any results.

Shopping for Shoes (or Cars?)

(GazoPa – November 25, 2008)



Capturing the overall shape is not enough!
There is a semantic abyss rather than just semantic gap.

Shopping for Shoes (cont)
(GazoPa – November 25, 2008)

Adding a tag (SHOE) has not helped.

An Auto-tagging Result
(From a new site, October 2008)

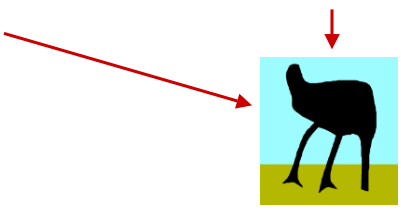
Rest, chairs, **architecture**, animals, Europe, church, boats, livestock, ports, **city**, Italy, the sea, **building**, boat, beach, **housing**, harbor, holiday

Another Auto-tagging Result
(From the same site, October 2008)

Mammals, show, Business Woman, animals, black, business, attitude, full, office workers, business, computers, office, smiles, close-up, businessman, adults, parents

Response of Test Site Owners*

1. "The Problem turned out to be much harder than we thought – we have given up on general CBIR …"
2. "We are still trying to improve our site …"
3. "Student who maintained the site has left", etc.
4. Silence!



* Also of Authors

Limitations of CBIR - Outline

- What is the true current state of the art?
- Methodological Problems of General CBIR
- Why is CBIR so Hard?
- What can we learn from the past?
- What can be done?

Methodological Issues: Solutions in Search of a Problem - 1

Many papers describing CBIR methods use trivial queries, for example:

- "Show all pictures with buildings" (rather than a particular building)
- "Show all pictures with people" (rather than a particular person)
- "Show all pictures with a lot of green"

- They do so because these are the only kind of queries that can be answered by the methods used.

Methodological Issues:
Solutions in Search of a Problem - 2

- A classifier is trained on a finite set of classes of objects. The retrieval system is limited to such classes. It cannot deal with an open collection of images.
- Such training will need labeled samples, but if we label all images we do not need CBIR.
- (Using trained classifiers is fine for application specific CBIR where object categories may be well defined.)

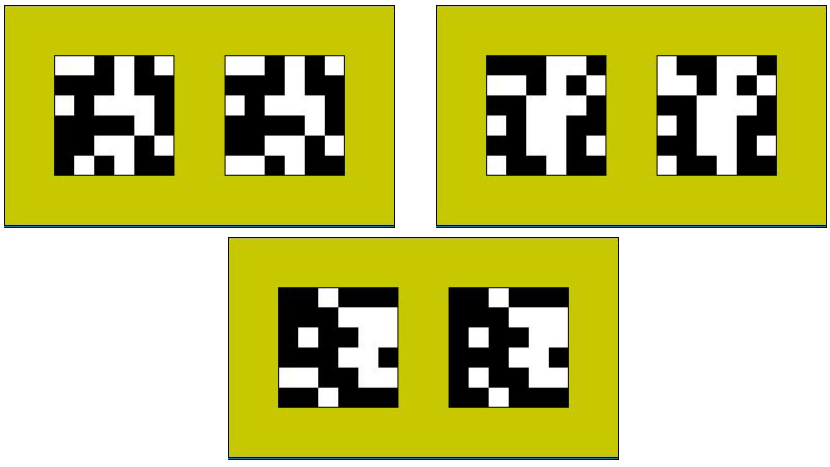
Methodological Issues:
Underestimating the Image Space

- Methods are developed over too "small" a set of samples: We deal, in effect, with hashing schemes. No collisions in small sets but many collisions in large sets.
- The possible number of images for a given size is huge. The number of binary images on a 10x10 array is
  $2^{100} > 10^{30} >$ trillions of trillions.
  For a 6x6 array it is
  $2^{36} > 64*10^9 = 64$ billions.
- Even "80 million images" (part of the title of a PAMI Nov. 2008 paper) cover too small a part of the space of all images.

Human Discriminating Ability

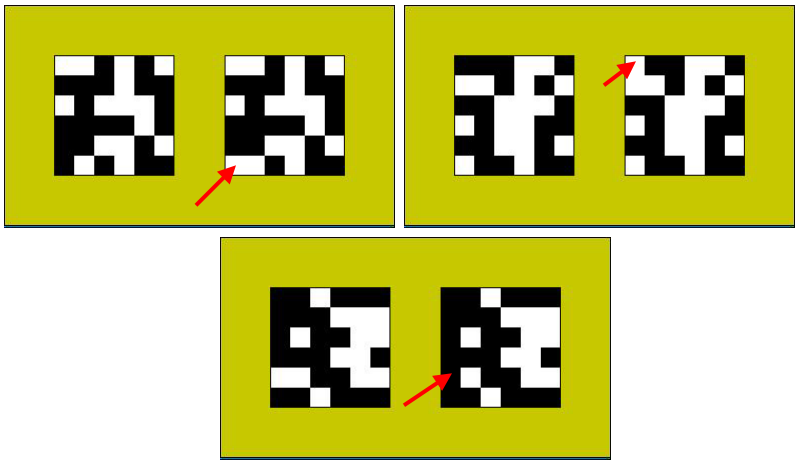
- People seem to be able to discriminate 6x6 arrangements, so that 64 billion distinct images seems a realistic lower bound.
- In the following two slides we show three pairs of random patterns that differ in just one location. Because of the lack of order this is the most difficult case for discrimination.
- The second version of the patterns marks the differences.

Some 6x6 Boards



See the next slide for help in finding differences within each pair

Some 6x6 Boards



Methodological Issues:
Confusion between Looking for Similar Images or Looking for Similar Objects

- Many papers (incl. some in the Nov. 2008 issue of PAMI) are vague on whether they search for similar 2D images or similar 3D objects contained in images.
- Difference in viewpoint/pose and illumination offer serious challenges to methods based on simple features.
- Even segmentation and techniques such as “salient” point matching cannot deal with viewpoint/pose issues, especially for “articulated” objects.

What is Needed in the Real World

- Infamous example: “Find all pictures of president Clinton and Monica Lewinski.”
- The images in the data cannot possibly be tagged because ML became famous only during the impeachment proceedings.
- The need to perform such open queries requires that we have general image similarity measures that allow for detailed matching. (Possibly comparing parts of images.) **Scene similarity measures** would be even better.
  - Text search works by similarity measures.

The big challenge: Perceptual versus Computational Similarity

- Two pictures may differ a lot in their pixel values but appear similar to an observer. (“They have the same meaning”.)
- Two pictures may differ in few pixels only but they have different meaning. (Face portraits of two different people in front of the same background.)
- *By the way:* Color is not particularly useful because of the isoluminance effect [Bach_02].

Two Images with Nearly Identical Color Histograms

More examples in [Appendix C]

Perceptual versus Computational Similarity – An Example

Perceptually close

Computationally close (similar pixel statistics)

Conceptual Similarity is Even Harder to Deal With

Conceptually Close

Computationally close
(Large areas with similar local pixel statistics)

In Summary:

- Pixel values have little to do with the human interpretation of an image.
  - Unfortunately, comparing images pixel-wise persists (papers in Nov. 2008 issue of PAMI)
- There is not just a semantic gap (to be patched with heuristics)
- There is a **SEMANTIC ABYSS** requiring new methodologies.

Limitations of CBIR - Outline

- What is the true current state of the art?
- Methodological Problems of General CBIR
- Why is CBIR so Hard?
- What can we learn from the past?
- What can be done?

Text versus Pictures

- In text files each byte (or two) is a numerical code for a character. Therefore strings of bytes correspond to words that, in turn, carry semantic meaning.
- In pictures each byte (or group thereof) represents the color at a particular location (pixel). Pixels are quite far from the components that have a semantic meaning.

By the way, we do not that well in text!

- If it is hard to search for concepts unless we can map concepts into words.
- Example 1: Find all articles critical of the government policy in dealing with the banking crisis.
- Example 2: Find all articles about a **dog named Lucy**. Amongst the Google returns was an article with the phrase: "**Lucy** and I spent the weekend alone together. We have a **dog named** Kyler."

Dealing with images has to be much harder than dealing with text.

- The human visual system has evolved from animal visual systems over a period of more than **200 million** years.

- Speech is barely over **100 thousand** years old.
- Written text is about **5 thousand** years old.

What Human-Vision Scientists Say

- Bela Julesz [Ju91]: “In real-life situations, bottom-up and top-down processes are interwoven in intricate ways," and "progress in psychobiology is ... hampered ... by our inability to find the proper levels of complexity for describing mental phenomena“
- V.S. Ramachandran: [RB98, p. 56] “Perceptions emerge as a result of reverberations of signals between different levels of the sensory hierarchy, indeed across different senses”. He then goes on to criticize the view that “sensory processing involves a one-way cascade of information (processing)”
- Richard Gregory [Gr98]: "Perceptions are predictive hypotheses, based on knowledge stored from the past". The idea goes back to Helmholtz (19th century). See also the discussion by Papathomas [Pap99].

A Neuroscience Sampler

- Research published in *Nature Neuroscience* in November 2008 has found “evidence for explicit neural code for complex three-dimensional object shape” in the brain of monkeys [YCBWC08] .
- An editorial in the same issue of the journal is titled *“So many pixels, so little time”* [Ma08] and points out that “the primate visual system is composed of 25-40 distinct areas, depending on how they are counted.”

Challenges to Machine Vision (and not just to CBIR)

- We need to replicate complex transformations that the (human/animal) brain has evolved to do over hundreds of millions of years.
- We have to deal with the fact the processing is not unidirectional and also affected by other factors besides input (context both inside and outside the image). Visual illusions (far more common than auditory illusions) attest to that fact.

Machine Vision must deal with "Multi-directional" Processing

- Awareness on the importance of reconciling "high" and "low" levels of knowledge in Machine Vision is old (see review by Tsotsos [Ts84]) but relatively little research has been done over the years.
- A search of Keith Price's Vision Literature site yields the following:
  - For "top-down bottom-up" 346 entries (many from the document processing area).
  - For "segmentation" 5750 entries.
- "Multi-directional" Processing is not only inspired from Neuroscience.

It is also good Engineering in its own right!

Limitations of CBIR - Outline

- What is the true current state of the art?
- Methodological Problems of General CBIR
- Why is CBIR so Hard?
- What can we learn from the past?
- What can be done?

The OCR Experience

- Considerable work in "simple" OCR algorithms that could run fast enough on computers of the 1960's or 1970's turned to be a waste of time because such techniques could not provide satisfactory performance.
  - By 1990 there were machines fast enough to implement more "complex" algorithms (invented 20 years earlier).
- Earlier scanners of 100 or 150 dpi tended to distort images of characters significantly, for example mapping of "a", "e", and "s" into a form topologically equivalent to an "8". There were efforts to recognize the characters on slight variations of shape.
  - This problem went away with the introduction of 300dpi resolution scanners.

Fingerprint Recognition

- Fingerprint recognition by computer is now practical and is used widely in law enforcement [Section 6].
- It represents a major CBIR success but lessons from it, the use of application specific features, seem to be ignored.

Matching Images to Maps

- Matching the location of an aerial photograph to a particular geographic area is of interest in surveillance.
- Over ten years ago we used a graph representation of the roads in aerial photographs and shape features of the branches to match the location using information in stored road maps.
- The project was sponsored by Grumman Data Systems and it is described in a paper by J. Hu and T. Pavlidis [HP96].

Limitations of CBIR - Outline

- What is the true current state of the art?
- Methodological Problems of General CBIR
- Why is CBIR so Hard?
- What can we learn from the past?
- What can be done?

Face the Situation

General Strategy

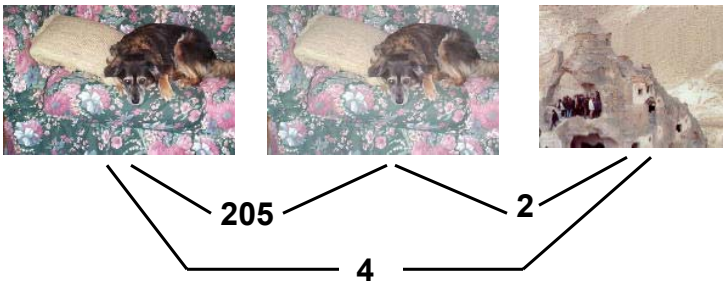
- Accept that practically significant results for real time general CBIR cannot be obtained unless there are major breakthroughs both in Image Analysis and in Computer Architecture.
  - As long as we do not have general segmentation methods that can identify objects on an image, it is unwise to pursue general CBIR.
- Instead we should pursue research focused on **specific CBIR** problems that satisfy certain feasibility criteria.
  - A list of criteria will be given after we look at two successful methodologies: SIFT key-point matching and graph matching.

SIFT

- SIFT (Scale Invariant Feature Transform) key-points [Lo04] seem helpful as a tool for 2D (but not 3D) matching.
  - When two **images** appear almost identical to the human observer they have a large number of matched key-points.
- SIFT key-point matching seems to be the best technique we have now, although it has some weaknesses.
- **Can we find transforms that are guaranteed (by analysis) to improve upon SIFT?**

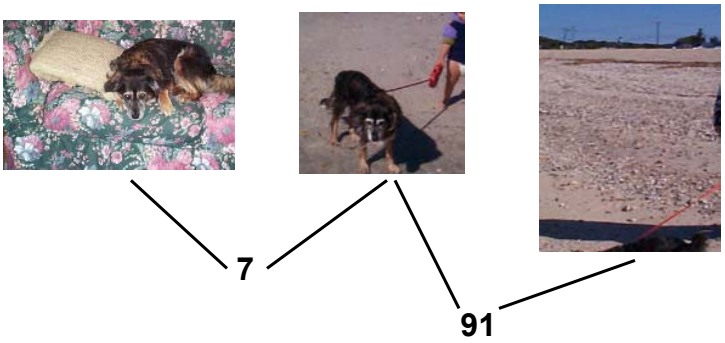
Number of matched SIFT Key-points – I
SIFT does well in 2D matching (see [Pa08] for details)

Meeting a Challenge



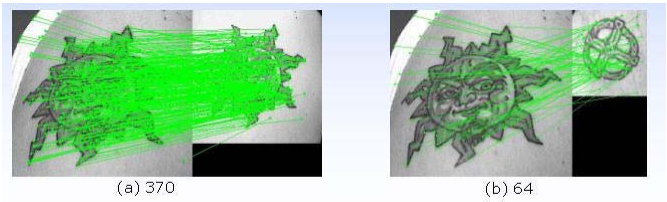
Results thanks to Ms. Jung-Eun Lee, a student of Prof. A.K. Jain at Michigan State University

Number of matched SIFT Key-points – II
SIFT fails in 3D object matching (see [Pa08] for details)



Results thanks to Ms. Jung-Eun Lee, a student of Prof. A.K. Jain at Michigan State University

Application of SIFT to Biometrics



Numbers of matching points between (a) similar and (b) different images

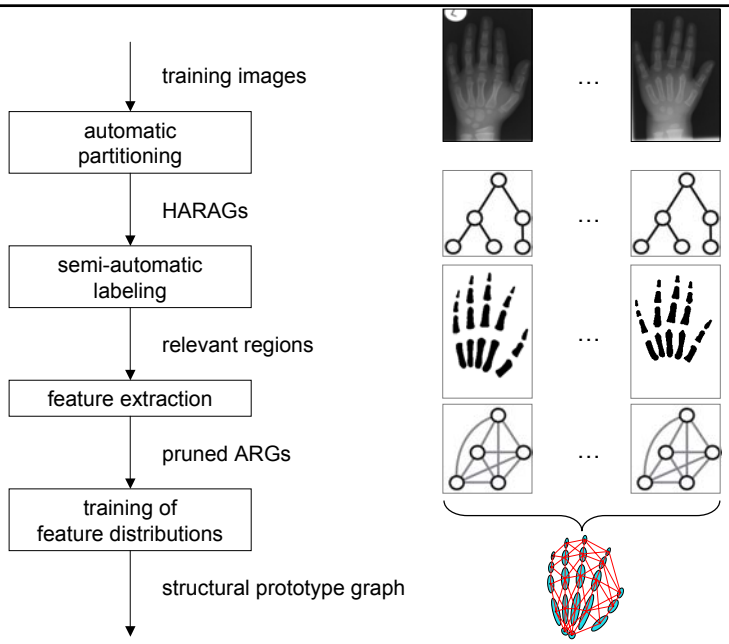
Courtesy Prof. Anil K. Jain from his presentation on "Scars, Marks, and Tattoos: Soft Biometric for Victim and Suspect Identification" [LJJ08]

Comments on Meeting a Challenge

- Because SIFT key-point matching requires no training, it can be applied on any set of images.
- Methods relying on training cannot possibly handle the challenge.
  - *"(our system) would be able to select photos of a particular dog … if given several examples with the dog and other examples without the dog especially if there were no other very similarly colored dogs in the data set."* e-mail from author of a training based paper.

HARAGs

- Matching of **H**ierarchical **A**ttributed **R**egion **A**djacency **G**raphs have been used in the matching of bones in hand X-rays by B. Fischer *et al* "Scene Analysis with Structural Prototypes for Content-Based Image Retrieval in Medicine", [FSGD08].
- The next slide is courtesy of Prof. Thomas M. Deserno of Aachen University, Germany.



Criteria for solvable CBIR problems

1. The mapping between semantics and image features is well defined.
2. Top level knowledge and/or context are known.
3. The accuracy requirements are well defined.
4. The computational requirements are well defined.
5. The matching of images requires careful scrutiny.

Mapping between Semantics and Image Features

- The representation by features should be possible so that only images with similar interpretation (and no others) are mapped onto the same set of features.
- Examples abound in the pattern recognition literature, including OCR and fingerprint recognition.
- If we deal with a specific application we may also be able to incorporate high level knowledge into the methodology.

Accuracy Requirements

- An application should have its own:
  - Absolute accuracy requirements.
  - Relative significance of false matches versus omitted true matches.
- A reminder from OCR
  - Older papers reported recognition rates of 97-98%, a rate useless in practice because it corresponds to over 50 errors per page. For a practical system the recognition rate should be at least 99.9% (2-3 error per page) and the *errors should be rejections (rather than substitutions)*. The tolerable rate for substitution errors is much lower.

Computational Requirements

- In many cases instant response is not needed.
  - In a medical application it may take well over an hour to produce an image, so waiting another hour to find matches in a database is not particularly onerous.
  - Auto-tagging can be done in the background, so speed is not critical.
- If a fast response is needed the database may be organized for fast retrieval.

The matching of images requires careful scrutiny

- Humans are not very good at careful scrutiny and it is likely that machines can match or even exceed human performance. (This has been an important factor in the success of automatic industrial inspection.)
- Medical, industrial, and biometric, amongst other applications, seem to fit in this category.

Where does Shopping CBIR fit?

- Users are supposed to submit a picture of a product and find through CBIR web sites where that product is sold.
- If the product is broadly defined, or if the list to be searched is small, CBIR may succeed, otherwise it will fail for the same reasons as Veggie-Vision.
  - Veggie-Vision could tell tomatoes from eggplants but not organic tomatoes from conventional tomatoes, so it never made it to the check-out counter.
- **The devil is in the details!**

Concluding Suggestions

- We should focus on matching 2D images rather than 3D scenes to avoid issues of pose, viewpoint, and large variations in scale.
- There are plenty of challenging applications needing 2D image matching: Medical, Biometric, Forensic, Industrial, Security, etc.
- It is better to really solve a special case of CBIR than pretend to solve the general CBIR problem.