A Benchmark for Image Retrieval using Distributed Systems over the Internet: BIRDS-I

Neil J. Gunther

Performance Dynamics Consulting

Giordano Beretta

Hewlett-Packard Laboratories

13th IS&T/SPIE Symposium on Electronic Imaging—Science and Technology

Systems & scalability

The Internet is a visual medium

- A picture is worth ten thousand words…
  - A screenful of text requires 24×80 = 1,920 bytes
  - A VGA size image requires 640×480×3 = 921,600 bytes
- …but requires almost 500 times as much bandwidth…
  - data compression is essential for images on the Internet
  - which compression is best for my image?
- Need to maintain user response time within acceptable range when aggregate Web traffic increases
- The Web is a system of mostly unspecified elements
- Need to *measure* the scalability of application services

User experience expectation

Internet bandwidth, processing power, data size

- Today's user experiences set very high expectations for imaging on the Internet
  - many users access the Internet in the office on fast workstations connected over fast links to the Internet
  - at home users often have fast graphics controllers for playing realistic computer games
  - increasingly, private homes are equipped with fast connections over DSL, cable modem, satellite, …
  - the latest video game machines are very powerful graphic workstations
  - the new generation grew up on video games & WWW
- People have about 2T bytes of own data and want to access all of the Web
  - at work, they expect concise answers immediately on multiple media

The real world

The nomadic workforce

- The new working world is mobile and wireless
  - a comprehensive fast fiber optics network provides a global backbone
  - the “last mile” is wireless
  - computers are becoming wearable (PDAs)
- Displays and power are the main two open problems
- To conserve power, wireless devices will have low effective transmission bandwidth and small display areas
  - humans generate 80 W while sleeping and 400 W walking
  - thermo-generator in a shirt can harvest 5 W
  - piezo-generator in a shoe heel can produce 8 W
- Concomitantly the new users are impatient

Image retrieval

Text or contents based

- Text-based image retrieval: images are annotated and a database management system is used to perform image retrieval on the annotation
  - drawback 1: labor required to manually annotate the images
  - drawback 2: imprecision in the annotation process
- Content-based image retrieval systems overcome these problems by indexing the images according to their visual content, such as color, texture, etc.
- A goal in CBIR research is to design representations that correlate well with the human visual system

Retrieval metrics

Exact queries are not possible for images (nor text)

- Recall (Sensitivity) = Number of relevant items retrieved / Number of relevant items in database
- Precision (Specificity) = Number of relevant items retrieved / Number of items retrieved
- CBIR algorithms must make a compromise between these two metrics: broad general vs. narrow specific query formulation
- CBIR algorithms tend to be very imprecise…
- …the result of a query requires ongoing *iterative refinement* (Leung, Viper)

Ground truth

- Precision and recall are determined with the aid of a data structure known as the ground truth
  - embodies the relevance judgments for queries
  - done only for the ground truth not for each image as in text based IR
- Two problems
  - appropriate image categories must be defined
  - images must be categorized according to those definitions
- Challenge: scalability
  - gestalt approach
  - decoupled ontologies
  - categorization supervised by experts
  - ground truth data structure created by program

Run rules

- The benchmark is run by executing scripts
  - no human involved at query time
  - aspects of CBIR applications such as the readability of the result or the user interface of the application are not the subject of this benchmark
- All files, including the image collection, the versioned ground truth files, and the scripts for each benchmark must be freely accessible on the World Wide Web
- An attribute of the personalized Web experience is near-instantaneous response, including the retrieval of images
  - any CBIR benchmark must provide metrics for response time performance

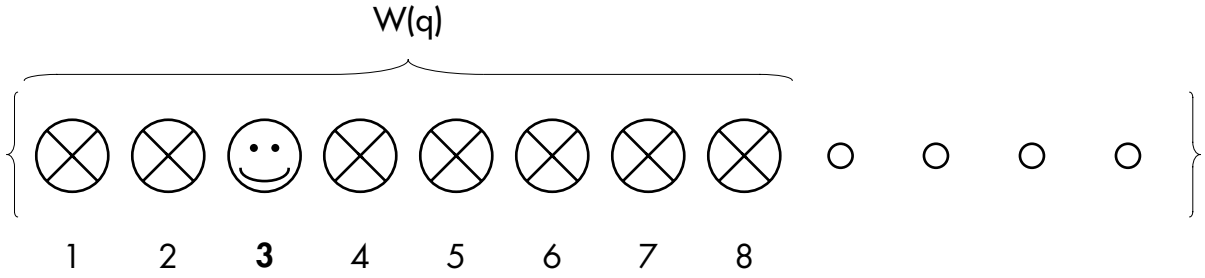
Benchmark metrics

- BIRDS-I uses a normalized rank similar to that proposed by Müller et al., but with the additional feature of penalizing missed images

- A single leading performance indicator should be provided with any system benchmark otherwise, a less reliable ad hoc metric simply will get invented for making first-order comparisons between CBIR benchmark results

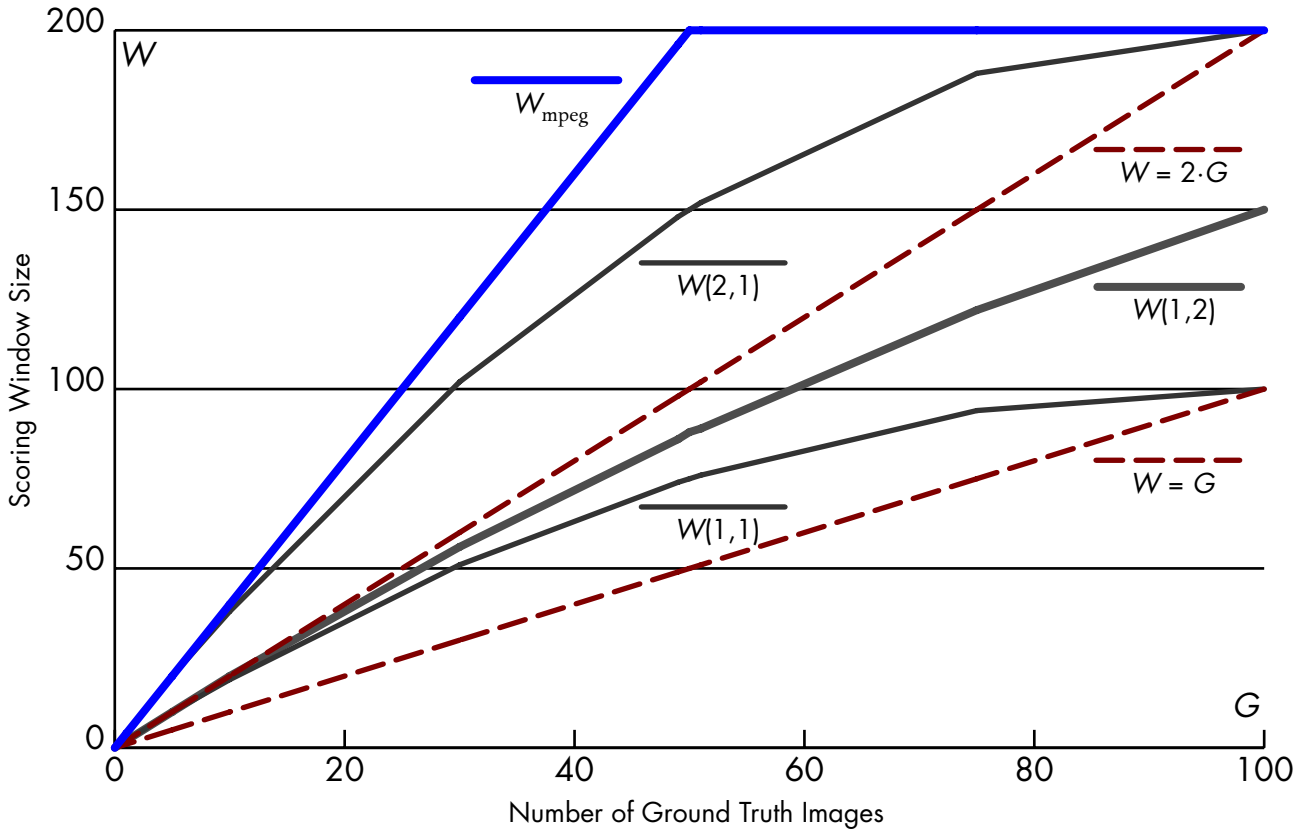
Window optimization problem

- A scoring window $W(q)$ is associated with the query q such that the returned images contained in $W(q)$ are ranked according to an index $r = 1, 2, \ldots, W$



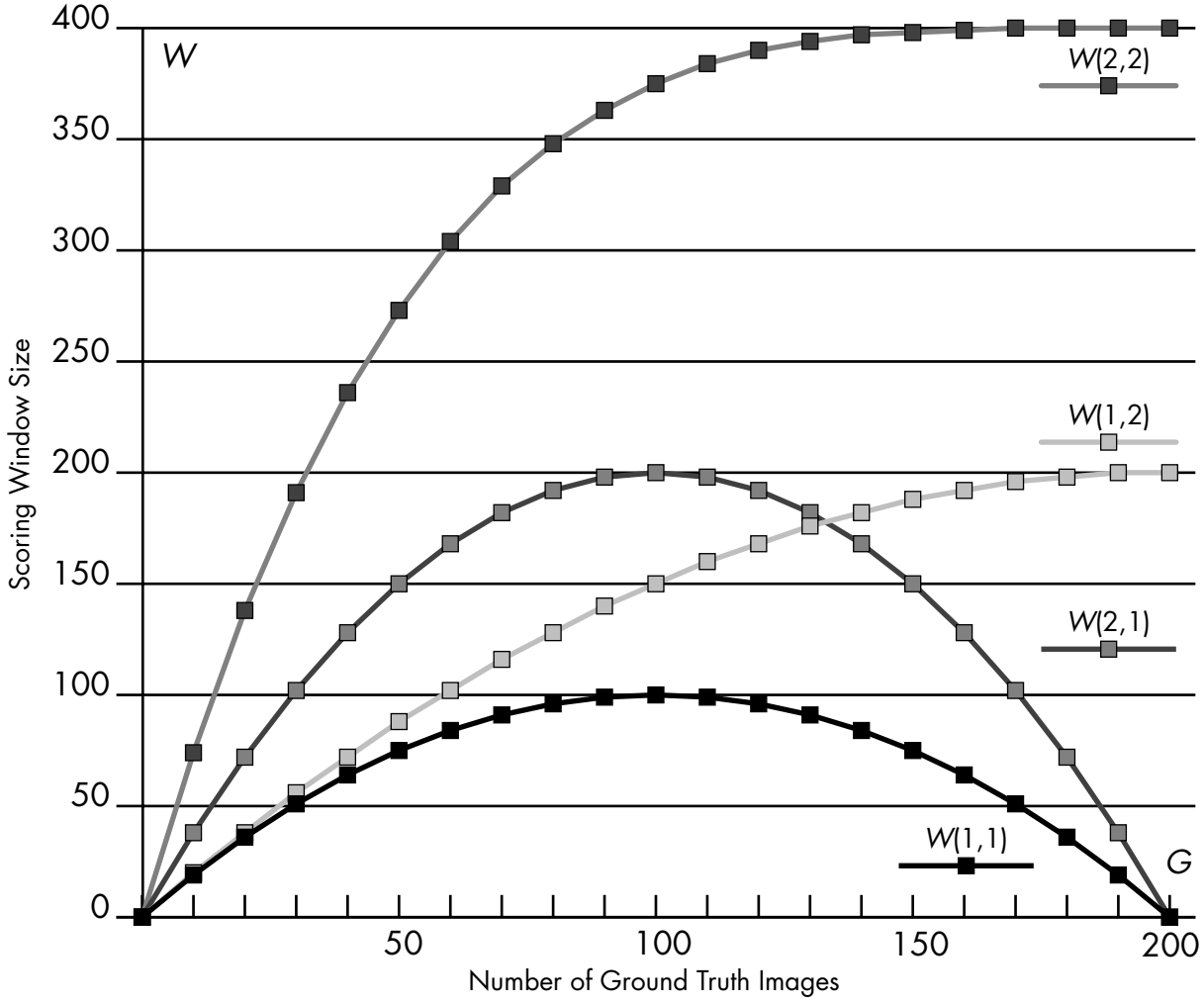
- Number of images returned is unknown
  - must select a search window of size $W(q)$ proportioned to each query's ground truth
  - determining the size for $W(q)$ is an optimization problem
  - should look for a (piecewise) continuous *convex* function from which to select the values for $W(q)$
  - family of inverted parabolic functions

Optimal retrieval window



the MPEG–7 function is labeled W_{mpeg}

Generalized class of convex scoring fns



Other CBIR benchmark activities

- MPEG-7 — B.S. Manjunath
- IEEE Multimedia — Clement Leung, Horace Ip
- Viper Team — Thierry Pun et al.
- Benchathlon

Summary

- BIRDS-I benchmark has been designed with the trend toward the use of small, personalized, wireless-networked systems clearly in mind
  - personalization with respect to the Web implies heterogeneous collections of images and this, in turn, dictates certain constraints on how these images can be organized and the type of retrieval accuracy measures that can be applied to CBIR performance
- BIRDS-I only requires controlled human intervention for compilation of the image collection, none for the ground truth generation & the assessment of retrieval accuracy
  - benchmark image collections need to be evolved incrementally toward millions of images and that scaleup can only be achieved through the use of computer-aided compilation
  - Two concepts we introduced into the scoring metric are a tightly optimized image-ranking window, and a penalty for missed images, each of which is important for the automated benchmarking of large-scale CBIR