Adaptive Cluster Distance Bounding for Similarity Search in Image Databases

Sharadh Ramaswamy and Kenneth Rose
Signal Compression Lab
{rsharadh,rose}@ece.ucsb.edu

Introduction

- Huge image databases are central in many apps. e.g. bio-imaging
- Images rep. by high-dim. features
- Content-based retrieval is inevitable
- Fast similarity search needed for quick navigation

http://www.cs.cmu.edu/~juny/Prof/images/CBIR.jpg

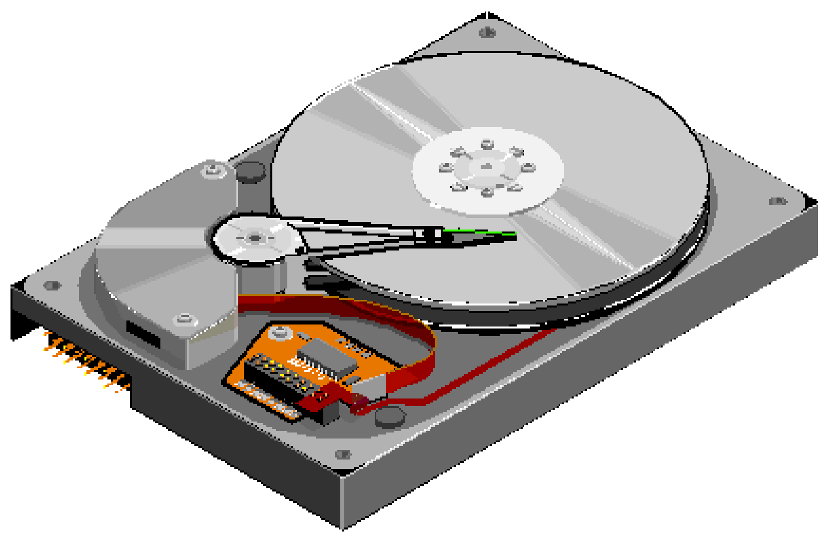
Image Features and Distances

- Popular image features
  - color histogram
  - texture descriptors
  - shape descriptors
- Common similarity measure – Euclidean distance (not perceptually optimal)
- Mahalanobis distances (thru' relevance feedback) possible

Storage on Hard-disks

- High-dim. features stored on a hard-drive
- Access thru' blocks/pages (fixed size)
- Sequential/serial or random access
- Random IOs more expensive per page
- Every access = 1 random IO + rest serial IOs

From Computer Desktop Encyclopedia
© 2005 The Computer Language Co. Inc.

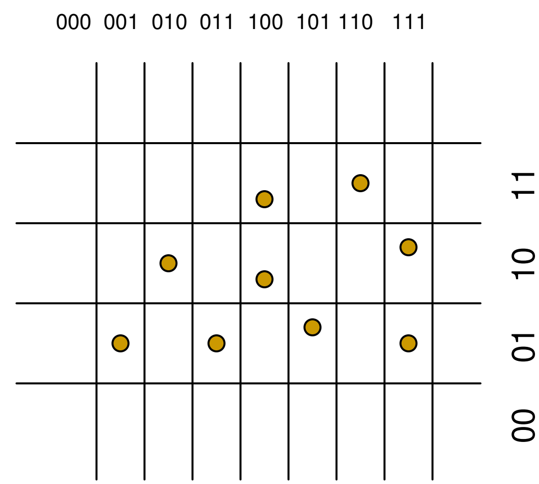


Multi-dimensional Indexes

- Tree-like indexes efficient in low dimensions e.g. R-tree
- ‘Curse of dimensionality’ hinders R-trees etc., creates large no. of random IOs
- Scan-like methods more effective at high dimensions
- Vector Approximation (VA)-File popular

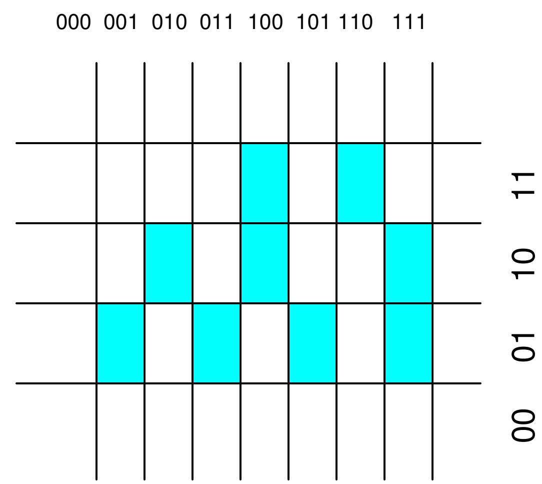
VA-File based Indexing

- Quantize each dimension uniformly
- Quantize each element of data-set



VA-File based Indexing

- Quantize each dimension uniformly
- Quantize each element of data-set
- Create approximation file
  - store quantization bit-strings for each element

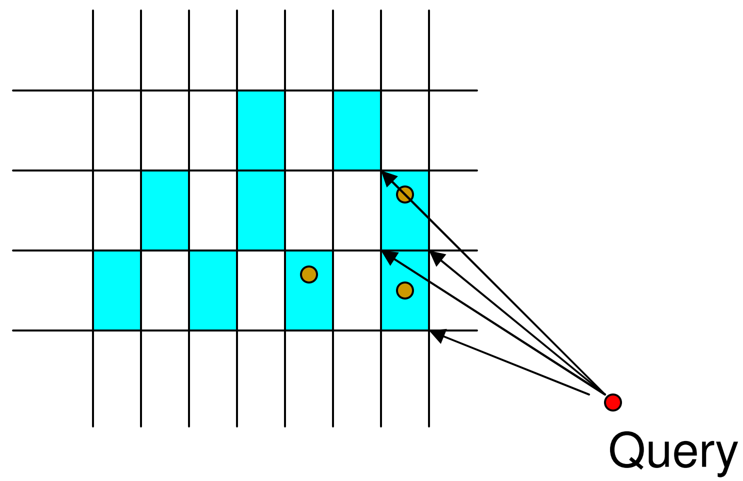


VA-File –Query processing

1. Read approximation file
2. Establish lower and upper distance bounds to occupied cells
3. Eliminate irrelevant cells
4. Access all survivors in order of lower bounds
5. If k^{th} lowest distance found so far, less than next lower bound,

   STOP (kNNs found)

   Else read next survivor.



VQ/Clustering for Indexing

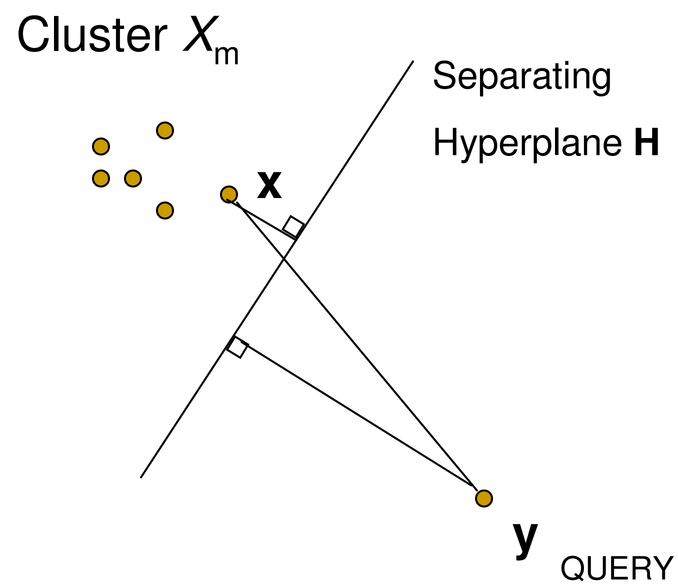
- VQ *is* optimal in compression
  - smaller preprocessing storage
- Similar feature vectors stored together
  - each cluster has several candidate vectors
  - better use of page access structure
- Extensively used in approx NN search
- Cluster-distance bounding for exact NN
  - bounds using MBRs and MBSs are loose

Bounding Query-Cluster Distance

- $d(y, X_m) = \min d(x,y)$
- $d(x,y) \geq d(y,H) + d(x,H)$

$\Rightarrow \min d(x,y) \geq d(y,H) + \min d(x,H)$

$\Rightarrow d(y, X) \geq d(y,H) + d(X_m,H)$

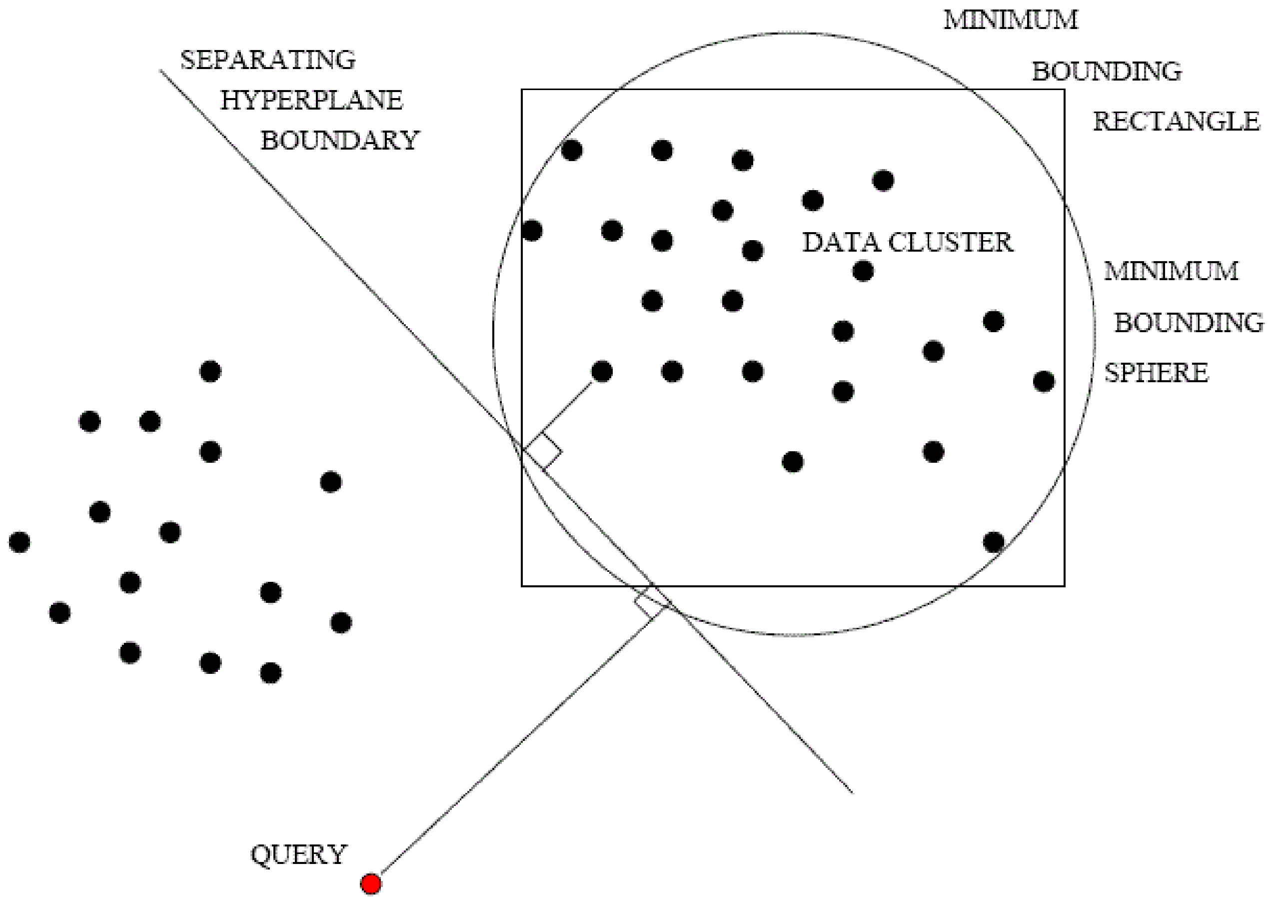


Distance-to-cluster ≥ Query-Hyperplane distance + Cluster-Hyperplane Distance

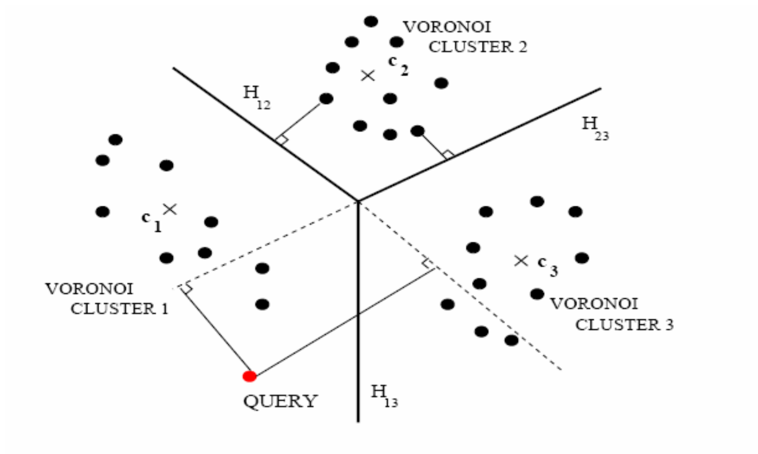
("Support")

Support evaluated offline and stored

Cluster Distance Bounding



Adaptive Cluster Distance Bounding



- Bound distance with multiple hyperplanes
  - use tightest distance bound
- Cluster boundaries are linear
  - use them as (separating) hyperplanes
  - no need to store hyperplanes

Proposed Indexing Scheme

1. Cluster data-set through VQ/K-means
   - "nearest neighbor" partitioning for linear boundaries
   - evaluate "offline" and store cluster supports
2. Bound query-cluster distance with hyperplane bound
3. Retrieve clusters in order of distance
   - IF kNN distance so far < distance to next cluster

     STOP (kNNs found)

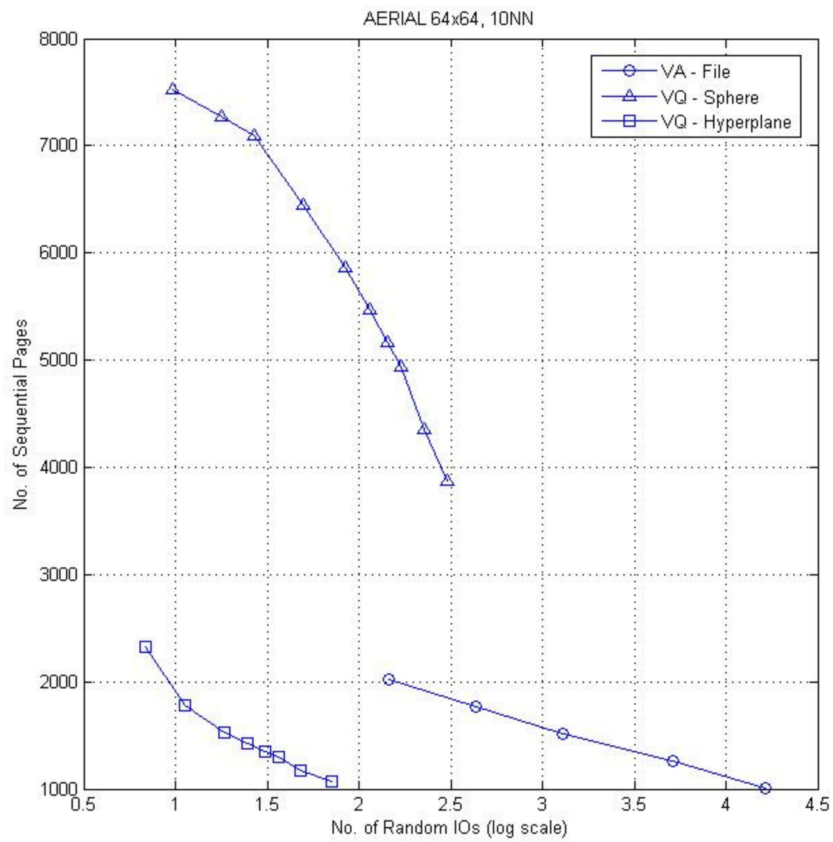
     ELSE read next cluster (till all clusters read)

Experiments & Results

- Data-sets - AERIAL (60 dim, 275K)
- - BIO-RETINA (62 dim, 208K)
- Clustering with GLA/K-means
- No. of clusters varied (20 – 600)
- VA-File Quantization varied (3-8 bits/dim)
- Page size – 8kB
- 2D Performance Metric –

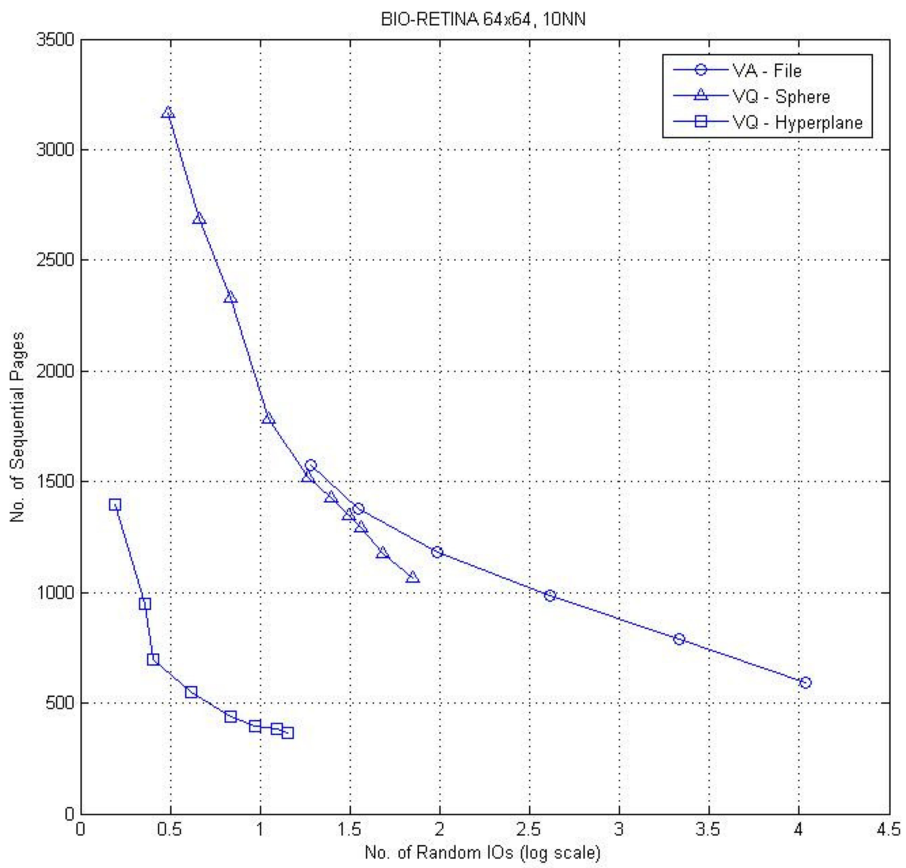
  (Random IOs ,Serial IOs)

IO Performance - AERIAL



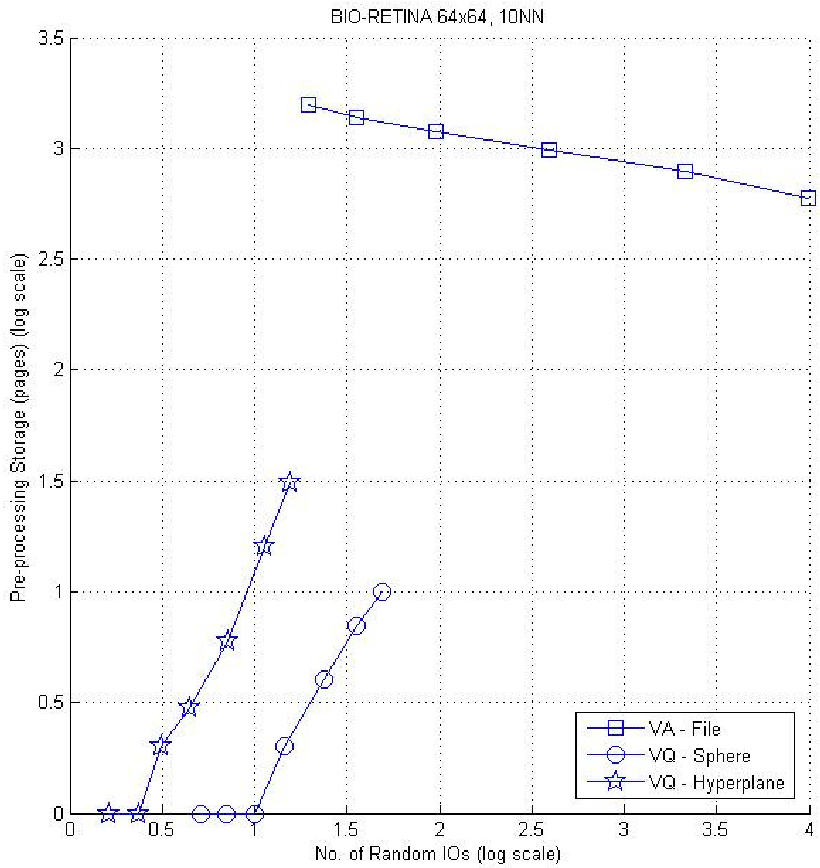
Random IOs reduced by 100X!!

IO Performance – BIO-RETINA



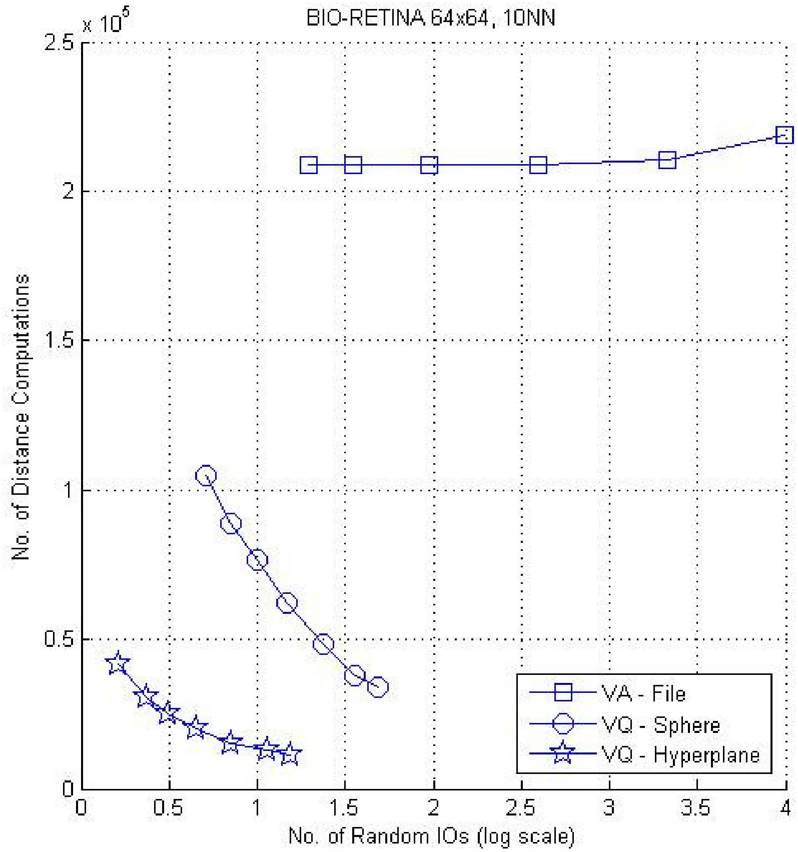
Random IOs reduced by 100X!!

Pre-processing Storage (BIO-RETINA)



Lower pre-proc. Storage!!

Computational Costs – BIO-RETINA



Lower computational costs!

Results & Future Work

- Real data-sets exhibit significant dependencies across dimensions
- VQ/Clustering exploits correlations
- Proposed hyperplane bound is tight and provides for efficient spatial filtering
- Huge gains in IO complexity possible over VA-File and MBS bounds
- To be extended towards Mahalanobis distances and relevance feedback…