Multimedia Data Mining

Data Mining

- Data Mining definition:
  - A class of database applications that look for hidden patterns in a group of data.
  - Finding rules of the game knowing the moves of the game
  - Unifying framework for data representation and problem solving in order to learn and discover from large amounts of different types of data.

Multimedia Data Mining

- Multimedia data types
  - any type of information medium that can be represented, processed, stored and transmitted over network in digital form
  - Multi-lingual text, numeric, images, video, audio, graphical, temporal, relational, and categorical data.
  - Relation with conventional data mining term

Definitions

- Subfield of data mining that deals with an extraction of implicit knowledge, multimedia data relationships, or other patterns not explicitly stored in multimedia databases
  - Influence on related interdisciplinary fields
  - Databases – extension of the KDD (rule patterns)
  - Information systems – multimedia information analysis and retrieval – content-based image and video search and efficient storage organization

Information model

- Data segmentation
  - Multimedia data are divided into logical interconnected segments (objects)
  - Pattern extraction
  - Mining and analysis procedures should reveal some relations between objects on the different level
  - Knowledge representation
  - Incorporated linked patterns

Generalizing Spatial and Multimedia Data

- Spatial data:
  - Generalize detailed geographic points into clustered regions, such as business, residential, industrial, or agricultural areas, according to land usage
  - Require the merge of a set of geographic areas by spatial operations
- Image data:
  - Extracted by aggregation and/or approximation
  - Size, color, shape, texture, orientation, and relative positions and structures of the contained objects or regions in the image
- Music data:
  - Summarize its melody: based on the approximate patterns that repeatedly occur in the segment
  - Summarized its style: based on its tone, tempo, or the major musical instruments played

What Is a Spatial Database System?

- Geometric, geographic or spatial data: space-related data
  - Example: Geographic space (2-D abstraction of earth surface), VLSI design, model of human brain, 3-D space representing the arrangement of chains of protein molecule.
- Spatial database system vs. image database systems.
  - Image database system: handling digital raster image (e.g., satellite sensing, computer tomography), may also contain techniques for object analysis and extraction from images and some spatial database functionality.
  - Spatial (geometric, geographic) database system: handling objects in space that have identity and well-defined extents, locations, and relationships.

Modeling Spatial Objects

- What needs to be represented?
- Two important alternative views
  - Single objects: distinct entities arranged in space each of which has its own geometric description
    - modeling cities, forests, rivers
  - Spatially related collection of objects: describe space itself (about every point in space)
    - modeling land use, partition of a country into districts

Modeling Single Objects: Point, Line and Region

- Point: location only but not extent
- Line (or a curve usually represented by a polyline, a sequence of line segment):
  - moving through space, or connections in space (roads, rivers, cables, etc.)
- Region:
  - Something having extent in 2D-space (country, lake, park). It may have a hole or consist of several disjoint pieces.

Spatial Association Analysis

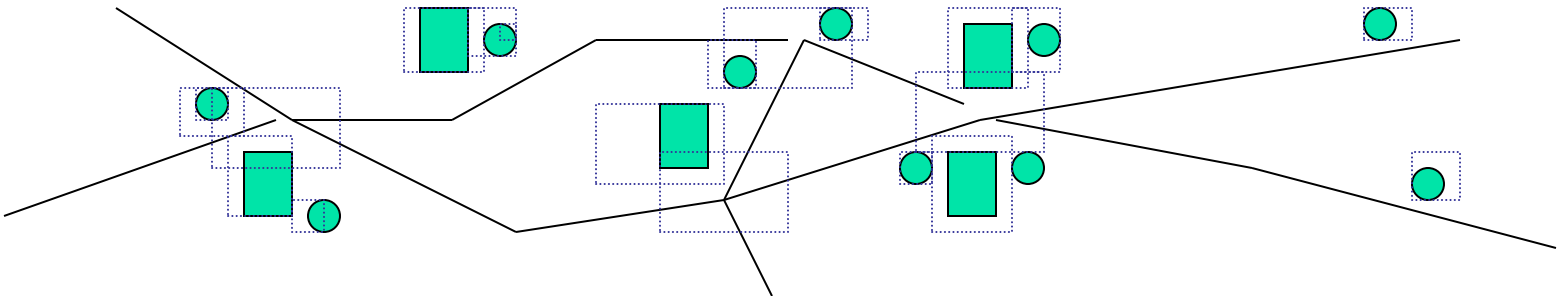
- Spatial association rule: $A \Rightarrow B$ [*s%, c%*]
  - A and B are sets of spatial or non-spatial predicates
    - Topological relations: *intersects, overlaps, disjoint,* etc.
    - Spatial orientations: *left_of, west_of, under,* etc.
    - Distance information: *close_to, within_distance,* etc.
  - *s%* is the support and *c%* is the confidence of the rule
- Examples

1) *is_a(x, large_town) ^ intersect(x, highway) → adjacent_to(x, water)*
[7%, 85%]

2) What kinds of objects are typically located close to golf courses?

Progressive Refinement Mining of Spatial Association Rules

- Hierarchy of spatial relationship:
  - *g_close_to*: *near_by*, *touch*, *intersect*, *contain*, etc.
  - First search for rough relationship and then refine it
- Two-step mining of spatial association:
  - Step 1: Rough spatial computation (as a filter)
    - Using MBR or R-tree for rough estimation
  - Step2: Detailed spatial algorithm (as refinement)
    - Apply only to those objects which have passed the rough spatial association test (no less than *min_support*)



Mining Spatial Co-location Rules

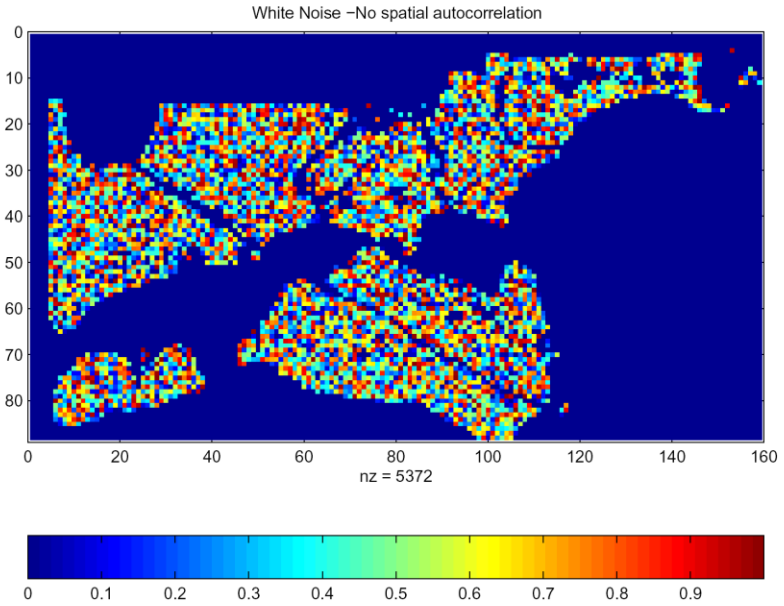
- Co-location rule is similar to association rule but explore more relying spatial auto-correlation
- It leads to efficient processing
- It can be integrated with progressive refinement to further improve its performance
- Spatial co-location mining idea can be applied to clustering, classification, outlier analysis and other potential mining tasks

Spatial Autocorrelation

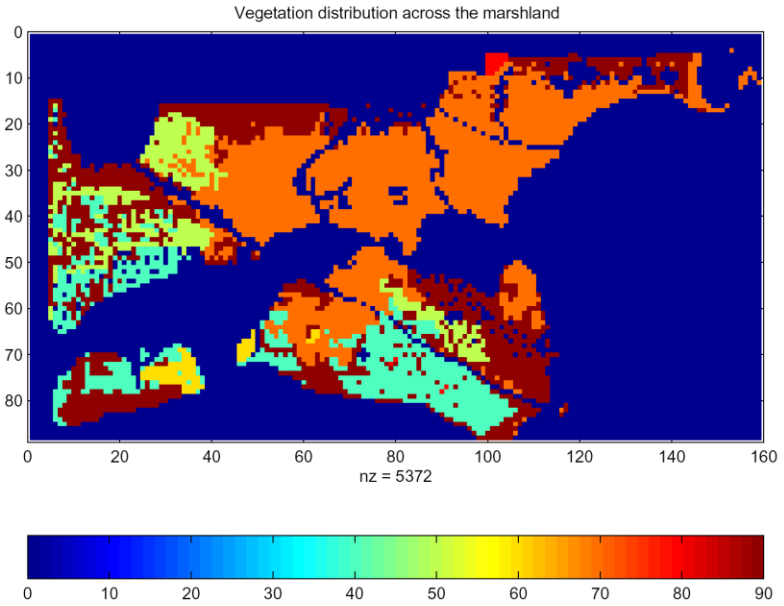
- Spatial data tends to be highly self-correlated
  - Example: Neighborhood, Temperature
  - Items in a traditional data are independent of each other, whereas properties of locations in a map are often "**auto-correlated**".
- First law of geography:

"*Everything is related to everything, but nearby things are more related than distant things*."

Spatial Autocorrelation (cont'd)



(a) Pixel property with independent identical distribution

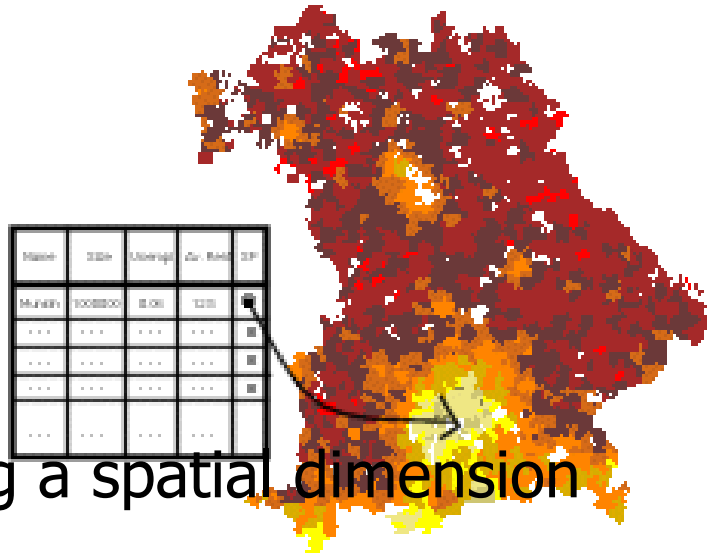


(b) Vegetation Durability with SA

Spatial Classification

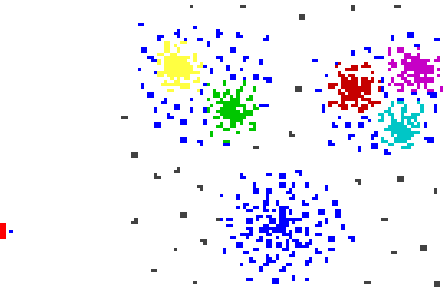
- Methods in classification
  - Decision-tree classification, Naïve-Bayesian classifier + boosting, neural network, logistic regression, etc.
  - Association-based multi-dimensional classification - Example: classifying house value based on proximity to lakes, highways, mountains, etc.
- Assuming learning samples are independent of each other
  - Spatial auto-correlation violates this assumption!
- Popular spatial classification methods
  - Spatial auto-regression (SAR)
  - Markov random field (MRF)

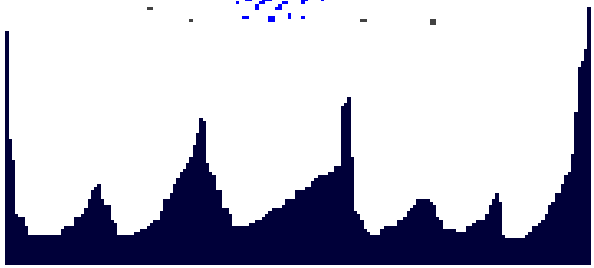
Spatial Trend Analysis



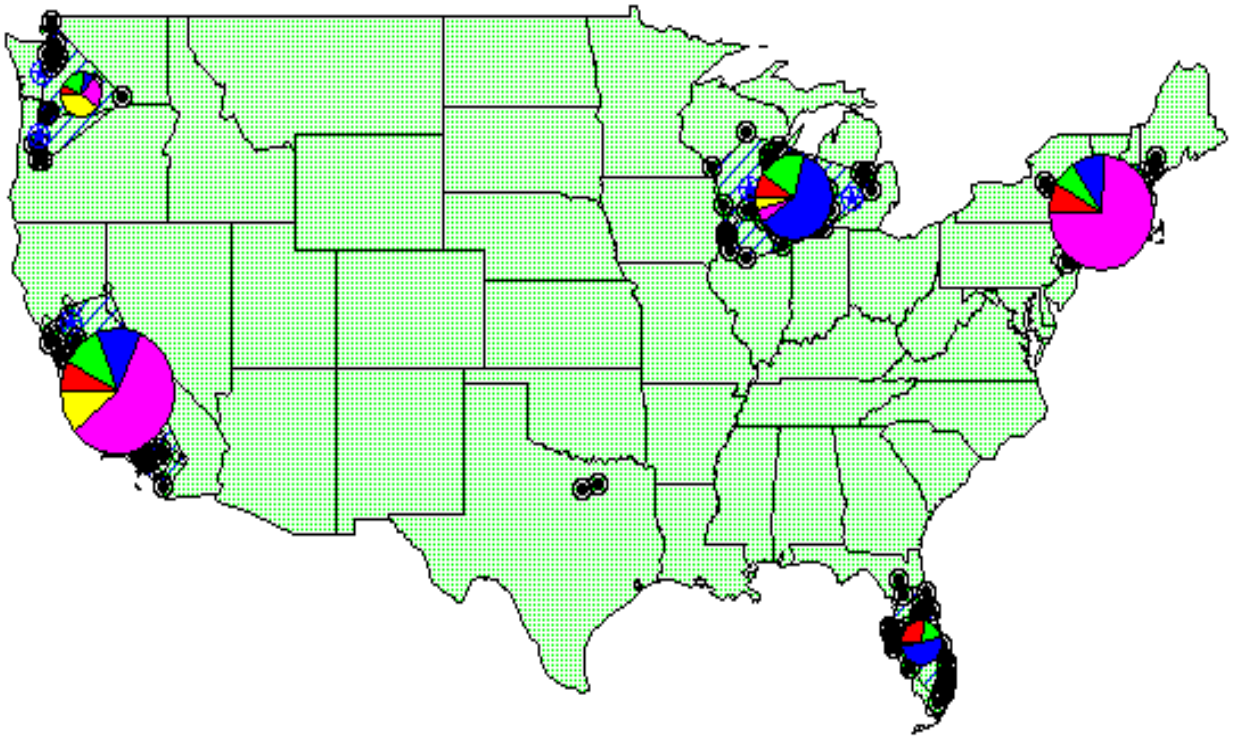
- Function
  - Detect changes and trends along a spatial dimension
  - Study the trend of non-spatial or spatial data changing with space
- Application examples
  - Observe the trend of changes of the climate or vegetation with increasing distance from an ocean
  - Crime rate or unemployment rate change with regard to city geo-distribution

Spatial Cluster Analysis





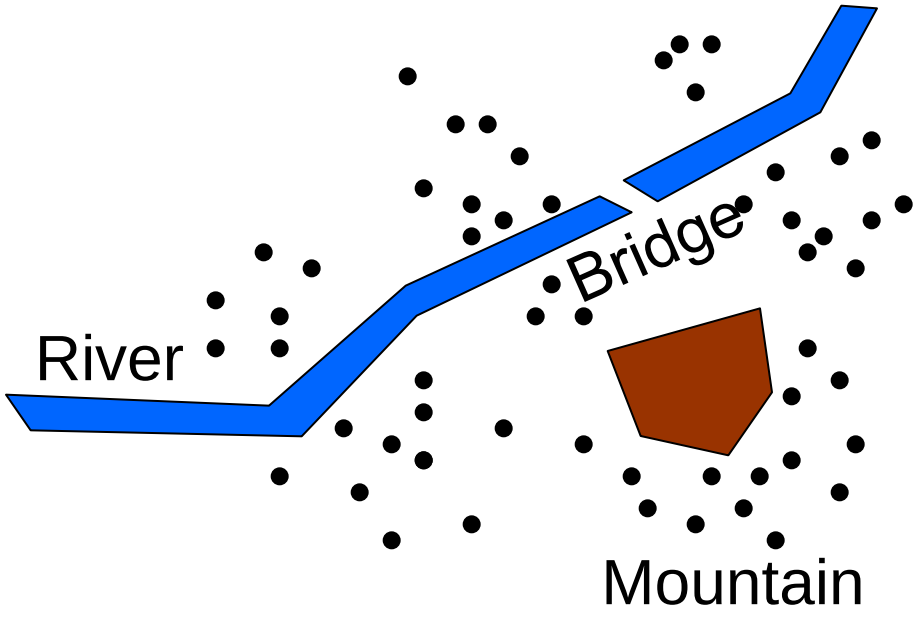
- Mining clusters—k-means, k-medoids, hierarchical, density-based, etc.
- Analysis of distinct features of the clusters



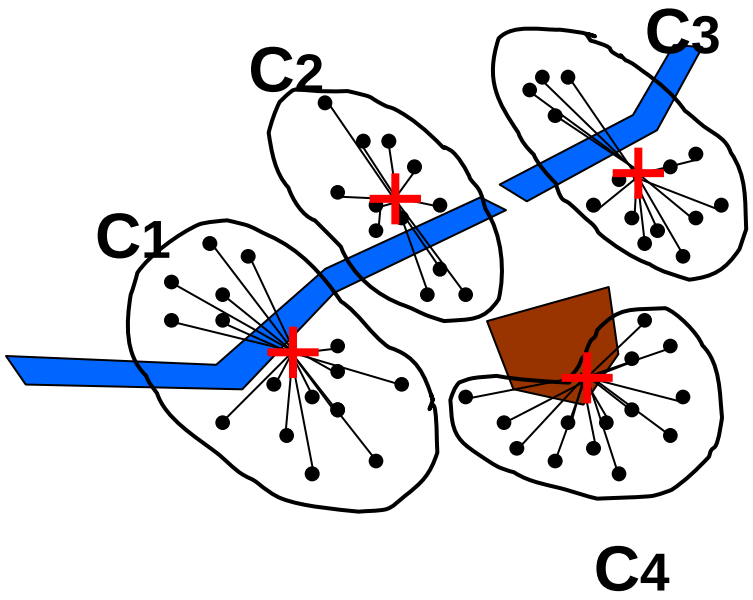
Constraints-Based Clustering

- Constraints on individual objects
  - Simple selection of relevant objects before clustering
- Clustering parameters as constraints
  - K-means, density-based: radius, min-# of points
- Constraints specified on clusters using SQL aggregates
  - Sum of the profits in each cluster > $1 million
- Constraints imposed by physical obstacles
  - Clustering with obstructed distance

Constrained Clustering: Planning ATM Locations



Spatial data with obstacles



Clustering *without* taking obstacles into consideration

Mining Spatiotemporal Data

- Spatiotemporal data
  - Data has spatial extensions and changes with time
  - Ex: Forest fire, moving objects, hurricane & earthquakes
- Automatic anomaly detection in massive moving objects
  - Moving objects are ubiquitous: GPS, radar, etc.
  - Ex: Maritime vessel surveillance
  - Problem: Automatic anomaly detection

Analysis: Mining Anomaly in Moving Objects

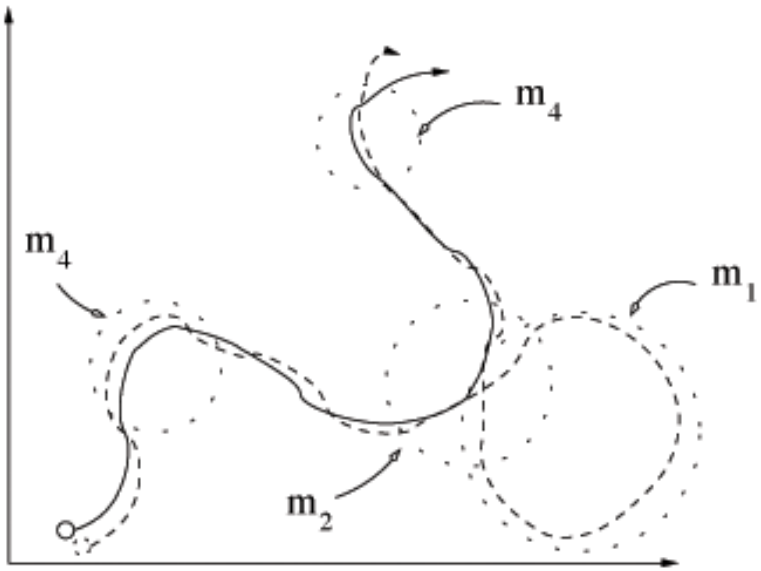
- Raw analysis of collected data does not fully convey "anomaly" information
- More effective analysis relies on higher semantic features
- Examples:
  - A speed boat moving quickly in open water
  - A fishing boat moving slowly into the docks
  - A yacht circling slowly around landmark during night hours

Framework: Motif-Based Feature Analysis

- Motif-based representation
  - A motif is a prototypical movement pattern
  - View a movement path as a sequence of motif expressions
- Motif-oriented feature space
  - Automated motif feature extraction
  - Semantic-level features
- Classification
  - Anomaly detection via classification
  - High dimensional classifier

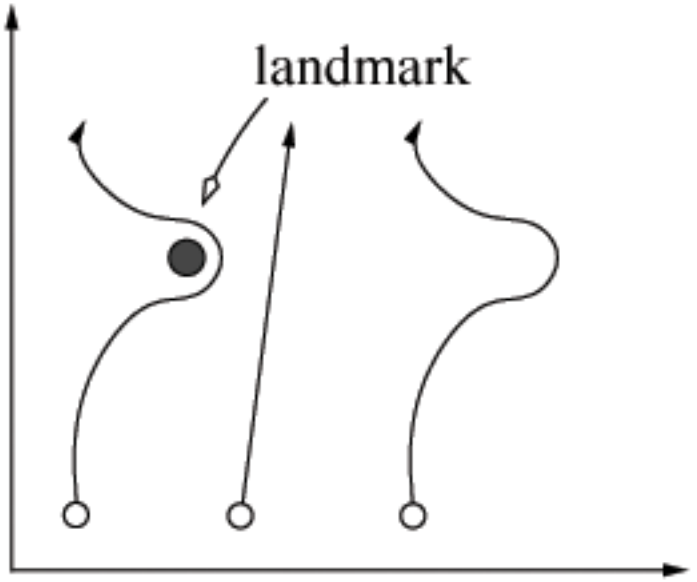
Movement Motifs

- Prototypical movement of object
  - *Right-turn, U-turn*
- Can be either defined by an expert or discovered automatically from data
  - Defined in our framework
- Extracted in movement paths
- Path becomes a set of motif expressions



Motif Expression Attributes

- Each motif expression has attributes (e.g., speed, location, size)
- Attributes express *how* a motif was expressed
- Conveys semantic information useful for classification
  - *a tight circle at 30mph near landmark Y.*
  - *A tight circle at 10mph in location X*



Motif-Oriented Feature Space

- Attributes describe *how* motifs are expressed
- Let there be A attributes, each path is a set of (A+1)-tuples

  $\{(m_i, v_1, v_2, \ldots, v_A), (m_j, v_1, v_2, \ldots, v_A)\}$
- Naïve Feature space construction
  1. Let each distinct $(m_j, v_1, v_2, \ldots, v_A)$ be a feature
  2. If path exhibits a particular motif-expression, its value is 1. Otherwise, its value is 0.

Similarity Search in Multimedia Data

- Description-based retrieval systems
  - Build indices and perform object retrieval based on image descriptions, such as keywords, captions, size, and time of creation
  - Labor-intensive if performed manually
  - Results are typically of poor quality if automated
- Content-based retrieval systems
  - Support retrieval based on the image content, such as color histogram, texture, shape, objects, and wavelet transforms

Queries in Content-Based Retrieval Systems

- Image sample-based queries
  - Find all of the images that are similar to the given image sample
  - Compare the feature vector (signature) extracted from the sample with the feature vectors of images that have already been extracted and indexed in the image database
- Image feature specification queries
  - Specify or sketch image features like color, texture, or shape, which are translated into a feature vector
  - Match the feature vector with the feature vectors of the images in the database

Approaches Based on Image Signature

- Color histogram-based signature
  - The signature includes color histograms based on color composition of an image regardless of its scale or orientation
  - No information about shape, location, or texture
  - Two images with similar color composition may contain very different shapes or textures, and thus could be completely unrelated in semantics
- Multifeature composed signature
  - Define different distance functions for color, shape, location, and texture, and subsequently combine them to derive the overall result

Wavelet Analysis

- Wavelet-based signature
  - Use the dominant wavelet coefficients of an image as its signature
  - Wavelets capture shape, texture, and location information in a single unified framework
  - Improved efficiency and reduced the need for providing multiple search primitives
  - May fail to identify images containing similar objects that are in different locations.

One Signature for the Entire Image?

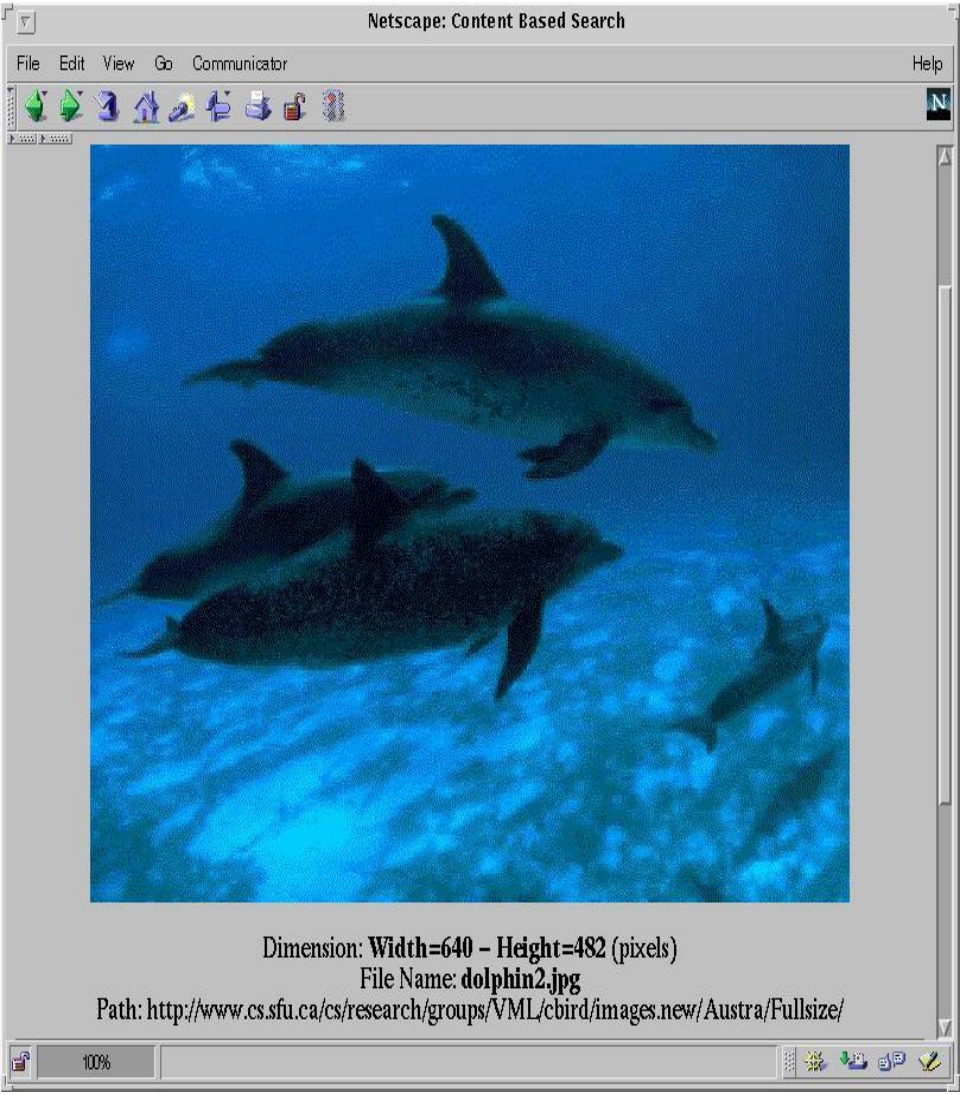
- Walnus: [NRS99] by Natsev, Rastogi, and Shim
- Similar images may contain similar regions, but a region in one image could be a translation or scaling of a matching region in the other

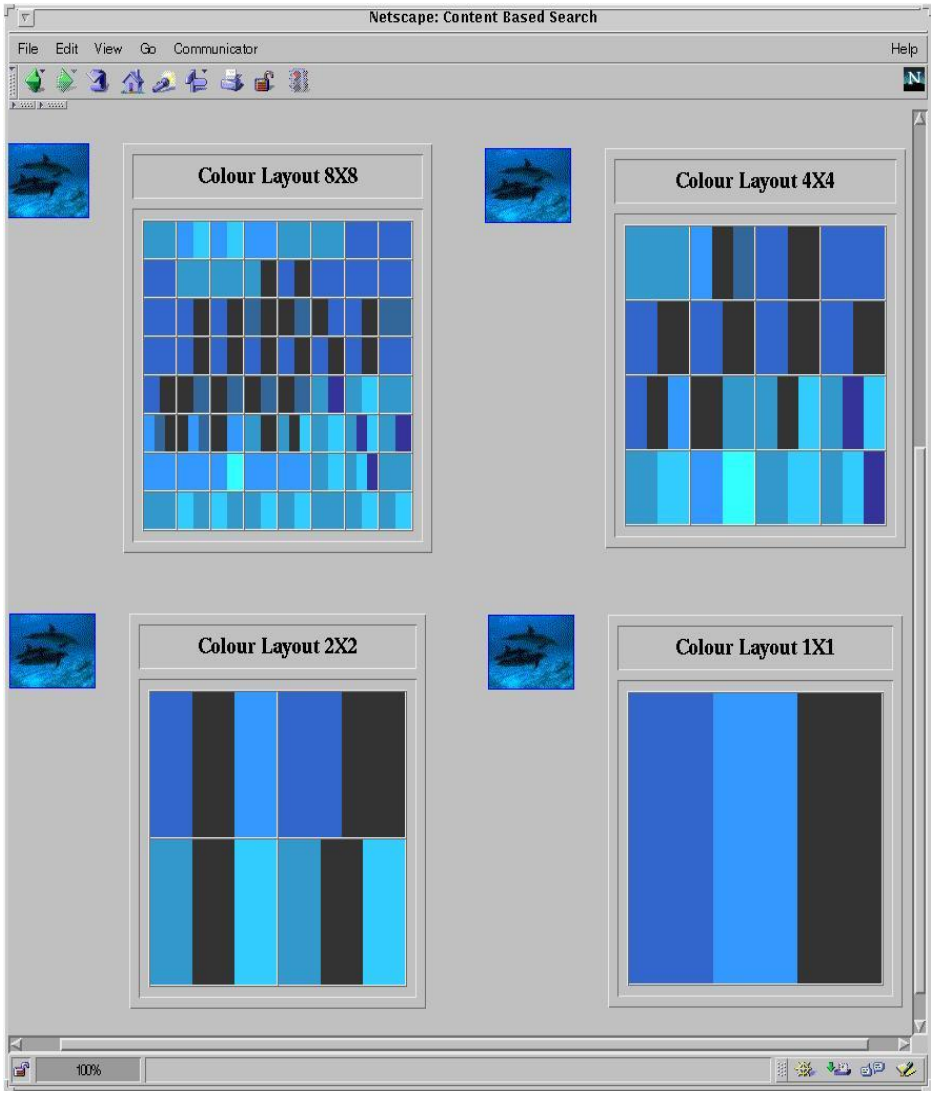
- Wavelet-based signature with region-based granularity
  - Define regions by clustering signatures of windows of varying sizes within the image
  - Signature of a region is the centroid of the cluster
  - Similarity is defined in terms of the fraction of the area of the two images covered by matching pairs of regions from two images

Multidimensional Analysis of Multimedia Data

- Multimedia data cube
  - Design and construction similar to that of traditional data cubes from relational data
  - Contain additional dimensions and measures for multimedia information, such as color, texture, and shape
- The database does not store images but their descriptors
  - Feature descriptor: a set of vectors for each visual characteristic
    - Color vector: contains the color histogram
    - MFC (Most Frequent Color) vector: five color centroids
    - MFO (Most Frequent Orientation) vector: five edge orientation centroids
  - Layout descriptor: contains a color layout vector and an edge layout vector

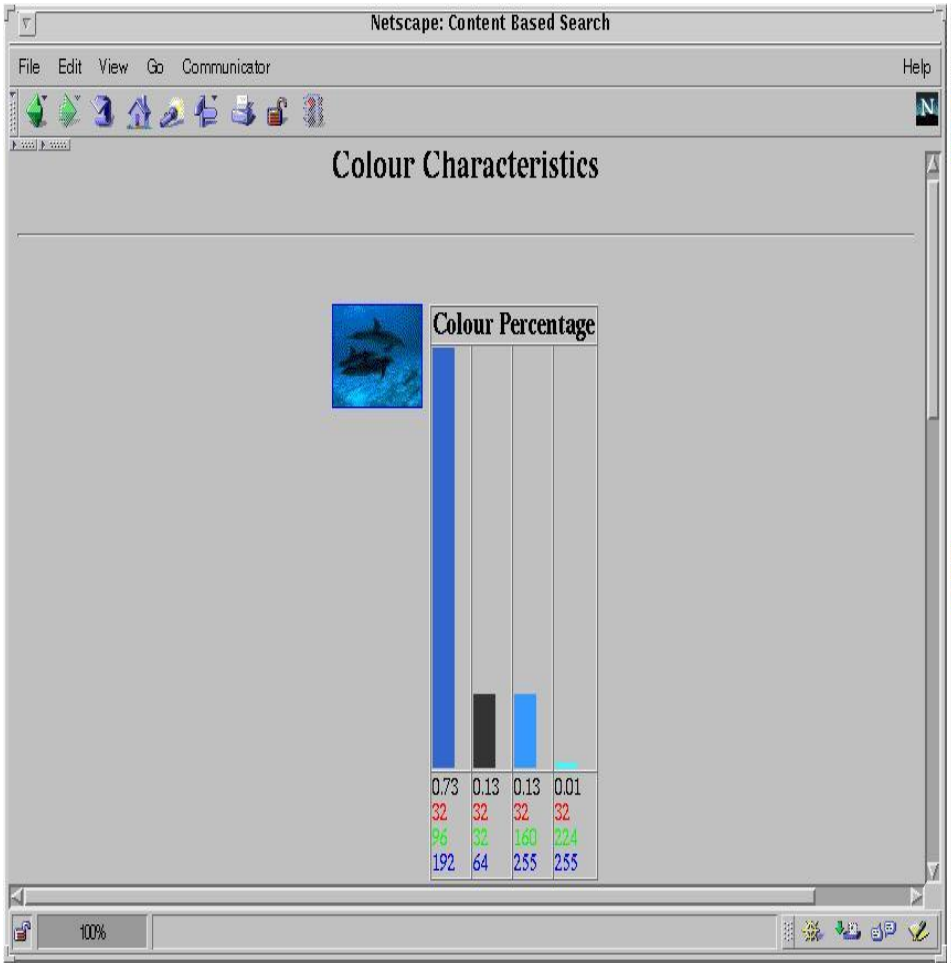
Multi-Dimensional Search in Multimedia Databases



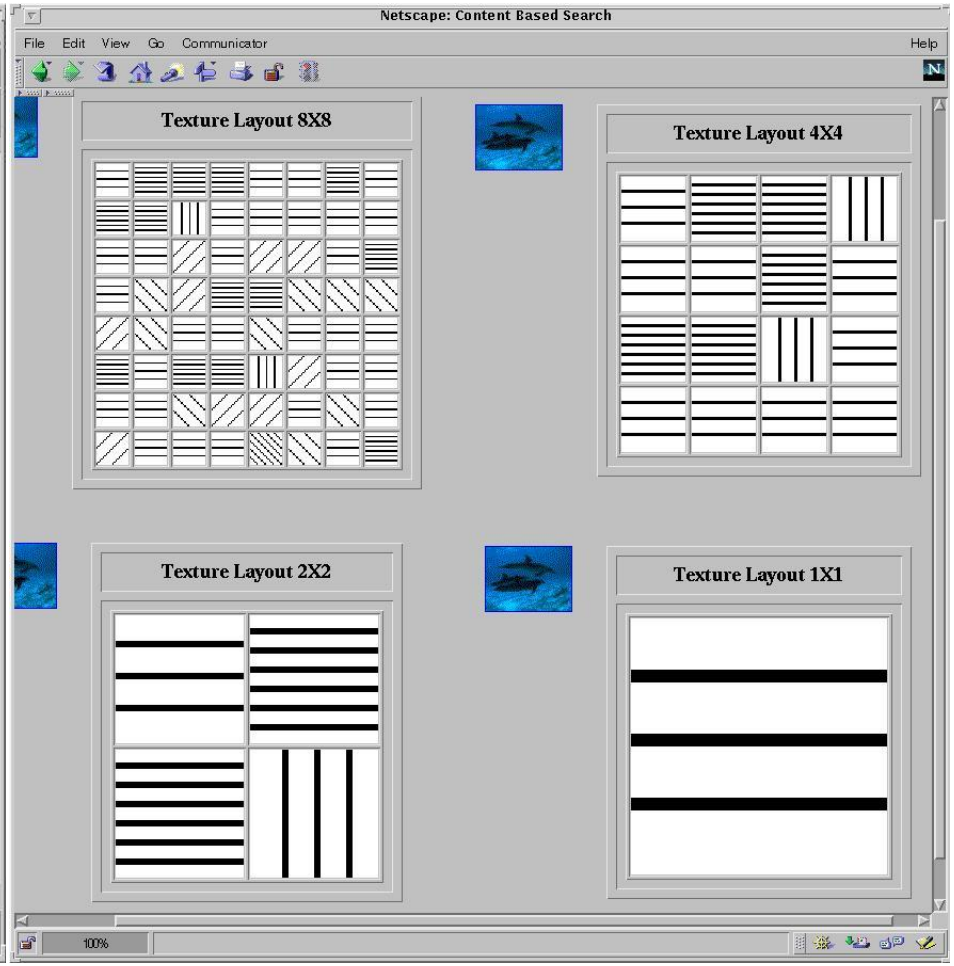


Multi-Dimensional Analysis in Multimedia Databases

Color histogram



Texture layout



Mining Multimedia Databases

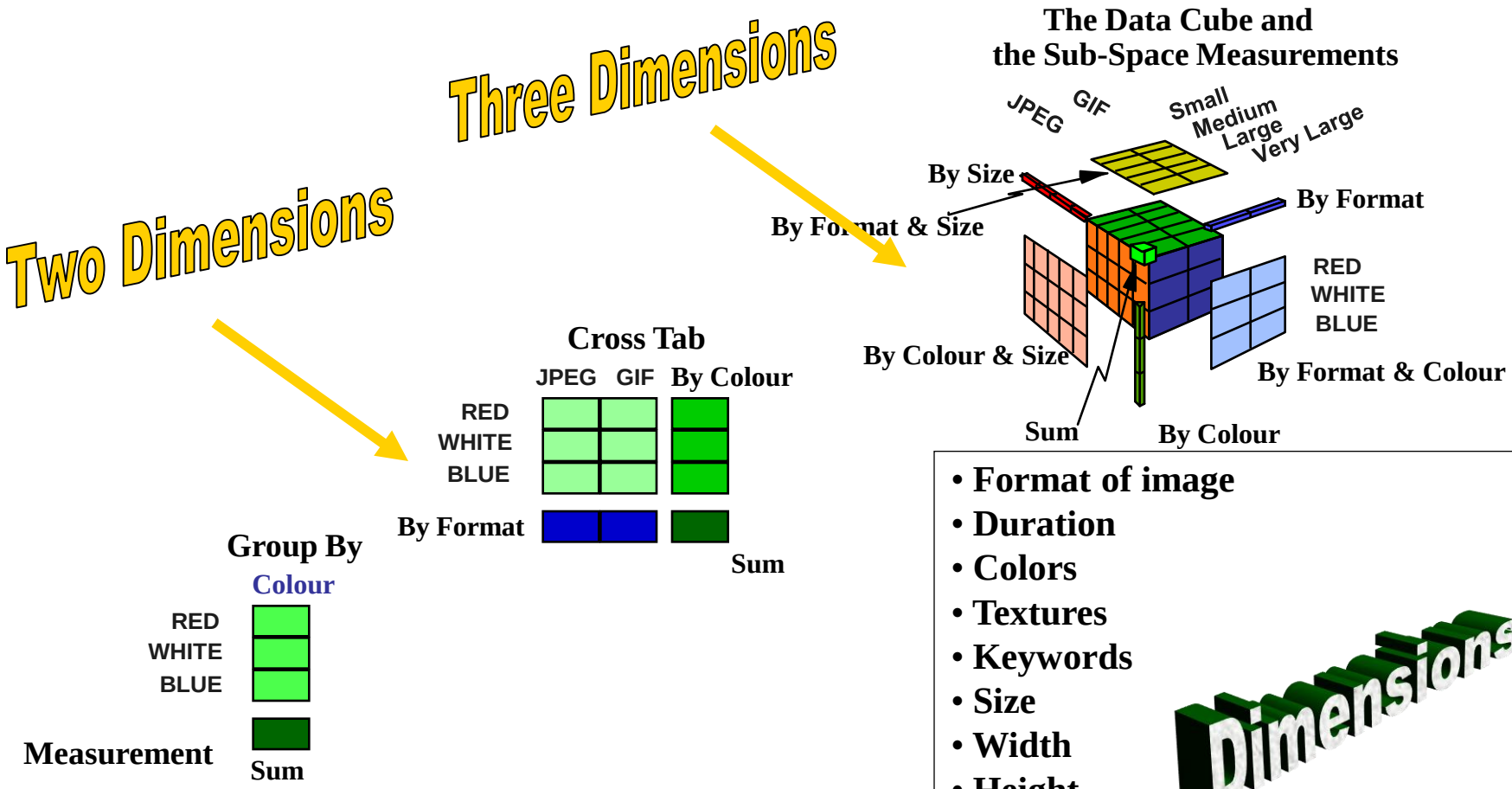
Refining or combining searches

Search for “blue sky”
(top layout grid is blue)

Search for “airplane in blue sky”
(top layout grid is blue and keyword = “airplane”)

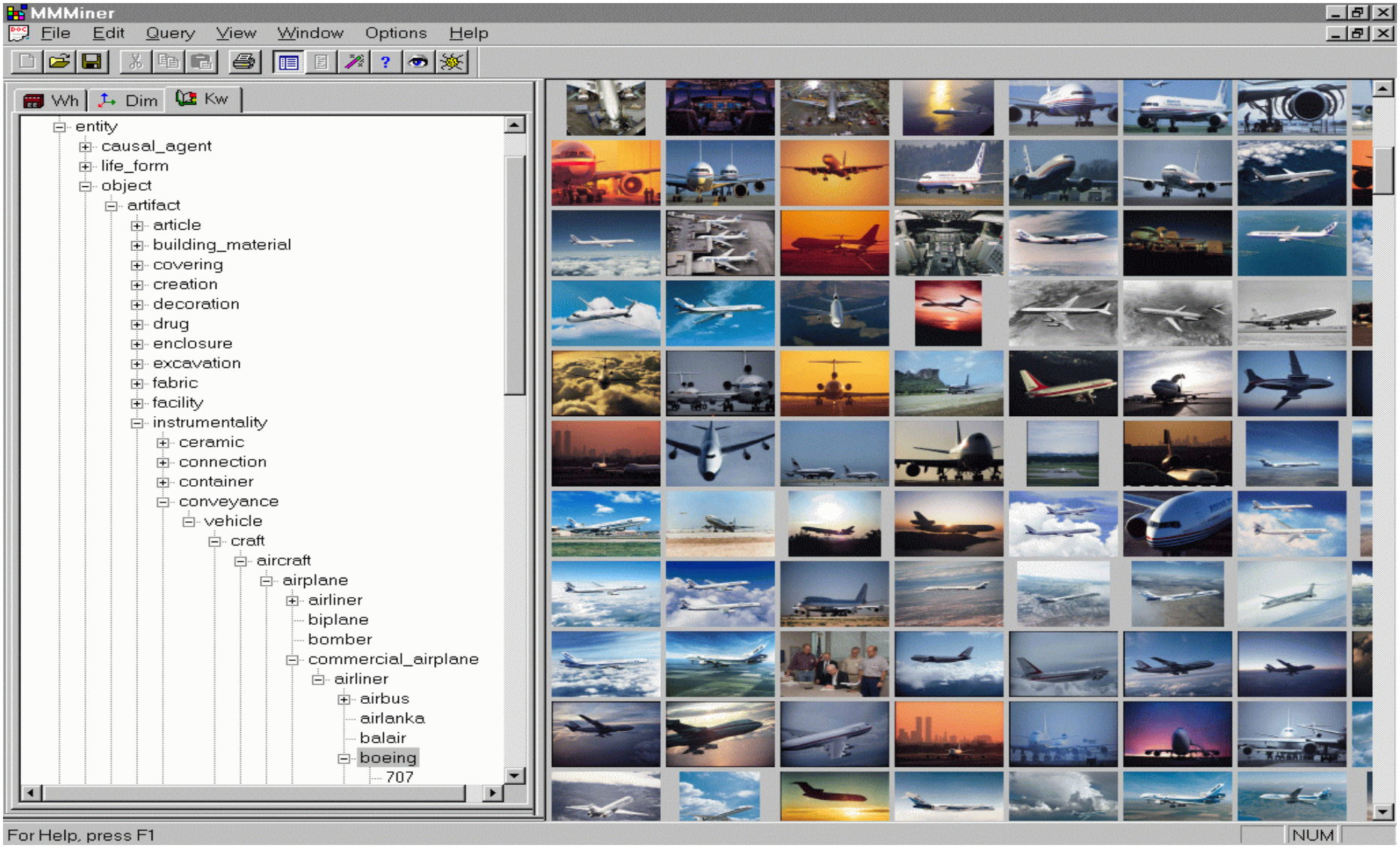
Search for “blue sky and green meadows”
(top layout grid is blue and bottom is green)

Mining Multimedia Databases

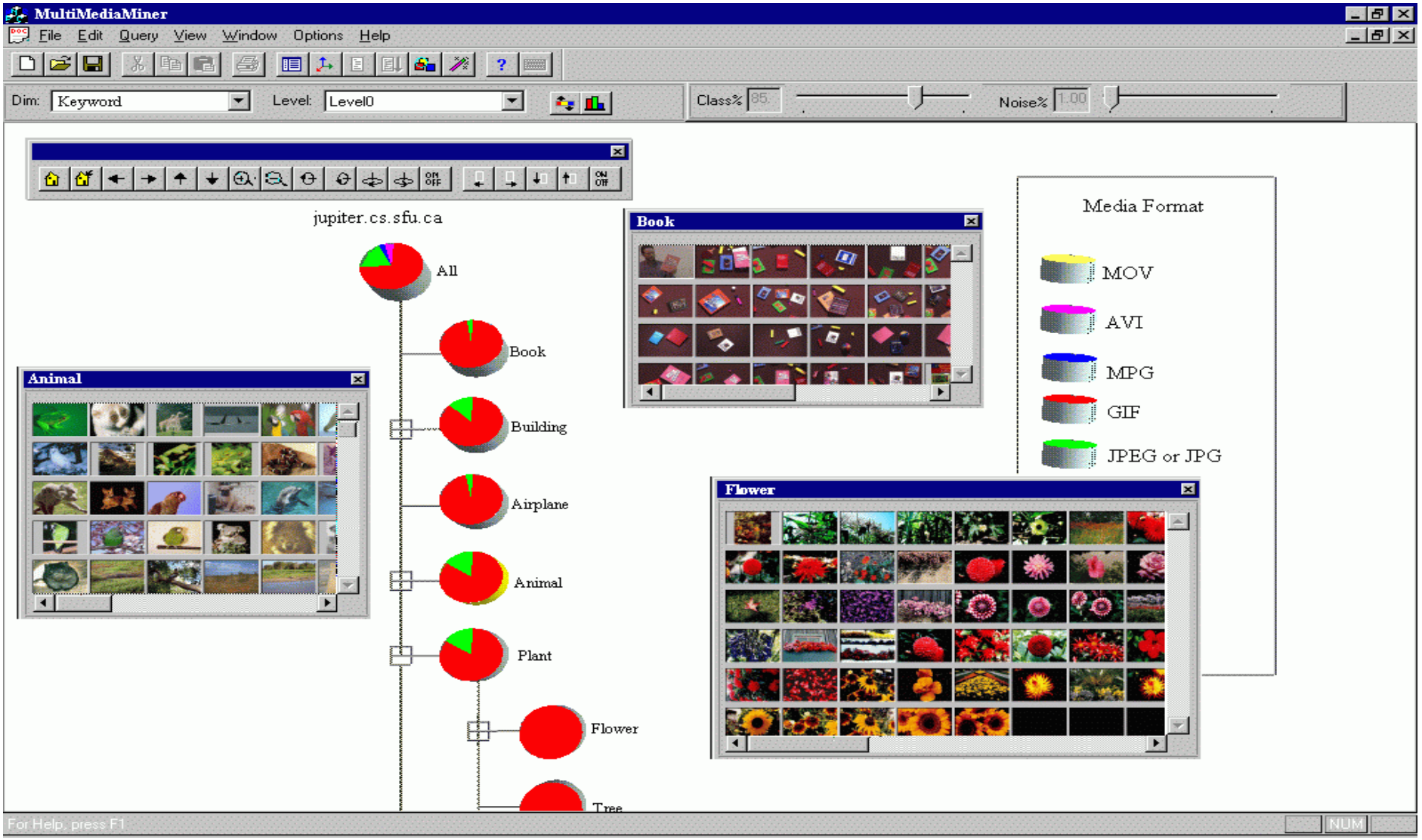


- Format of image
- Duration
- Colors
- Textures
- Keywords
- Size
- Width
- Height
- Internet domain of image
- Internet domain of parent pages
- Image popularity

Mining Multimedia Databases in MultiMediaMiner



Classification in MultiMediaMiner



Mining Associations in Multimedia Data

- Associations between image content and non-image content features
  - "If at least 50% of the upper part of the picture is blue, then it is likely to represent sky."
- Associations among image contents that are not related to spatial relationships
  - "If a picture contains two blue squares, then it is likely to contain one red circle as well."
- Associations among image contents related to spatial relationships
  - "If a red triangle is between two yellow squares, then it is likely a big oval-shaped object is underneath."

Mining Associations in Multimedia Data

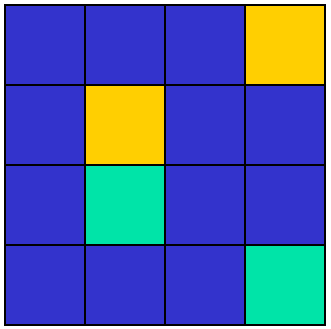
- Special features:
  - Need # of occurrences besides Boolean existence, e.g.,
    - "Two red square and one blue circle" implies theme "air-show"
  - Need spatial relationships
    - Blue on top of white squared object is associated with brown bottom
  - Need multi-resolution and progressive refinement mining
    - It is expensive to explore detailed associations among objects at high resolution
    - It is crucial to ensure the completeness of search at multi-resolution space

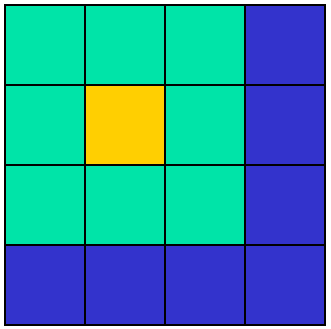
Mining Multimedia Databases

Spatial Relationships from Layout

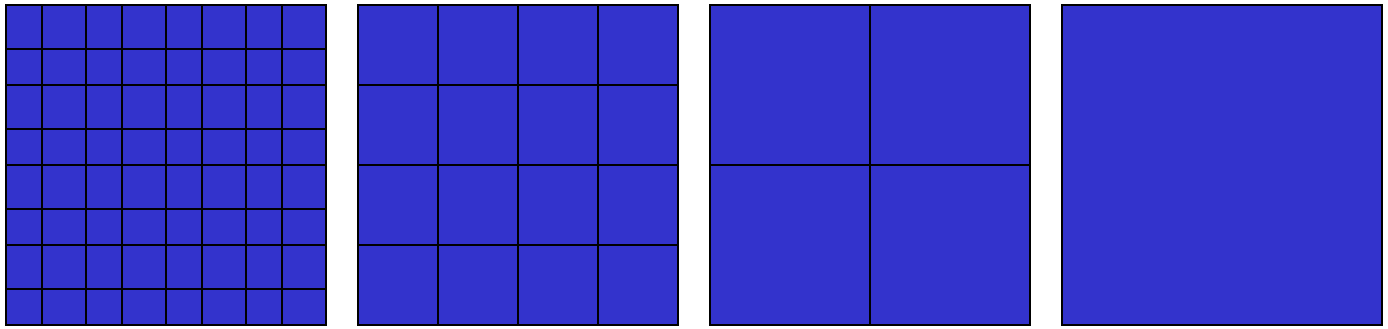
property **P1** ***on-top-of*** property **P2**

property **P1** ***next-to*** property **P2**



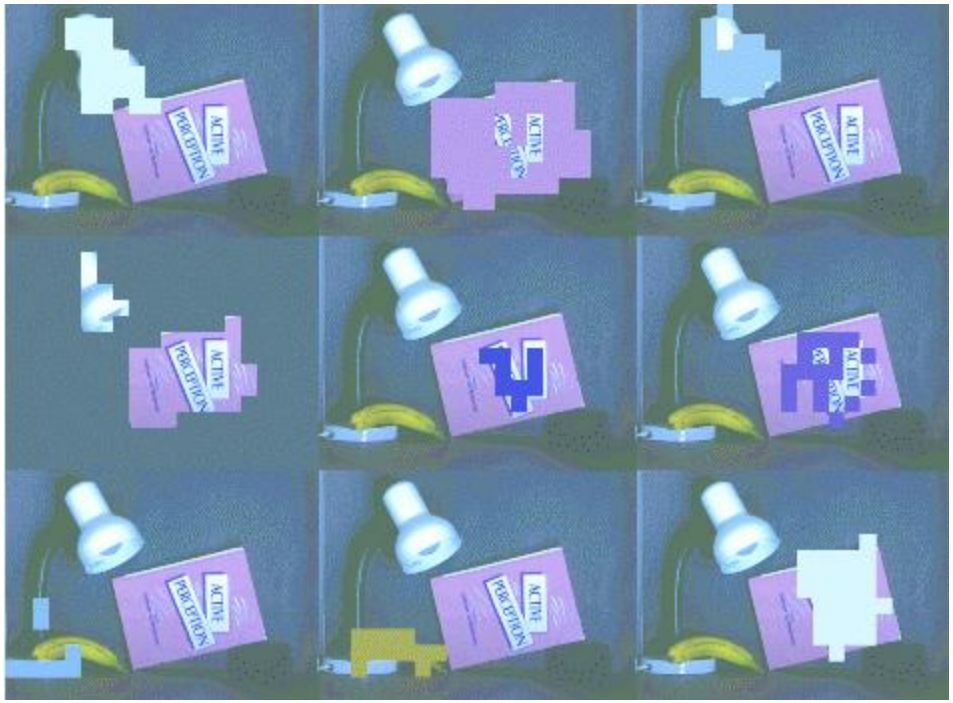


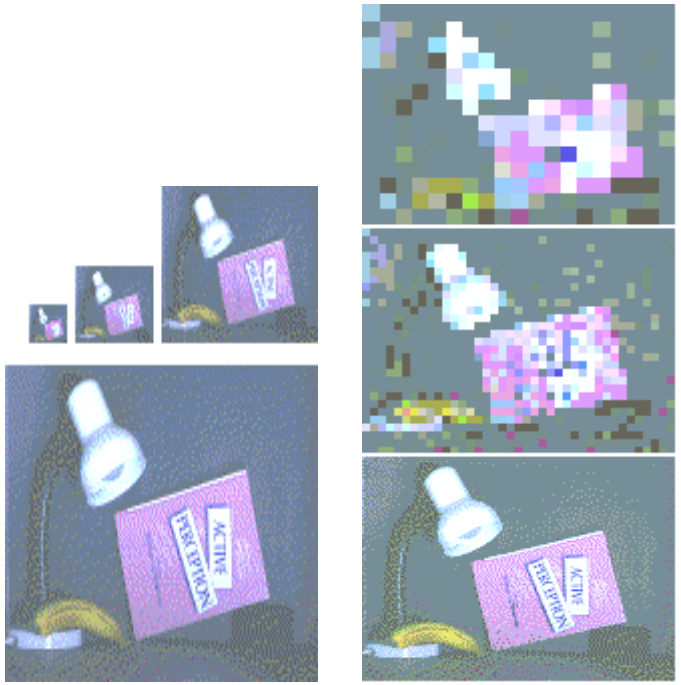
Different Resolution Hierarchy



Mining Multimedia Databases

From Coarse to Fine Resolution Mining





Challenge: Curse of Dimensionality

- Difficult to implement a data cube efficiently given a large number of dimensions, especially serious in the case of multimedia data cubes
- Many of these attributes are set-oriented instead of single-valued
- Restricting number of dimensions may lead to the modeling of an image at a rather rough, limited, and imprecise scale
- More research is needed to strike a balance between efficiency and power of representation

Summary

- Mining object data needs feature/attribute-based generalization methods
- Spatial, spatiotemporal and multimedia data mining is one of important research frontiers in data mining with broad applications
- Spatial data warehousing, OLAP and mining facilitates multidimensional spatial analysis and finding spatial associations, classifications and trends
- Multimedia data mining needs content-based retrieval and similarity search integrated with mining methods