Outline

• Motivation / problem definition
  • Main idea / time sequences
    • images
    • sub-pattern matching
    • automatic feature extraction / FastMap

Problem

Given a large collection of (multimedia) records (eg. stocks)
Allow fast, similarity queries
Applications

- time series: financial, marketing (click-streams!), ECGs, sound;
- images: medicine, digital libraries, education, art
- higher-d signals: scientific db (eg., astrophysics), medicine (MRI scans), entertainment (video)

Sample queries

- find medical cases similar to Smith's
- Find pairs of stocks that move in sync
- Find pairs of documents that are similar (plagiarism?)
- find faces similar to ‘Tiger Woods’

Detailed problem defn.:

Problem:
- given a set of multimedia objects,
- find the ones similar to a desirable query object

- for example:
distance function: by expert
(eg, Euclidean distance)

Types of queries

- whole match vs sub-pattern match
- range query vs nearest neighbors
- all-pairs query

Design goals

- Fast (faster than seq. scan)
- ‘correct’ (ie., no false alarms; no false dismissals)
Outline

- Motivation / problem definition
- Main idea / time sequences
  - images
  - sub-pattern matching
  - automatic feature extraction / FastMap

Main idea

- Eg., time sequences, ‘whole matching’, range queries, Euclidean distance

Main idea

- Seq. scanning works - how to do faster?
Idea: ‘GEMINI’

(GEneric Multimedia INdexIng)
Extract a few numerical features, for a ‘quick and dirty’ test

‘GEMINI’ - Pictorially

GEMINI

Solution: Quick-and-dirty filter:
• extract \( n \) features (numbers, eg., avg., etc.)
• map into a point in \( n \)-d feature space
• organize points with off-the-shelf spatial access method (‘SAM’)
• discard false alarms
GEMINI

Important: Q: how to guarantee no false dismissals?
A1: preserve distances (but: difficult/impossible)
A2: **Lower-bounding lemma**: if the mapping ‘makes things look closer’, then there are **no** false dismissals

• ‘proof’ of lower-bounding lemma

Lower-bounding:
Makes objects look closer to each other (& to query object)
**GEMINI**

- ‘proof’ of lower-bounding lemma

Lower-bounding:
Makes objects look closer to each other (& to query object)
=> ONLY false alarms

---

**GEMINI**

Important:
Q: how to extract features?
A: “if I have only one number to describe my object, what should this be?”

---

**Time sequences**

Q: what features?
Time sequences

Q: what features?
A: Fourier coefficients (etc)

Time sequences - results

- keep the first 2-3 Fourier coefficients
- faster than seq. scan
- NO false dismissals (see book)

<table>
<thead>
<tr>
<th>time</th>
<th># coeff. kept</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>total</td>
</tr>
<tr>
<td></td>
<td>cleanup-time</td>
</tr>
<tr>
<td></td>
<td>r-tree time</td>
</tr>
</tbody>
</table>

Time sequences - improvements:

- improvements/variations: [Kanellakis +Goldin], [Mendelzon+Rafiei]
- could use Wavelets, or DCT
- could use segment averages [Yi+2000]
Outline

- Motivation / problem definition
- Main idea / time sequences
- images
  - sub-pattern matching
  - automatic feature extraction / FastMap

Images - color

What is an image? 
A: 2-d array

Color histograms, and distance function
Mathematically, the distance function is:

\[
\text{distance}_{\text{histogram}}(\vec{x}, \vec{y}) = \begin{bmatrix}
R_R & R_P & \ldots \\
R_R & R_P & \ldots \\
\ldots & \ldots & \ldots \\
\end{bmatrix} (\vec{x} - \vec{y})^T \\
\ldots = (\vec{x} - \vec{y}) \times (\vec{x} - \vec{y})^T
\]

Problem: ‘cross-talk’:
- Features are not orthogonal ->
- SAMs will not work properly
  
- Q: what to do?
- A: feature-extraction question

Possible answers:
- avg red, avg green, avg blue

It turns out that this lower-bounds the histogram distance ->
- no cross-talk
- SAMs are applicable
Images - color

Multimedia - Detailed outline

- multimedia
  - Motivation / problem definition
  - Main idea / time sequences
  - images (color; shape)
  - sub-pattern matching
  - automatic feature extraction / FastMap

Images - shapes

- distance function: Euclidean, on the area, perimeter, and 20 'moments'
- (Q: how to normalize them?)
Images - shapes

• distance function: Euclidean, on the area, perimeter, and 20 'moments'
• (Q: how to normalize them?
• A: divide by standard deviation)

• distance function: Euclidean, on the area, perimeter, and 20 'moments'
• (Q: other 'features' / distance functions?

• distance function: Euclidean, on the area, perimeter, and 20 'moments'
• (Q: other 'features' / distance functions?
• A1: turning angle
• A2: dilations/erosions
• A3: ... )
Conclusions

• GEMINI works for any setting (time sequences, images, etc)
• uses a ‘quick and dirty’ filter
• faster than seq. scan
• (but: how to extract features automatically?)

Outline

• Motivation / problem definition
• Main idea / time sequences
• images
• sub-pattern matching
• automatic feature extraction / FastMap

FastMap

Automatic feature extraction:
• Given a dissimilarity function of objects
• Quickly map the objects to a (k-d) ‘feature’ space.
• (goals: indexing and/or visualization)
FastMap

- Multi-dimensional scaling (MDS) can do that, but in $O(N^{**2})$ time
Main idea: projections

We want a linear algorithm: FastMap [SIGMOD95]

FastMap - next iteration

Results

Documents /cosine similarity -> Euclidean distance (how?)
Results

Applications: time sequences
- given $n$ co-evolving time sequences
- visualize them + find rules [ICDE00]

Applications - financial
- currency exchange rates [ICDE00]
Variations

- Isomap [Tenenbaum, de Silva, Langford, 2000]
- LLE (Local Linear Embedding) [Roweis, Saul, 2000]
- MVE (Minimum Volume Embedding) [Shaw & Jebara, 2007]

Conclusions

- GEMINI works for multiple settings
- FastMap can extract ‘features’ automatically (-> indexing, visual d.m.)
Main references

- Myron Flickner, et al: 
  *Query by Image and Video Content: the QBIC System* 
- Journal of Intelligent Inf. Systems, 3, 3/4, pp. 231-262, 1994 (An earlier, more technical version of the IEEE Computer '95 paper.)
References


References

- Lawrence Saul & Sam Roweis. An Introduction to Locally Linear Embedding (draft)

References