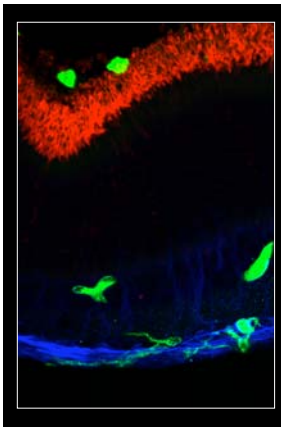


Indexing spatially sensitive distance measures

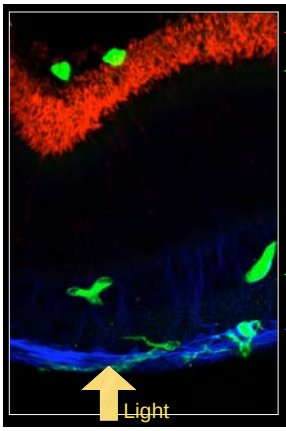
using multi-resolution lower bounds

Vebjorn Ljosa
Arnab Bhattacharya
Ambuj K. Singh

University of California,
Santa Barbara

EDBT 2006
Proceedings p. 865

1



Photoreceptors
(labeled by anti-rhodopsin)

Macrophages
(labeled by isolectin B4)

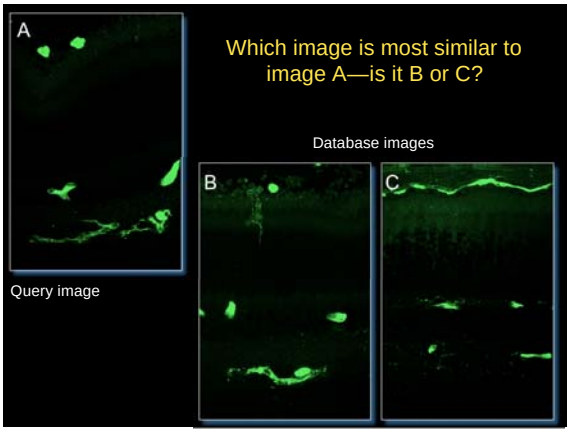
Confocal micrograph of cat retina
(by Geoff Lewis, Fisher lab, UCSB)

Microglia and blood vessels
(labeled by isolectin B4)

Müller cells
(labeled by anti-glial fibrillary acidic protein)

Light

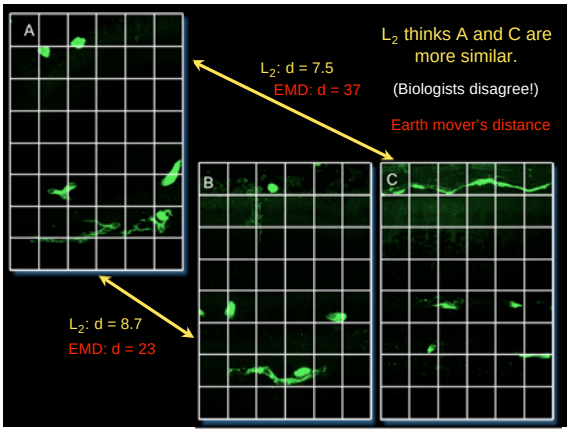
2



Which image is most similar to image A—is it B or C?

Database images

Query image



L_2 thinks A and C are more similar.

L_2: d = 7.5
EMD: d = 37

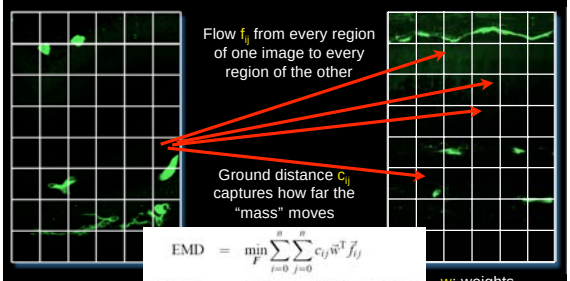
(Biologists disagree!)

Earth mover's distance

L_2: d = 8.7
EMD: d = 23

The earth mover's distance (EMD)

[Werman et al., 1985; Peleg et al., 1989; Rubner et al. 2000]



Flow f_{ij} from every region of one image to every region of the other

Ground distance c_{ij} captures how far the "mass" moves

LP-problem:

$$\text{EMD} = \min_{F} \sum_{i=0}^{n} \sum_{j=0}^{n} c_{ij} \vec{w}^{\mathrm{T}} \vec{f}_{ij}$$

$$\text{subject to} \quad \vec{f}_{ij} \geq [0 \cdots 0]^{\mathrm{T}} \text{ (element-wise)},$$

$$\forall i, \sum_{j=0}^{n} \vec{f}_{ij} = \vec{a}_i, \qquad \forall j, \sum_{i=0}^{n} \vec{f}_{ij} = \vec{b}_j$$

w: weights

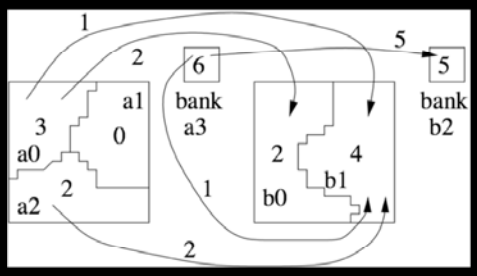
… but it's slow!

Outline

- EMD on feature vectors
  - Formulation with special "bank" region
  - Decompose EMD on feature vectors into many smaller LP-problems
    - Faster, uses less memory
- Lower bound for EMD
  - Spatially motivated
  - Faster to compute, using summary of image
  - Multiple resolutions
  - Range and k-NN query algorithms that use the lower bounds
    - Sequential scan and M-tree index structure
- Experimental results

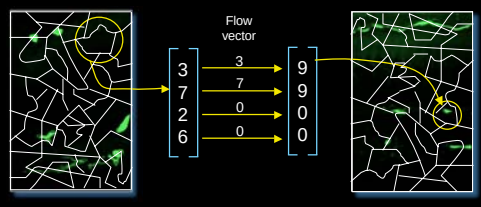
6

Introducing the “bank” region



Decomposing the EMD

Flows are between the same dimensions of the feature vectors.



In other words, there is no crosstalk.

Reducing crosstalk

- No crosstalk between independent dimensions
  - Color Layout Descriptor: DCT coefficients
  - Orthogonal bases found by PCA
- Sometimes crosstalk only between some dimensions
  - Concatenated feature vectors for two independent proteins
  - Cluster dimensions, no crosstalk between dimensions in different clusters

Decomposition makes the LP-problem smaller

97 regions 97 regions

$$\min_{F} \sum_{i=0}^{n}\sum_{j=0}^{n} c_{ij} \sum_{k=1}^{m} w_k f_{ijk} = \sum_{k=1}^{m} \min_{F_k} \sum_{i=0}^{n}\sum_{j=0}^{n} c_{ij} w_k f_{ijk}$$

12 dimensions

Large LP problem

41 s
37 MB main memory

97 x 97 x 12 variables

12 small LP problems

2.9 s in total
5 kB main memory

97 x 97 variables for each problem

Outline

- EMD on feature vectors
  - Formulation with special “bank” region
  - Decompose EMD on feature vectors into many smaller LP-problems
    - Faster, uses less memory
- Lower bound for EMD
  - Spatially motivated
  - Faster to compute, using summary of image
  - Multiple resolutions
  - Range and k-NN query algorithms that use the lower bounds
    - Sequential scan and M-tree index structure
- Experimental results

Decomposition helps, but is not enough

$$\rho_{AB} = \sum_{k=1}^{m} \min_{F_k} \sum_{i=0}^{n}\sum_{j=0}^{n} c_{ij} w_k f_{ijk}$$

12 dimensions

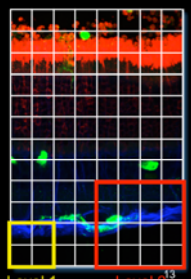
97 regions (8 x 12 tiles + bank)

Decomposed:
97 x 97 variables, 12 dimensions

Number of variables quadratic in the number of regions

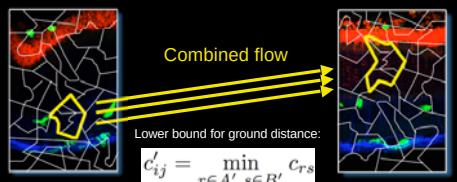
Spatially motivated lower bound

- Cannot just use larger regions
  - Regions must be small enough to fit in layers of tissue
- Idea
  - Compute distance using larger regions
  - Modify the distance function so this distance is a lower bound for the EMD using smaller regions
  - Multiple resolutions



13

Lower bound by assuming best-case ground distance



Lower bound for ground distance:

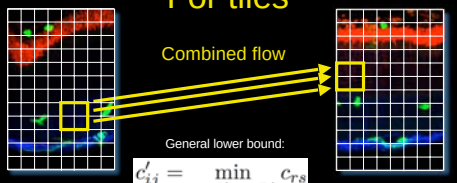
$$c'_{ij} = \min_{r \in A'_i, s \in B'_j} c_{rs}$$

$$\min_{F} \sum_{i=0}^{n} \sum_{j=0}^{m} c_{ij} \boldsymbol{w}^{\mathrm{T}} \boldsymbol{f}_{ij} \geq \min_{F'} \sum_{i=0}^{n'} \sum_{j=0}^{m'} c'_{ij} \boldsymbol{w}^{\mathrm{T}} \boldsymbol{f}'_{ij}$$

Proof is by decomposing flows and using c' ≤ c for each.

14

For tiles



General lower bound:

$$c'_{ij} = \min_{r \in A'_i, s \in B'_j} c_{rs}$$

For tiles, the ground distance has a simple formula:

$$c'_{i'j'} = \begin{cases} \left[\max\{0, 2|i'_x - j'_x| - 1\}^2 + \max\{0, 2|i'_y - j'_y| - 1\}^2\right]^{\frac{1}{2}} & \text{if } i', j' \neq n' \\ 0 & \text{if } i' = j' = n' \\ \alpha & \text{otherwise} \end{cases} \quad (4)$$

15

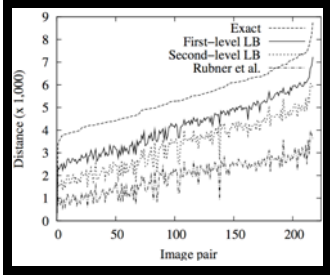
Outline

- EMD on feature vectors
  - Formulation with special "bank" region
  - Decompose EMD on feature vectors into many smaller LP-problems
    - Faster, uses less memory
- Lower bound for EMD
  - Spatially motivated
  - Faster to compute, using summary of image
  - Multiple resolutions
  - Range and k-NN query algorithms that use the lower bounds
    - Sequential scan and M-tree index structure
- Experimental results

16

Properties of the lower bound

Distance from image to each of 217 other images



Fast
- Full EMD: 2.9 s
- 1st level: 60 ms
- 2nd level: 4 ms

Tight
- 1st level: 25% lower
- 2nd level: 44% lower
- Rubner: 68% lower

[Rubner et al., 2000]

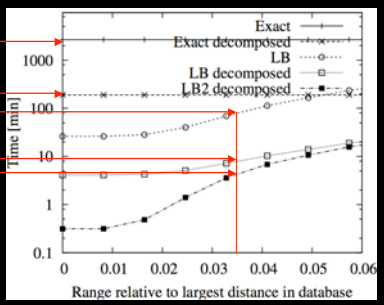
8 x 12 tiles, Color Layout Descriptor (12-D)

17

Effect of decomposition and lower bound

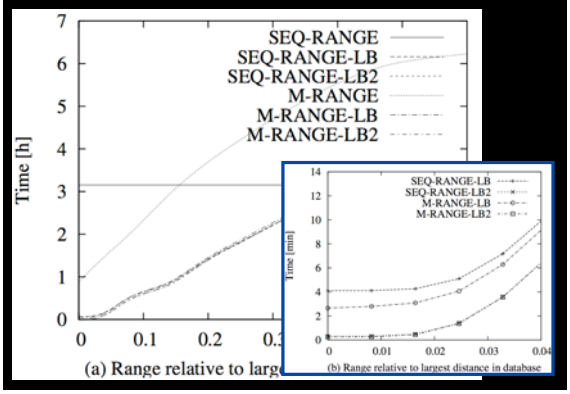
Range search on 3,932 retinal images

40 h
3.2 h
1.5 h
9 min
5 min

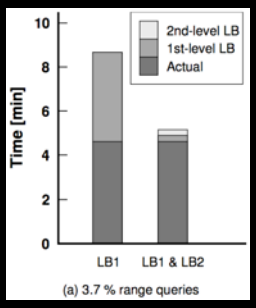


18

Effect of lower bound and index structure — Range queries



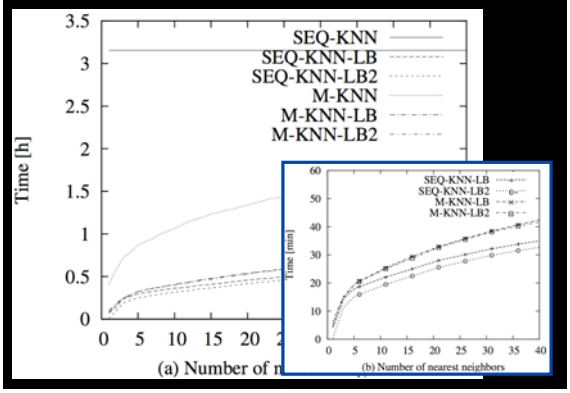
Breakdown of range query time



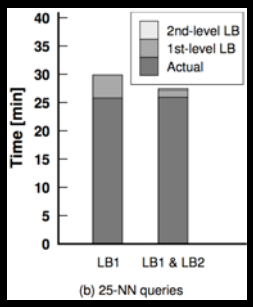
Search without lower bounds takes 3.2 h (not shown).

20

Effect of lower bound and index structure — k-NN queries



Breakdown of 25-NN query time



Search without lower bounds takes 3.2 h (not shown).

22

Summary

- EMD is a useful distance measure
  - Combines feature distance and spatial distance
- Techniques for speeding up EMD computation
  - Reduce search times up to 500 times
    - From 40 hours to 5 minutes
  - Make EMD viable, even when many regions are necessary
- Future
  - Integrate into BISQUE, the database infrastructure of the Bioimage project at UCSB (www.bioimage.ucsb.edu)
  - Other datasets
  - Crosstalk
  - Compare with new lower bounds (Assent et al., ICDE 2006)

Images by Geoff P. Lewis (Lab. of S.K. Fisher, UCSB)
Supported in part by grant ITR-0331697 from NSF.

23