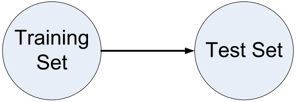
Pattern Recognition in Diffusion Spaces

Yosi Keller

School of Engineering, Bar-Ilan University [1]

[1] This research was supported by NSF grant 0534052 and ONR grant N00014-04-1-0725.

Outline I



Pattern recognition

1. Spectral embeddings provide numerical tools for pattern recognition:
   - Dimensionality reduction
   - Density invariant embeddings
   - Embedding extension scheme
2. Spectral embeddings can be interpreted in several ways:
   - Dimensionality reduction: *geometric approach*
   - Diffusion distances: *Markov walks approach*

Outline II

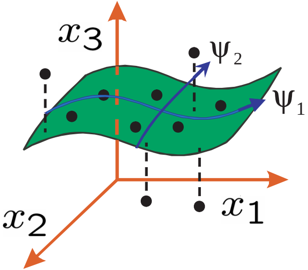
- Data adaptive basis and extension of functions over data sets: *signal processing over manifolds*
- Spectral relaxation of integer optimization problems: *graph/set alignment and set spaces*

Collaborators

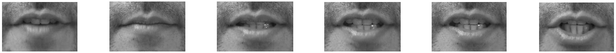
- Ronald Coifman, Yale.
- Stephane Lafon, Google.
- Amit Singer, Yale
- Steven W. Zucker, Yale
- Michael Chertok, Bar Ilan

Curse of dimensionality & Dimensionality Reduction

- Density estimation is difficult
- Computational cost of many algorithms grows exponentially with the dimension.
- Certain signals are in essence low-dimensional and their high dimensional representation is due to over sampling and noise.
- The high dimensional representation obscures the underlying low dimensional structure.
- Certain tasks only require low-dimensional representations.



Example: audio visual lipreading



Two input channels:

1. Video frames: R^{16000}: 25fps.
2. Audio frames: each 40ms sampled in 32khz: R^{1280}.

We aim to recognize a limited vocabulary: $\{$“0”,“1”,...,“9”$\}$.
Each word consists of a *set* of samples: *set recognition*
We can also consider the recognition of single samples.

The role of random projections

[Johnson-Lindenstrauss84]

- A fast algorithm for dimensionality reduction $n \rightarrow O(\log(n))$ while preserving the L_2 distances.
- In practice, since spectral embeddings schemes use L_2 distances as inputs, $n = O\left(10^4\right)$ at most.
- This also applies to metric trees [Liu,Moore,Gray,Yang2004].

What's wrong with L_2 distances?

"Short distances good, long distances bad"

This is because the data lies on a low dimensional manifolds: in this example the two rotation angles.

Kernel methods

Definition

Given a dataset $\{x_i\}_{i=1..n}$:

1. Apply a p.s.d. kernel k to $\{x_i\}$. For instance: $w_{ij} = \exp(-\frac{d(x_i,x_j)}{\varepsilon}), \varepsilon > 0$.
2. Compute the eigenvectors of W: $w_{ij} = \sum_{l\geq 0} \lambda_l \psi_l(i)\psi_l(j)$,
3. The embedding is given by $\Psi(x_i) : x_i \mapsto \left(\lambda_1\psi_1(x_i), \lambda_2\psi_2(x_i), \ldots\right)$ $\|\Psi_t(x_i) - \Psi_t(z_i)\|^2_{L_2} = \sum_{l=0}^{n-1} \lambda_l^{2t}(\psi_l(x) - \psi_l(z))^2 =$ $w_{ii} + w_{jj} - 2w_{ij} = D(x,z)^2$

Kernel methods survey

- MDS - Cox and Cox. *Multidimensional Scaling*. Chapman & Hall, 2nd edition, 2001.
- ISOMAP - Josh. Tenenbaum, Vin de Silva, John Langford 2000, *A Global Geometric Framework for Nonlinear Dimensionality Reduction*
- LLE - S. Roweis and L. Saul, *Nonlinear Dimensionality Reduction by Locally Linear Embedding,* Science 2000
- Hessian LLE - D. Donoho and C. Grimes, *Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data,* PNAS 2003
- Laplacian Eigenmaps - M. Belkin and P. Niyogi, *Laplacian Eigenmaps and Spectral Techniques for Embedding and Clustering*. NIPS2001
- Diffusion maps - *Geometric diffusions as a tool for harmonic analysis and structure definition of data*, PNAS2005

Geometric interpretation: Laplacian eigenmaps[Belkin-Nyogi, NIPS2002]

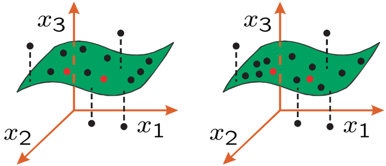
The embedding preserves the infinitesimal geometry of a low dimensional *manifold* M. The distortion of an embedding f is locally bounded by $|f(z) - f(x)| \leq d_M(z,x) \, \|\nabla_M f(x)\|$ and

$$\int_M \|\nabla_M f\|^2 = \int_M f \cdot \Delta_M f \approx \sum_{i,j} W_{ij} \left(f_i - f_j\right)^2 = f^T L f$$
$$f = f^T L f, \text{ s.t. } f^T D f = 1, f^T D \underline{1} = 0$$
$$x_i \mapsto \left(\psi_1(x_i), \psi_2(x_i), \ldots\right)$$

The Graph Laplacian becomes a natural choice for a kernel.

Density invariant embeddings I

[Coifman,Lafon,et. al,PNAS2005],[Keller,Lafon,Coifman,PAMI2006]



Same manifold, same points in $\mathbb{R}^3$, different densities

$\Downarrow$

different embeddings in $\mathbb{R}^2$

Density invariant embeddings II

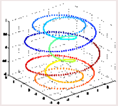
[Coifman,Lafon,et. al,PNAS2005],[Keller,Lafon,Coifman,PAMI2006]

Density invariant embedding

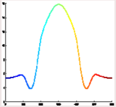
Given $w_{ij} = \exp(-\frac{\left\|x_i,x_j\right\|^2}{\varepsilon})$, define $q_i \triangleq \sum_j w_{ij}$,and form the new kernel $\widetilde{w}_{ij} = \frac{w_{ij}}{q_i q_j}$.

Now continue with the regular graph Laplacian using $\widetilde{w}_{ij}$.

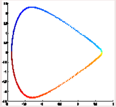
Curve



Density

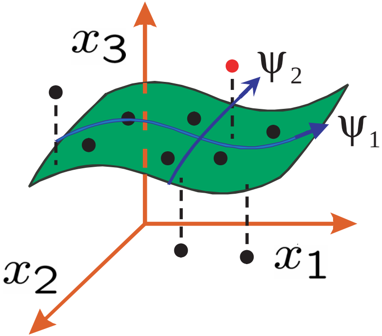


Graph Laplacian



Laplace-Beltrami

Out of sample extension I
[Lafon,Keller,Coifman,PAMI2006]



Given the parametrization $\{\psi_l\}$ computed using $N \gg 1$ samples, extend the embedding to the new point y, without re-embedding the $N+1$ data set.

- This is based on a Nyström extension with in an extension kernel.
- Differs from [Fowlkes, Belongie, Chung, Malik,PAMI2004].

Out of sample extension II
[Lafon,Keller,Coifman,PAMI2006]

Spectral low pass extension

Given a p.s.d. symmetric $\widetilde{k}$ with $\widetilde{\varepsilon} \gg \varepsilon$ and its eigensystem $\left\{\widetilde{\psi}_l,\widetilde{\lambda}_l\right\}$. $\widetilde{\psi}_l$ can be approximated beyond $x \in \bar{X}$ by

$$\widetilde{\psi}_l(y) = \frac{1}{\widetilde{\lambda}_l} \sum_{z \in X} \widetilde{k}(y,z)\widetilde{\psi}_l(z),\ \lambda_l > C,\ \forall y$$

and used to extend $\{\psi_l\}$

$$\psi_l = \sum_l \langle \psi_l, \widetilde{\psi}_l \rangle_X \widetilde{\psi}_l,\ \forall y$$

The extension kernel $\widetilde{k}$ differs from the embedding kernel k.

Out of sample extension III
[Lafon,Keller,Coifman,PAMI2006]

The $\frac{1}{\widetilde{\lambda}_l}$ term limits the number of eigenfunctions ψ_l that can be extended, since the eigenvalues of the laplacian are decaying. This is a natural MDL criterion for kernel based learning.

The width of the extension kernel $\widetilde{\varepsilon}$

	Kernel	Approx.	Extens.	Task	Learning set
$\widetilde{\varepsilon} \longrightarrow 0$	narrow	good	poor	interpolation	large
$\widetilde{\varepsilon} \gg 0$	wide	poor	good	extrapolation	small

[Lafon-Coifman,ACHA2006]: *Geometric harmonics*: iterative refinement of $\widetilde{\varepsilon}$ by optimizing the trade-off between the approximation and extension errors.
The approximation error depends on the extended function, ψ_l in this case.

Implementation issues for massive datasets

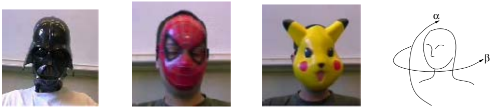
These schemes utilize two numerical workhorses:

- Embedding
  1. n by n Gauss transform
  2. SVD: $n \times n$ matrix
- Extension to m points
  1. m by n Gauss transform

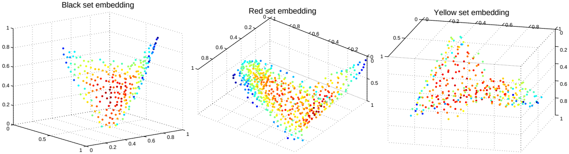
Solution

- *SVD: Random projections based approaches [Tygert,Rokhlin,Martinsson]*
- *Gauss transform: FMM, FGT [Greengard,Strain], IFGT [Yang, Duraiswami, Gumerov], Dual trees [Gray,Moore]*

High dimensional data alignment
[Lafon,Keller,Coifman,PAMI2006]



Input: 3000 frames each being $\mathbb{R}^{10000}$.
Output: align the heads based on a common low dimensional manifold.
Problem: each manifold is sampled with a different density $\rightarrow$ **Laplace Beltrami**.

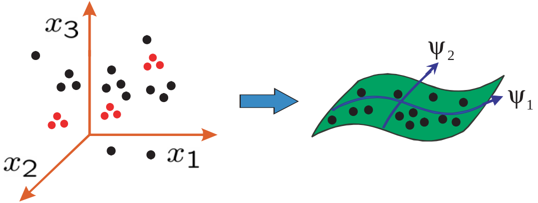


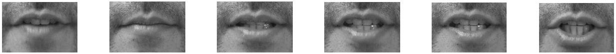
The low-dimensional embeddings are then aligned.

Audio-Visual lip reading I
[Lafon,Keller,Coifman,PAMI2006]

We split the learning process into two parts:

1. Embedding/learning the global manifold

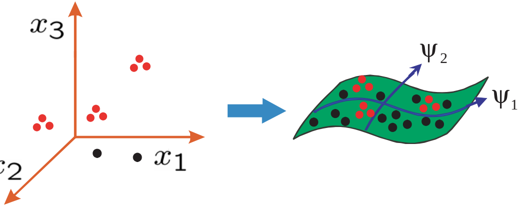




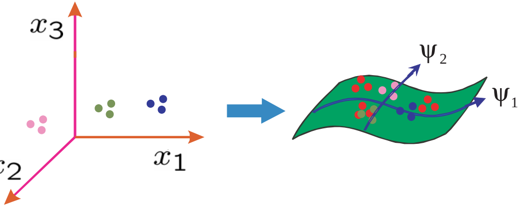
6000 frames

Audio-Visual lip reading II
[Lafon,Keller,Coifman,PAMI2006]

2. Using the labeled samples to compute signatures.



3. Identification



Audio-Visual lip reading III
[Lafon,Keller,Coifman,PAMI2006]

Results

Channel	"0"	"1"	"2"	"3"	"4"	"5"	"6"	"7"	"8"	"9"
Visual	0.90	0.99	0.90	0.94	0.93	0.81	0.87	0.74	0.75	0.82
Audio	0.75	0.94	0.87	0.90	0.96	0.86	0.93	0.81	0.80	0.92

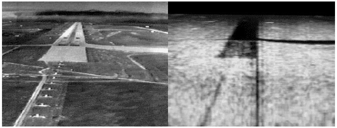
Remarks

- We used the L_2 metric for visual data and the cepstrum of the auditory data.
- Dimensionality reduction $\mathbb{R}^{16000} \rightarrow \mathbb{R}^{10}$.
- Recognition based on Hausdorff distance in $\mathbb{R}^{10}$.

Multisensor based recognition: lip reading I
[Keller,Lafon,Zucker2007]

Can we unify different sensors for signal recognition and analysis?

The good: Multisensor data is often of complementary nature.
The bad:



[Irani-Anandan,ICCV1998]

- Different sensors are related by unknown non-linear relationships

Multisensor based recognition: lip reading II
[Keller,Lafon,Zucker2007]

- Correspond to a common low dimensional manifold.
- Different sensors have different sampling rates and resolutions

Our approach

1. Embed each of the channels separately using the Laplace Beltrami.
2. Append the embeddings to get new coordinates.
3. Apply a pattern recognition scheme.

Results

	"0"	"1"	"2"	"3"	"4"	"5"	"6"	"7"	"8"	"9"
Audio	0.75	0.94	0.87	0.90	0.96	0.86	0.93	0.81	0.80	0.92
Visual	0.90	0.99	0.90	0.94	0.93	0.81	0.87	0.74	0.75	0.82
Both	0.90	0.99	0.96	0.99	0.96	0.97	0.90	0.93	0.95	0.96

Inducing Random walks on graphs I

Given a dataset $\{x_i\}$:

1. Apply a p.s.d. kernel k to $\{x_i\}$. For instance:
   $w_{ij} = \exp(-\frac{\|x_i, x_j\|^2}{\varepsilon})$. $\varepsilon > 0$ is the scale factor.
   w_{ij} describes the **infinitesimal** geometry of X up to $2\sqrt{\varepsilon}$.
2. Compute the Markov matrix $P = D^{-1}W$, $d_{ii} = \sum_j w_{ij}$
3. Compute the eigenvectors of $P_t = P^t$:

$$p_t(x,y) = \sum_{l \geq 0} \lambda_l^t \psi_l(x) \phi_l(y)\,,$$

The embedding is given by

Inducing Random walks on graphs II

$$x_i \mapsto \Psi_t(x_i) = \left(\lambda_1^t \psi_1(x_i), \lambda_2^t \psi_2(x_i), \dots\right)$$

If W is symmetric and $w_{ij} \geq 0$ then P and the graph Laplacian $L = D - W$ share the same eigenvectors.

Interpretations:

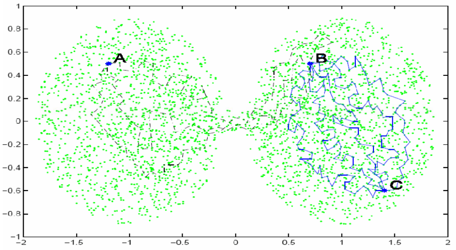
1. The mixing time of Markov random chains, *Spectral Graph Theory*, Fan R. K. Chung.
2. Lumpable Markov chains and piecewise constant right eigenvectors Meila and Shi, *A random walks view of spectral segmentation,* AISTATS 2001.
3. Stability analysis, *On Spectral Clustering: Analysis and an algorithm*, Y. Ng, M. Jordan, and Y Weiss, NIPS2001.

Diffusion distances
[Lafon,Coifman]

What is the distance measure induced by using the graph laplacian as a kernel?

$$\|\Psi_t(x_i) - \Psi_t(z_i)\|_{L_2}^2 = \sum_{l=0}^{n-1} \lambda_l^{2t} (\psi_l(x) - \psi_l(z))^2 =$$

$$\sum_{y \in \Omega} \frac{(p_t(x,y) - p_t(z,y))^2}{\phi_0(y)} = D_t(x,z)^2$$



Clustering Vs. Pattern recognition

Different approaches answer different questions

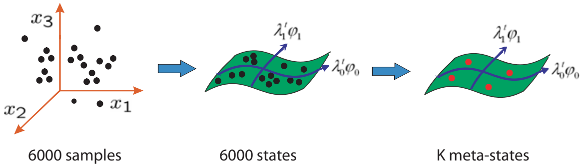
- Laplacian eigenmaps [Belkin,Nyogi] and Diffusion distances [Lafon,Coifman]:
  - What is the meaning of spectral dimensionality reduction?
  - Applicable to **pattern recognition**.
- Lumpable Markov chains[Meila,Shi], Spectral clustering [Y.Ng,M. Jordan,Y. Weiss]:
  - Why does spectral clustering work?
  - Applicable to **clustering**.

Revealed Markov models
[Keller,Singer,Coifman]

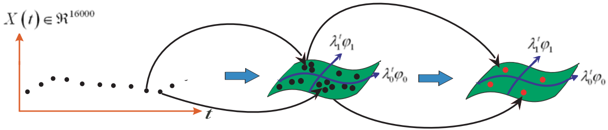
- Given a time series $x\left(t\right)$ $t \in \left[t_0, \ldots, t_{\max}\right]$
- Initially, $x\left(t\right)$ is considered a data set $\left\{x\right\}_t$
- Spectral embedding induces a random walk on the data - each sample is mapped to a state.
- The diffusion distance allows us to quantize the state space optimally:
  given the target states number, we minimize the L_2 quantization distortion in the embedding space (KMEANS).

Spectral embeddings and Markov walks II

We merge Markov states into K metastates:

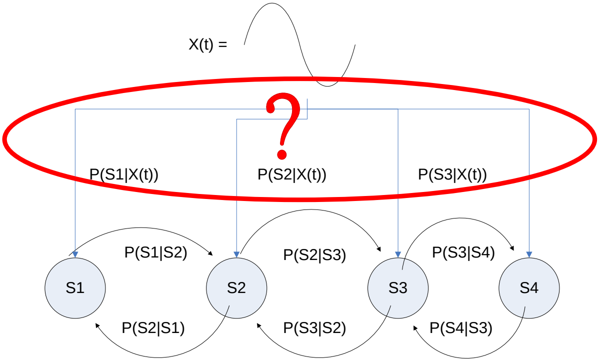


Now we can compute the transition probabilities $\{\pi_{ij}\}$, $i, j = 1..K$



Learning

We compute the state and transition probabilities: *Learning*

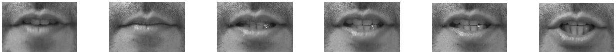


Pattern recognition: Audio-Visual lip reading

Separating the state-space from the density

Two inputs:

1. The 6000 frames are recordings of the speaker reading an article - **used to model the state-space**.



2. 50 sequences of the speaker speaking the digits 1...10 - **used to learn the dynamics of each digit**.

Similar to Bag-of-words models [Hoffman 1999; Blei, Ng & Jordan, 2004; Teh, Jordan, Beal & Blei, 2004], documents are represented as probability densities over the Bag-of-words (the state space).

Maximum likelihood classifiers I

Let: $D = \{d_1, ..., d_{10}\} = \{\text{“1”}, ..., \text{“10”}\}$
$\{S_i\}$ – a set of input states

Maximizing the state probabilities

$$k^* = \max_k \sum_i \log P\left(S_i | d = d_k\right)$$

Maximizing the transition probabilities

$$k^* = \max_k \sum_i \log P\left(S_i | S_{i-1}, d = d_k\right)$$

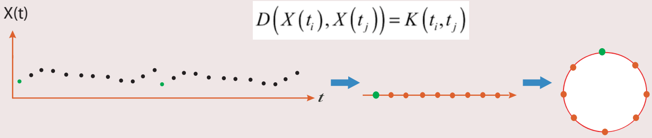
	“0”	“1”	“2”	“3”	“4”	“5”	“6”	“7”	“8”	“9”
State	0.80	0.91	0.90	0.86	0.95	0.90	0.96	0.69	0.83	0.92
Trans	0.88	1.00	0.95	0.86	1.00	1.00	0.96	0.82	0.72	0.92

Spectral embedding vectors as an adaptive basis [Keller]

The kernel k is p.s.d, hence its eigenvectors (the embedding) form an orthonormal system $\{\psi_i\}$.

Example

Given a periodic discrete signal $x(n)$, $n = 1 \dots N$, and a time invariant metric $D(x(t_1), x(t_2)) = D(|t_1 - t_2|)$



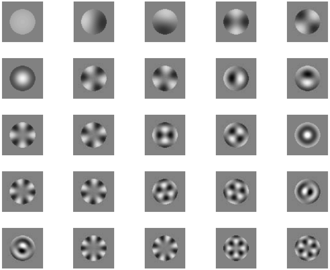
- A circle is parameterized by $e^{\frac{2\pi i}{N} n}$, $n = 1 \dots N$.
- For any kernel K, the Laplacian is a circulant matrix, diagonalized by the Fourier basis.

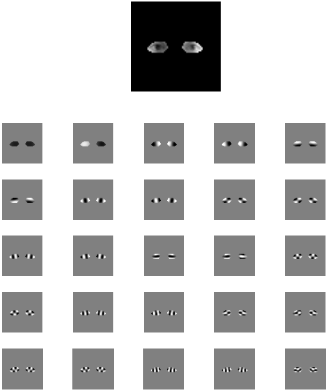
Sinc interpolation of periodic functions

- Start with the Fourier basis $\{\psi_i\}$ on the circle
  $f(x) = \sum_i a_i \psi_i(x)$, where $a_i = \langle f, \psi_i(x) \rangle$
- Extend $\{\psi_i\}$ from $X \to y$ using the Nystrom extension
  $\psi_i(y) = \frac{1}{\lambda_i} \sum_x K(x,y)\, \psi_i(x)$
- Extend f from $X \to y$
  $f(y) = \sum_i a_i \psi_i(y) = \sum_i a_i \frac{1}{\lambda_i} \sum_x K(x,y)\, \psi_i(x) =$
  $\sum_x K(x,y) \sum_i \frac{1}{\lambda_i} a_i \psi_i(x)$
- Set $K(x,y) = \frac{\sin(\pi(y-x))}{\sin\left(\frac{\pi(y-x)}{N}\right)} \to \lambda_i = 1$. K is the Dirichlet kernel.
  $f(y) = \sum_x \frac{\sin(\pi(y-x))}{\sin\left(\frac{\pi(y-x)}{N}\right)} \sum_i a_i \psi_i(x) = \sum_x \frac{\sin(\pi(y-x))}{\sin\left(\frac{\pi(y-x)}{N}\right)} f(x)$

Fourier bases on irregularly-shaped domains

[Saito2005]





Pattern recognition as a Function extension problem

Definition

Let f be the classification function: $f(x) = \begin{cases} x \in C & 1 \\ x \notin C & \text{-1} \end{cases}$

Definition

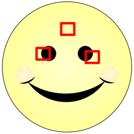
Given a kernel k with and its eigensystem $\{\psi_l, \lambda_l\}$ and a scalar function f

1. Compute the inner products $a_i = \langle \psi_l, f \rangle$
2. Extend the eigenfunction $\widetilde{\psi}_l(y) = \frac{1}{\lambda_l} \sum_{z \in X} k(y,z)\psi_l(z), y \notin X$
3. Extend the function f: $\widetilde{f} = \sum_l \langle \widetilde{\psi}_l, f \rangle_X \widetilde{\psi}_l, \forall y$

Example: eye detection I

Features extraction: SIFT [Lowe2003]:

- Strong translation and illumination invariance.
- Weak rotation invariance.
- Local scale estimation - strong scale invariance.
- Dominant angle estimation - strong rotation invariance.
- Estimation of local moments - affine invariance [Schaffalitzky,Zisserman2002],



Do not learn what you already know

Example: eye detection II

Learning

- We consider each SIFT descriptor as a sample in R^{128}.
- We collect a *learning set* of $\{P_i\}_1^N$ patches $\{f_i\}_1^N \in \{-1, 1\}$.
- Embed $\{P_i\}_1^N$ and compute $\{\psi_k\}$.
- Compute the inner products $\alpha_k = \langle f, \psi_k \rangle$

Example: eye detection III

Recognition

- Given an input face, we extract the patches $\left\{\widehat{P}_k\right\}_1^{\widehat{N}}$.
- $\{\psi_l\}$ and $\{f_i\}_1^N$ are extended to $\left\{\widehat{P}_k\right\}_1^{\widehat{N}}$.
- Use α_k to extend $\{f_k\}_1^{\widehat{N}}$ to $\left\{\widehat{P}_i\right\}_1^{\widehat{N}}$.
- The classification is given by $\begin{cases} f_k > 0 & \widehat{P}_k \in C \\ f_k < 0 & \widehat{P}_k \notin C \end{cases}$

Show demo

Example: eye detection IV

Comparison to Yann LeCun's talk

- Yann advocated using raw pixels:
  - The learning set becomes larger
  - The invariance is inserted later via the learning of the metric
  - It easy to normalize the intensity of patches - this is not the general case: recall the 9 Vs. 4 example.
- Holistic Vs. Gestalt approaches to pattern recognition.

Other application: image colorization

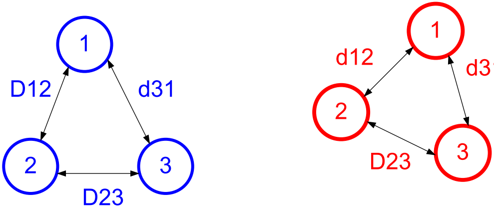
Input

Output

Original

Spectral embedding as a relaxation of integer optimization problems
[Chertok,Keller]

Consider the following set alignment problem



We are given two sets of points: S_k $k = 1, 2$ and the relative distances within each set: $d^k_{i,j}$ $i, j = 1..|S_k|$.

Definition

Alignment vector

$x = \begin{bmatrix} 0 & 1 & 0 | & 1 & 0 & 0 | & 0 & 0 & 1 \end{bmatrix}^T$

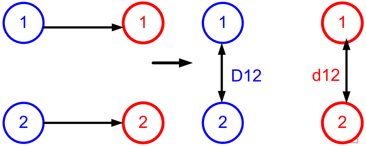
Definition

Assignment $C_{i\hat{i}} : S_1^i \rightarrow S_2^{\hat{i}}$

Definition

Assignment cost for pairs

$d\left(C_{i\hat{i}}, C_{j\hat{j}}\right) = d\left(d_{ij}, d_{\widehat{ij}}\right) = \left|d_{ij} - d_{\widehat{ij}}\right|$: what is the cost of both of the assignments being valid?



Definition

Assignment affinity matrix $a_{\hat{i}i,\hat{j}j} = \exp\left(-d\left(d_{ij}, d_{\hat{i}\hat{j}}\right)/\sigma\right), \sigma > 0$

We can now define the total assignment affinity and maximize it:

$$x^* = \arg\max_x X^T A X$$

where X is an assignment vector: binary + constraints
This is a difficult optimization problem: So **Relax** and solve

$$x^* = \arg\max_x X^T A X, X \in \mathbb{R}$$

X is the eigenvector corresponding to the largest eigenvalue.
It is easy to show that A is p.s.d.
x^* is discretized into x_d.

References

- A spectral technique for correspondence problems using pairwise constraints [Leordeanu-Hebert ICCV2005]
- Balanced Graph Matching. Cour-Srinivasan-Shi [NIPS2006]

A spectral clustering interpretation

1. Given a set of m_1 and m_2 points in $\mathbb{R}^n$ we constructed $m_1 \cdot m_2$ pairings $\{C_{i\hat{i}}\}$.
2. We computed the $(m_1 \cdot m_2) \times (m_1 \cdot m_2)$ affinity matrix and the **non**-normalized cut.
3. We assume that true assignments form a well connected component with in the graph.

The importance of cross-set similarity measures

The dimensions of the assignment affinity matrix A grow quadratically with the number of points. To align two 500 strong sets, we get $A_{500^2 \times 500^2}$.
But, in many applications we are given a cross set similarity measure: $d\left(x_i, \widehat{x_i}\right)$, and the number of possible assignment can be reduced.

In images analysis: local descriptors such as SIFT[Lowe2003], MSCR[Matas2002].

Results
Assignment score

$$S = X_d^T A X_d.$$

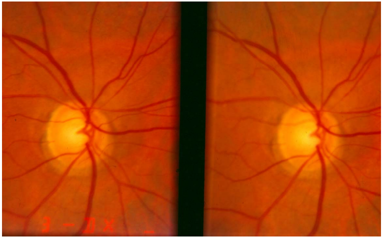
We applied this approach to speech recognition (audio only) in R^{10}.

audio	87%
visual	86%
both	95.1
RMM (audio only)	93%
spectral (audio only)	96.3%

Recall that the L_2 distances are diffusion distances.

Work in progress: automatic eye inspection

Joint work with Sina Farsiu and Mohammed El Mallah (Duke).



Certain eyes diseases manifest themselves as geometric deformations of blood vessels over years (5-10 years).

Future work

- Set spaces
- Shape embedding
- Shape recognition
- Automatic translation
- Non-rigid registration and tracking

Thanks You!