High-Dimesional Spectral Feature Selection for 3D Object Recognition Based on Reeb Graphs

Panel Talk

S+SSPR August 18 -20 2010
Cesme,TURKEY

,
Boyan Bonev &
Francisco Escolano UA (Spain)
Daniela Giorgi &
Silvia Biasotti IMATI (Italy)

High-Dimensional Feature Selection

Background. Feature selection in CVPR is mainly motivated by: (i) dimensionality reduction for improving the performance of classifiers, and (ii) feature-subset pursuit for a better understanding/description of the patterns at hand (either using generative or discriminative methods).

Exploring the domain of images [Zhu et al.,97] and other patterns (e.g. micro-array/gene expression data) [Wang and Gotoh,09] implies dealing with thousands of features [Guyon and Elisseeff, 03]. In terms of Information Theory (IT) this task demands pursuing the most informative features, not only individually but also capturing their statistical interactions, beyond taking into account pairs of features. This is a big challenge since the early 70's due to its intrinsic combinatorial nature.

Wrappers and Filters

Wrappers vs Filters

1. *Wrapper feature selection* (WFS) consists in selecting features according to the classification results that these features yield (e.g. using Cross Validation (CV)). Therefore wrapper feature selection is a classifier-dependent approach.
2. *Filter feature selection* (FFS)is classifier independent, as it is based on statistical analysis on the input variables (features), given the classification labels of the samples.
3. Wrappers build classifiers each time a feature set has to be evaluated. This makes them more prone to overfitting than filters.
4. In FFS the classifier itself is built and tested after FS.

FS is a Complex Task

The Complexity of FS

The only way to assure that a feature set is optimum (following what criterion?) is the exhaustive search among feature combinations. The *curse of dimensionality* limits this search, as the complexity is

$$O(z), \quad z = \sum_{i=1}^{n} \binom{n}{i}$$

Filters design is desirable in order to avoid overfitting, but they must be as multivariate as Wrappers. However, this implies to design and estimate a cost function capturing the high-order interaction of many variables. In the following we will focus on FFS.

Mutual Information Criterion

A Criterion is Needed

The primary problem of feature selection is the criterion which evaluates a feature set. It must decide whether a feature subset is suitable for the classification problem, or not. The optimal criterion for such purpose would be the Bayesian error rate for the subset of selected features:

$$E(\vec{S}) = \int_{\vec{S}} p(\vec{S}) \left(1 - \max_i(p(c_i|\vec{S}))\right) d\vec{S},$$

where $\vec{S}$ is the vector of selected features and $c_i \in C$ is a class from all the possible classes C existing in the data.

Mutual Information Criterion (2)

Bounding Bayes Error

An upper bound of the Bayesian error, which was obtained by Hellman and Raviv (1970) is:

$$E(\vec{S}) \leq \frac{H(C|\vec{S})}{2}.$$

This bound is related to mutual information (MI), because mutual information can be expressed as

$$I(\vec{S}; C) = H(C) - H(C|\vec{S})$$

and $H(\vec{C})$ is the entropy of the class labels which do not depend on the feature subspace $\vec{S}$.

Mutual Information Criterion (3)

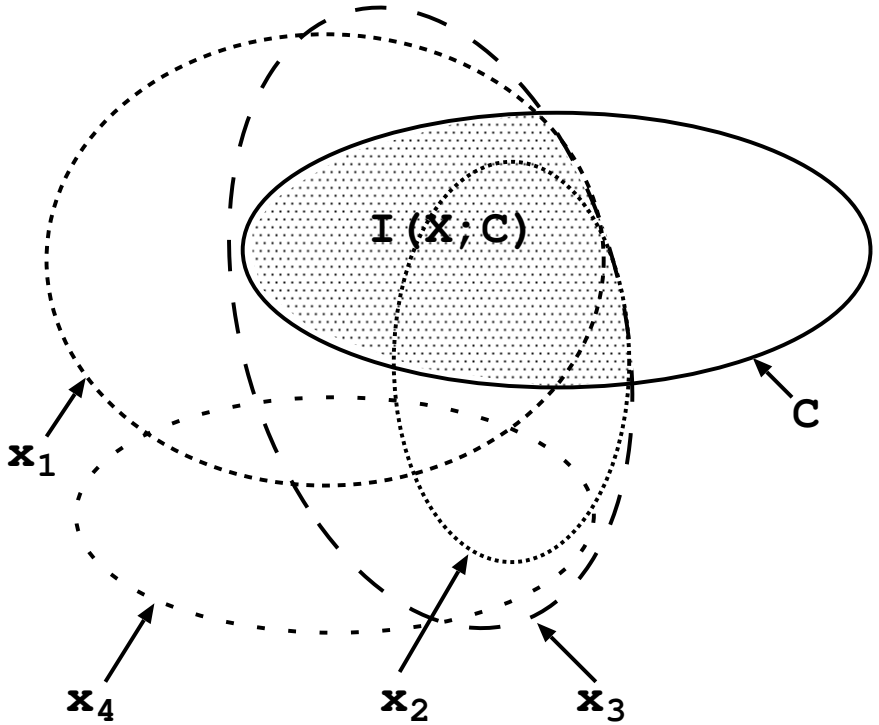

Figure: Redundancy and MI

Mutual Information Criterion (4)

MI and KL Divergence

From the definition of MI we have that:

$$\begin{aligned} I(X;Y) &= \sum_{y\in Y}\sum_{x\in X} p(x,y)\log\frac{p(x,y)}{p(x)p(y)} \\ &= \sum_{y\in Y} p(y)\sum_{x\in X} p(x|y)\log\frac{p(x|y)}{p(x)} \\ &= E_Y(KL\left(p(x|y)||p(x)\right)). \end{aligned}$$

Then, maximizing MI is equivalent to maximizing the expectation of the KL divergence between the class-conditional densities $P(\vec{S}|\vec{C})$ and the density of the feature subset $P(\vec{S})$.

Mutual Information Criterion (5)

The Bottleneck

- Estimating mutual information between a high-dimensional continuous set of features, and the class labels, is not straightforward, due to the entropy estimation.
- Simplifying assumptions: (i) *Dependency is not-informative about the class*, (ii) *redundancies between features are independent of the class* [Vasconcelos & Vasconcelos, 04]. The basic idea is to simplify the computation of the MI. For instance:

$$I(\vec{S}^*; C) \approx \sum_{x_i \in \vec{S}} I(x_i; C)$$

- More realistic assumptions involving interactions are needed!

The mRMR Criterion

Definition

Maximize the relevance $I(x_j; C)$ of each individual feature $x_j \in \vec{F}$ and simultaneously minimize the redundancy between x_j and the rest of selected features $x_i \in \vec{S}, i \neq j$. This criterion is known as the min-Redundancy Max-Relevance (mRMR) criterion and its formulation for the selection of the m-th feature is [Peng et al., 2005]:

$$\max_{x_j \in \vec{F} - \vec{S}_{m-1}} \left[I(x_j; C) - \frac{1}{m-1} \sum_{x_i \in \vec{S}_{m-1}} I(x_j; x_i) \right]$$

Thus, second-order interactions are considered!

The mRMR Criterion (2)

Properties

- Can be embedded in a forward greedy search, that is, at each iteration of the algorithm the next feature optimizing the criterion is selected.
- Doing so, this criterion is equivalent to a first-order using the Maximum Depedency criterion:

$$\max_{\vec{S} \subseteq \vec{F}} I(\vec{S}; C)$$

Then, the m-th feature is selected according to:

$$\max_{x_j \in \vec{F} - \vec{S}_{m-1}} I(\vec{S}_{m-1}, x_j; C)$$

Efficient MD

Need of an Entropy Estimator

- Whilst in mRMR the mutual information is incrementally estimated by estimating it between two variables of one dimension, in MD the estimation of $I(\vec{S}; C)$ is not trivial because $\vec{S}$ could consist of a large number of features.
- If we base MI calculation in terms of the conditional entropy $H(\vec{S}|\vec{C})$ we have:

$$I(\vec{S}; \vec{C}) = H(\vec{S}) - H(\vec{S}|\vec{C}).$$

To do this, $\sum H(\vec{S}|C = c)p(C = c)$ entropies have to be calculated, Anyway, we must calculate multi-dimensional entropies!

Leonenko's Estimator

Theoretical Bases

Recently Leonenko et al. published an extensive study [Leonenko et al, 08] about Rényi and Tsallis entropy estimation, also considering the case of the limit of $\alpha \to 1$ for obtaining the Shannon entropy. Their construction relies on the integral

$$I_\alpha = E\{f^{\alpha-1}(\vec{X})\} = \int_{\mathbb{R}^d} f^\alpha(x)dx,$$

where $f(.)$ refers to the density of a set of n i.i.d. samples $\vec{X} = \{X_1, X_2, \ldots, X_N\}$. The latter integral is valid for $\alpha \neq 1$, however, the limits for $\alpha \to 1$ are also calculated in order to consider the Shannon entropy estimation.

Leonenko's Estimator (2)

Entropy Maximization Distributions

- The α-entropy maximizing distributions are only defined for $0 < \alpha < 1$, where the entropy H_α is a concave function. The maximizing distributions are defined under some constraints. The uniform distribution maximizes α-entropy under the constraint that the distribution has a finite support. For distributions with a given covariance matrix the maximizing distribution is Student-t, if $d/(d+2) < \alpha < 1$, for any number of dimensions $d \geq 1$.
- This is a generalization of the property that the Gaussian distribution maximizes the Shannon entropy H.

Leonenko's Estimator (3)

Rényi, Tsallis and Shannon Entropy Estimates

For $\alpha \neq 1$, the estimated I is:

$$\hat{I}_{N,k,\alpha} = \frac{1}{N} \sum_{i=1}^{n} (\zeta_{N,i,k})^{1-\alpha},$$

with

$$\zeta_{N,i,k} = (N-1) C_k V_d (\rho_{k,N-1}^{(i)})^d,$$

where $\rho_{k,N-1}^{(i)}$ is the Euclidean distance from X_i to its k-th nearest neighbour from among the resting $N-1$ samples.
$V_d = \pi^{\frac{d}{2}} / \Gamma(\frac{d}{2}+1)$ is the volume of the unit ball $\mathcal{B}(0,1)$ in $\mathbb{R}^d$ and C_k is $C_k = [\Gamma(k)/\Gamma(k+1-\alpha)]^{\frac{1}{1-\alpha}}$.

Leonenko's Estimator (4)

Rényi, Tsallis and Shannon Entropy Estimates (cont.)

- The estimator $\hat{I}_{N,k,\alpha}$ is asymptotically unbiased, which is to say that $E\hat{I}_{N,k,\alpha} \to I_q$ as $N \to \infty$. It is also consistent under mild conditions.
- Given these conditions, the estimated Rényi entropy H_α of f is
$$\hat{H}_{N,k,\alpha} = \frac{\log(\hat{I}_{N,k,\alpha})}{1-\alpha}, \quad \alpha \neq 1,$$
and for the Tsallis entropy $S_\alpha = \frac{1}{q-1}(1 - \int_x f^\alpha(x)dx)$ is
$$\hat{S}_{N,k,\alpha} = \frac{1-\hat{I}_{N,k,\alpha}}{\alpha-1}, \quad \alpha \neq 1.$$
and as $N \to \infty$, both estimators are asymptotically unbiased and consistent.

Leonenko's Estimator (5)

Shannon Entropy Estimate

The limit of the Tsallis entropy estimator as $\alpha \to 1$ gives the Shannon entropy H estimator:

$$\hat{H}_{N,k,1} = \frac{1}{N} \sum_{i=1}^{N} \log \xi_{N,i,k},$$

$$\xi_{N,i,k} = (N-1) e^{-\Psi(k)} V_d(\rho_{k,N-1}^{(i)})^d,$$

where $\Psi(k)$ is the digamma function:
$\Psi(1) = -\gamma \simeq 0.5772, \Psi(k) = -\gamma + A_{k-1}, \;\; A_0 = 0, A_j = \sum_{i=1}^{j} \frac{1}{i}.$

Leonenko's Estimator (6)

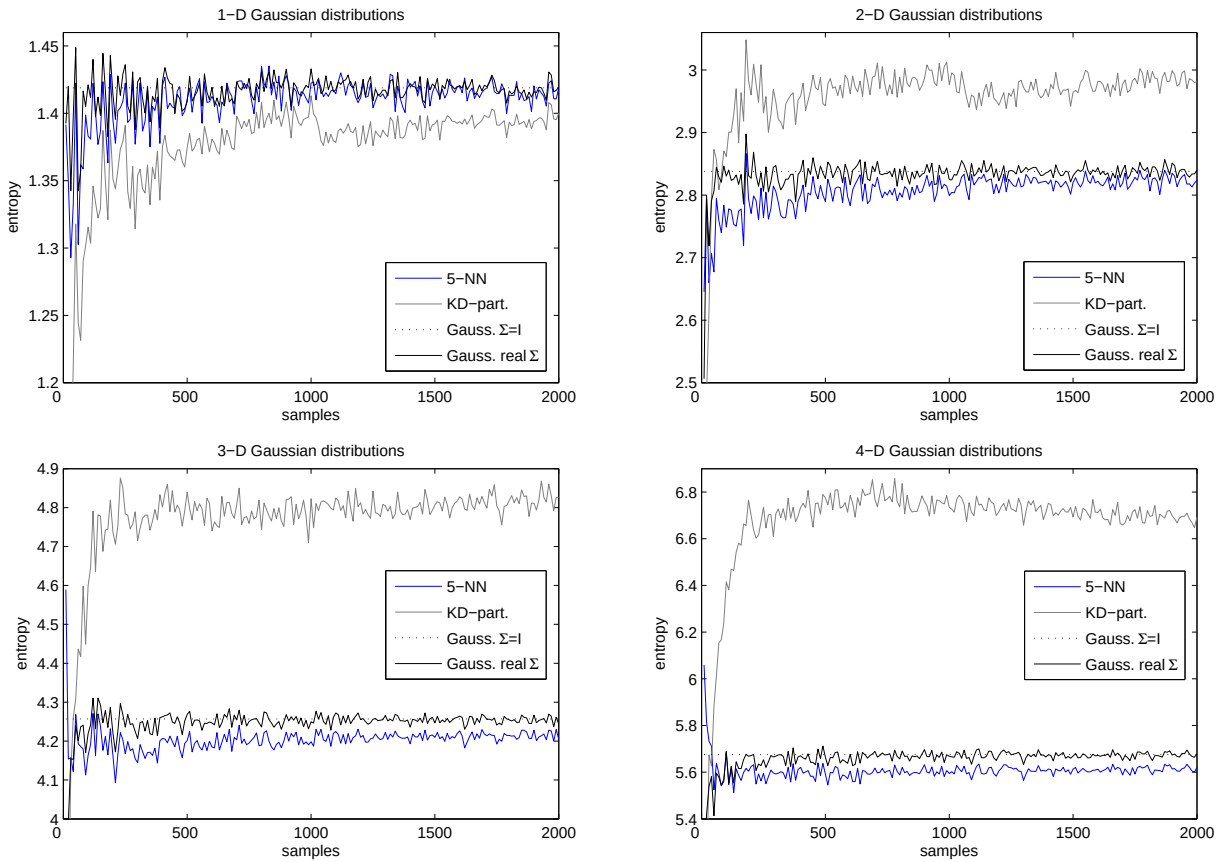

Figure: Comparison: k-NN vs k-d partitioning estimator for Gaussians.

Structural Recognition

Reeb Graphs

- Given a surface $\mathcal{S}$ and a real function $f : \mathcal{S} \to \mathbb{R}$, the *Reeb graph* (RG), represents the topology of $\mathcal{S}$ through a graph structure whose nodes correspond to the critical points of f.
- The *Extended Reeb Graph* (ERG) [Biasotti, 05] is an approximation of the RG by using of a fixed number of level sets (63 in this work) that divide the surface into a set of regions; critical regions, rather than critical points, are identified according to the behaviour of f along level sets; ERG nodes correspond to critical regions, while the arcs are detected by tracking the evolution of level sets.

Structural Recognition (2)

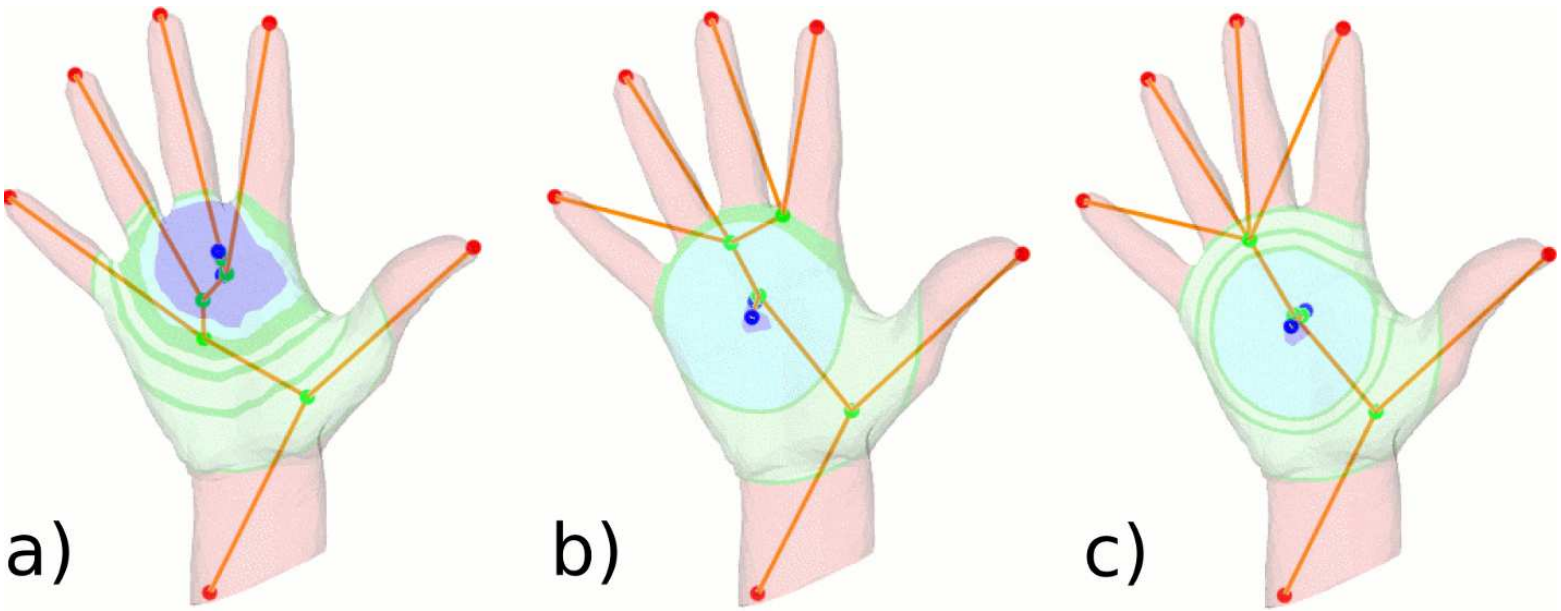

Figure: Extended Reeb Graphs. Image by courtesy of Daniela Giorgi and Silvia Biasotti. a) Geodesic, b) Distance to center, b) bSphere

Structural Recognition (3)

Catalog of Features

- Subgraph node centrality: quantifies the degree of participation of a node i in a subgraph.

$$C_S(i) = \sum_{k=1}^{n} \phi_k(i)^2 e^{\lambda_k},$$

where where $n = |V|$, λ_k the k-th eigenvalue of $\mathbf{A}$ and ϕ_k its corresponding eigenvector.

- Perron-Frobenius eigenvector: ϕ_n (the eigenvector corresponding to the largest eigenvalue of $\mathbf{A}$). The components of this vector denote the importance of each node.
- Adjacency Spectra: the magnitudes $|\phi_k|$ of the (leading) eigenvalues of $\mathbf{A}$ have been been experimentally validated for graph embedding [Luo et al.,03].

Structural Recognition (4)

Catalog of Features (cont.)

- Spectra from Laplacians: both from the un-normalized and normalized ones. The Laplacian spectrum plays a fundamental role in the development of **regularization kernels** for graphs.
- Friedler vector: eigenvector corresponding to the first non-trivial eigenvalue of the Laplacian (ϕ_2 in connected graphs). It encodes the connectivity structure of the graph (actually is the core of graph-cut methods).
- Commute Times either coming from the un-normalized and normalized Laplacian. They encode the path-length distribution of the graph.
- The Heat-Flow Complexity trace as seen in the section above.

Structural Recognition (5)

Backwards Filtering

- The entropies $H(\cdot)$ of a set with a large n number of features can be efficiently estimated using the k-NN-based method developed by Leonenko.
- Thus, we take the data set with all its features and determine which feature to discard in order to produce the smallest decrease of $I(\vec{S_{n-1}}; \vec{C})$.
- We then repeat the process for the features of the remaining feature set, until only one feature is left [Bonev et al.,10].
- Then, the subset yielding the minimum 10-CV error is selected.
- Most of the spectral features are histogrammed into a variable number of bins: $9 \cdot 3 \cdot (2 + 4 + 6 + 8) = 540$ features.

Experimental Results

Elements

- SHREC database (15 classes $\times$ 20 objects). Each of the 300 samples is characterized by 540 features, and has a class label $l \in \{$*human*, *cup*, *glasses*, *airplane*, *chair*, *octopus*, *table*, *hand*, *fish*, *bird*, *spring*, *armadillo*, *buste*, *mechanic*, *four-leg*$\}$
- The errors are measured by 10-fold cross validation (10-fold CV). MI is maximized as the number of selected features grows. A high number of features degrades the classification performance.
- However the MI curve, as well as the selected features, do not depend on the classifier, as it is a purely information-theoretic measure.

Structural Recognition (7)

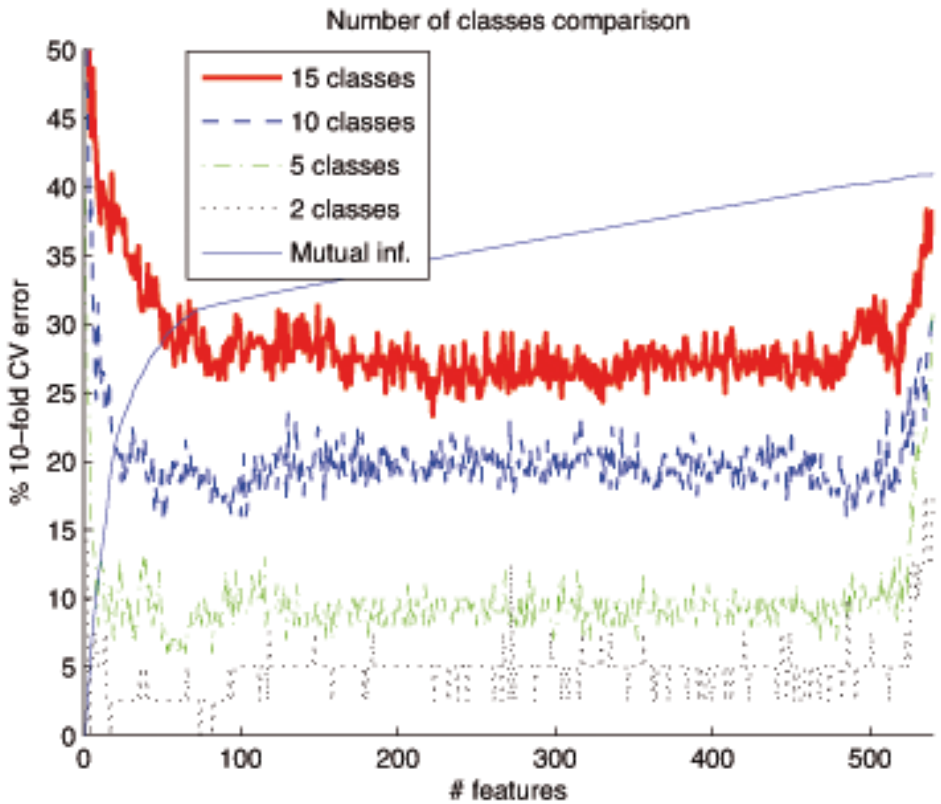

Figure: Classification error vs Mutual Information.

Structural Recognition (8)

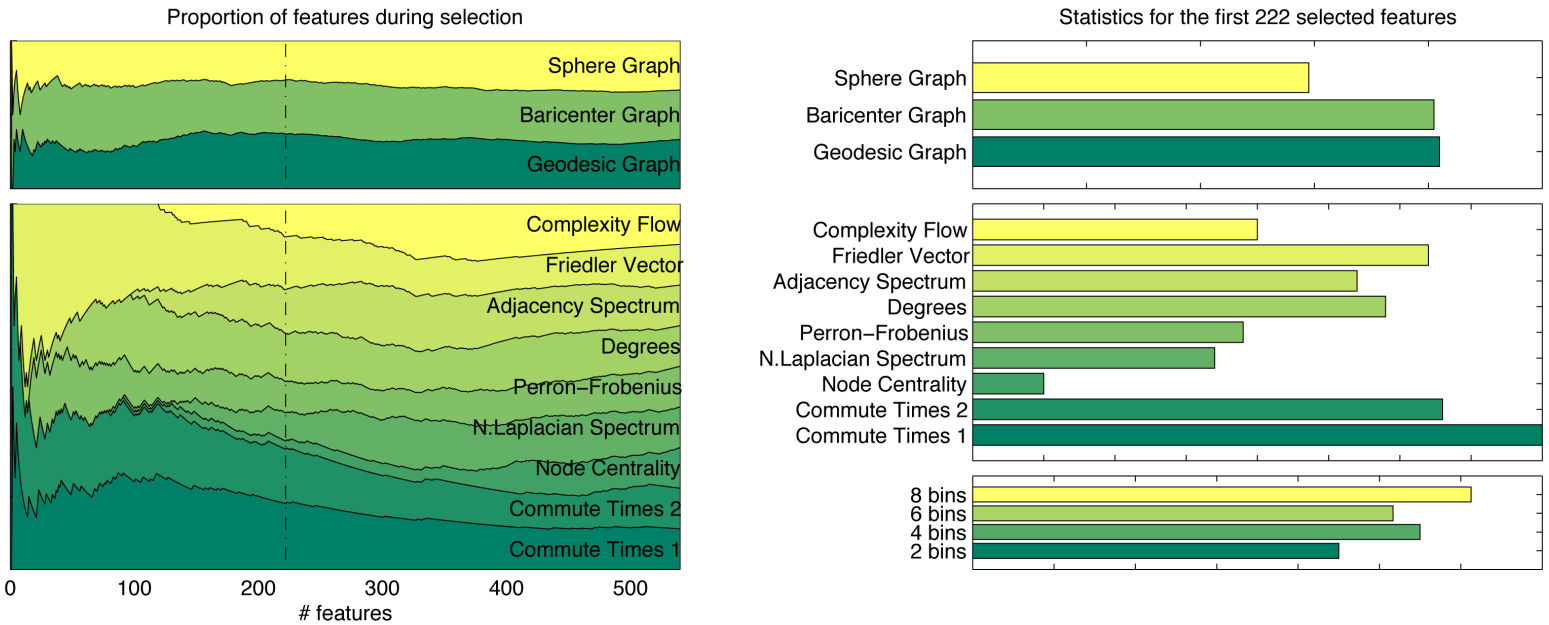

Figure: Feature selection on the 15-class experiment (left) and the feature statistics for the best-error feature set (right).

Structural Recognition (9)

Analysis

- For the 15-class problem, the minimal error $(23,3\%)$ is achieved with a set of 222 features.
- The coloured areas in the plot represent how much a feature is used with respect to the remaining ones (the height on the Y axis is arbitrary).
- For the 15-class experiment, in the feature sets smaller than 100 features, the most important is the Friedler vector, in combination with the remaining features. Commute time is also an important feature.Some features that are not relevant are the node centrality and the complexity flow. Turning our attention to the graphs type, all three appear relevant.

Structural Recognition (10)

Analysis (cont)

- We can see that the four different binnings of the features do have importance for graph characterization.
- These conclusions concerning the relevance of each feature cannot be drawn without performing some additional experiments with different groups of graph classes.
- We perform our different 3-class experiments. The classes share some structural similarities, for example the 3 classes of the first experiment have a head and limbs.
- Although in each experiment the minimum error is achieved with very different numbers of features, the participation of each feature is highly consistent with the 15-class experiment.

Structural Recognition (11)

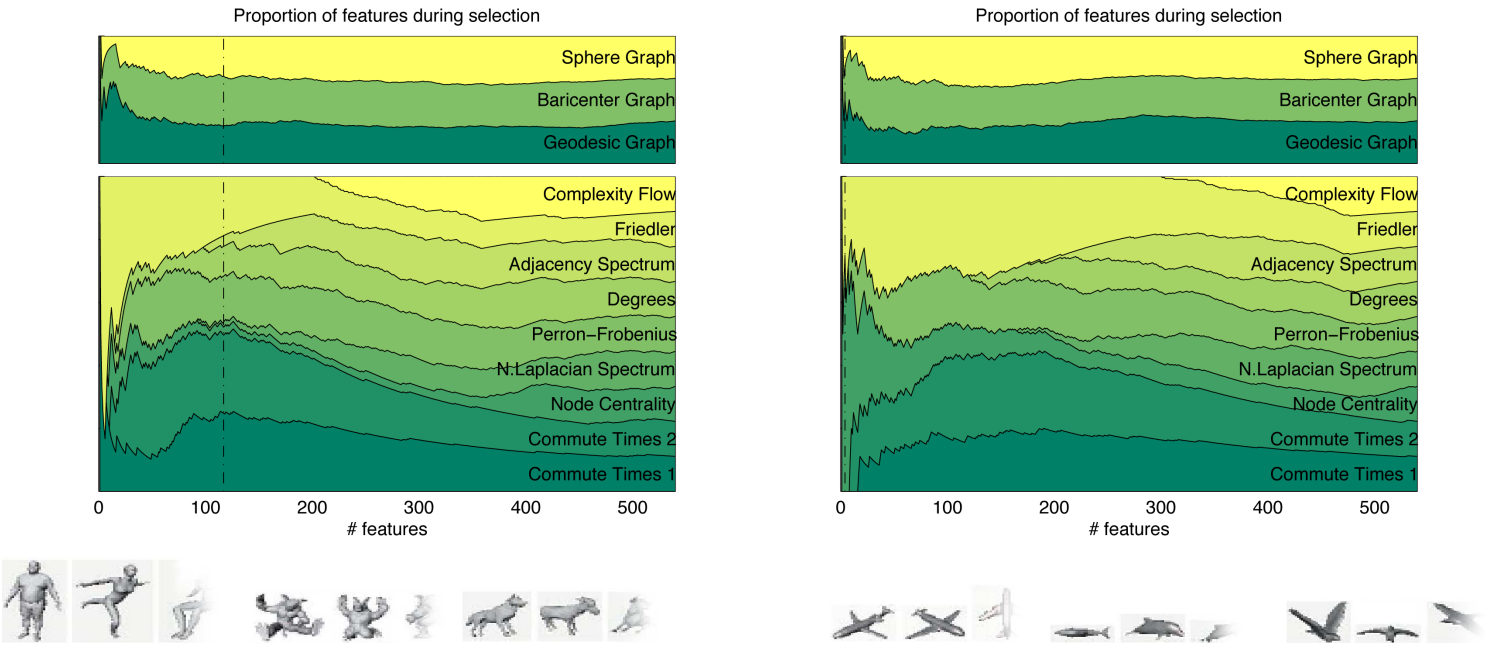

Figure: Feature Selection on 3-class experiments: Human/Armadillo/Four-legged, Aircraft/Fish/Bird

Structural Recognition (12)

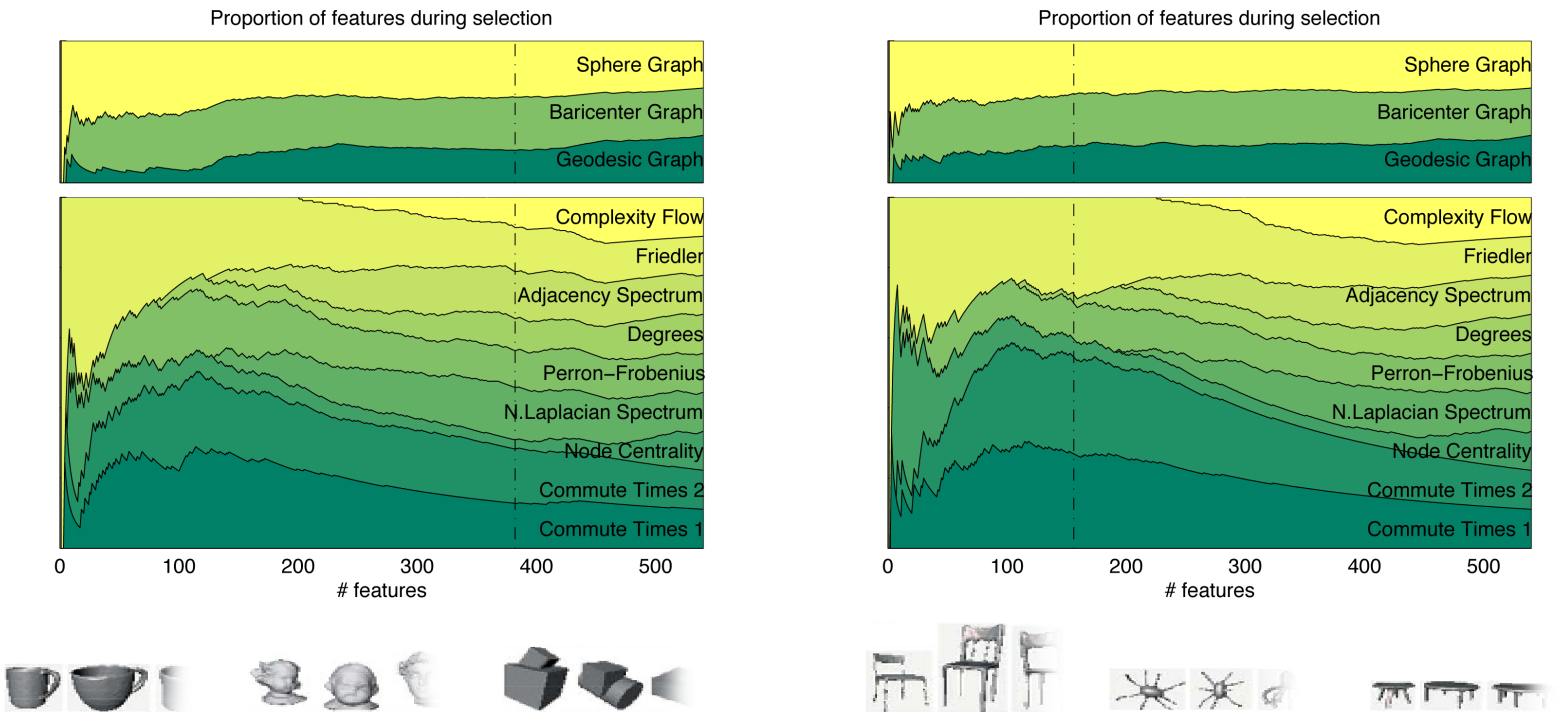

Figure: Feature Selection on 3-class experiments: Cup/Bust/Mechanic, Chair/Octopus/Table.

Structural Recognition (14)

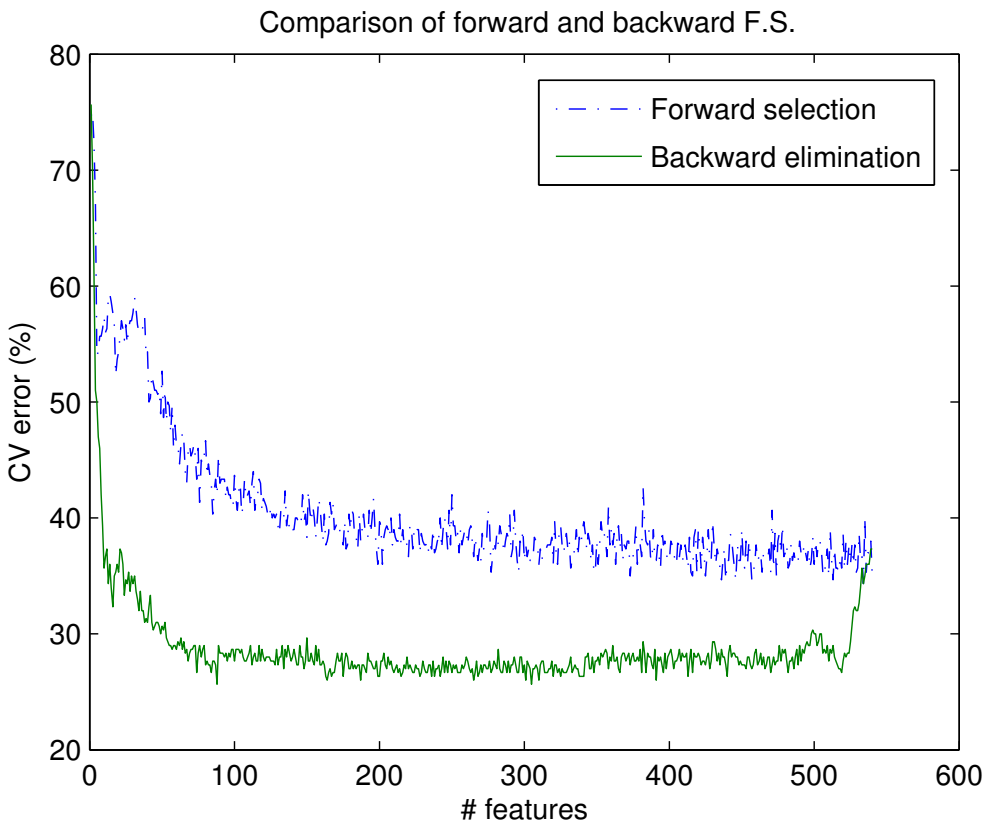

Figure: Forward vs Backwards FS

Structural Recognition (15)

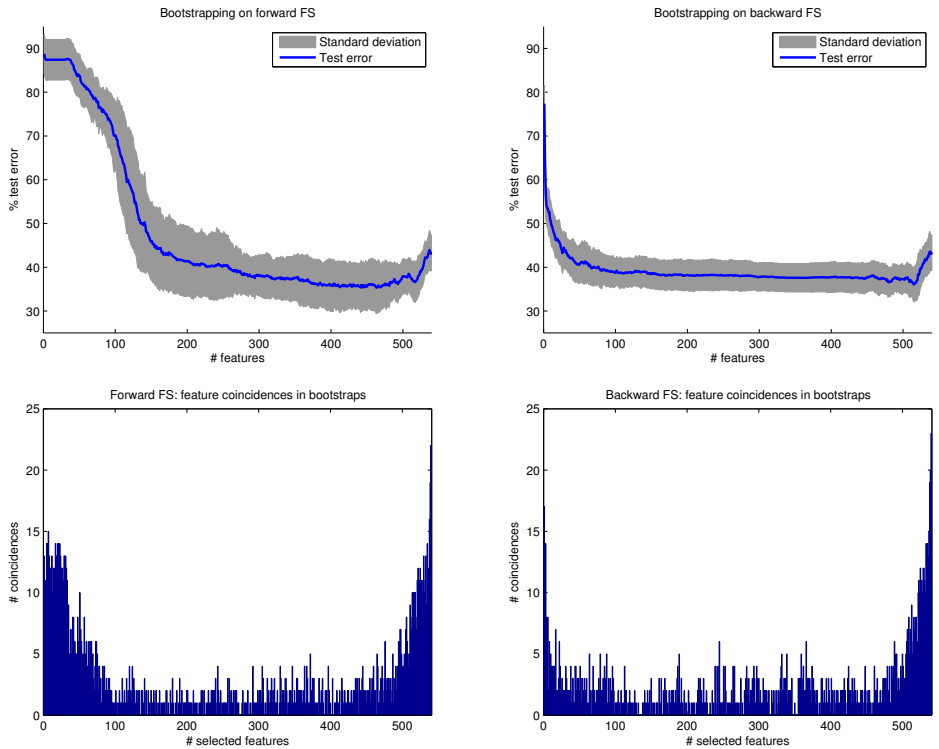

Figure: Bootstrapping experiments on forward (on the left) and backward (on the right) feature selection, with 25 bootstraps.

Structural Recognition (16)

Conclusions

- The main difference among experiments is that node centrality seems to be more important for discerning among elongated sharp objects with long sharp characteristics. Although all three graph types are relevant, the sphere graph performs best for blob-shaped objects.
- Bootstrapping shows how better is the Backwards method.
- The IT approach not only performs good classification but also yield the role of each spectral feature in it!
- Given other points of view (size functions, eigenfunctions,...) it is desirable to exploit them!

Questions?

Challenges

- It is straightforward to deal with attributed graphs complementing structured, and how is it introduced in the algorithm? (is modifying Laplacian enough?)
- What is the limit of the bypass entropy estimator and, consequently of the number of features we can extract?
- What happens when it is needed to measure Mutual Information between multidimensional variables, not in classification but in regression problems?
- Does the method provides an insight of how to incorporate other graph descriptors?

References

- [Zhu et al.,97] Zhu, S.C., Wu, Y.N. and Mumford, D (1997). Filters, Random Fields And Maximum Entropy (FRAME): towads an unified theory for texture modeling. IJCV 27 (2) 107–126
- [Wang and Gotoh, 2009] Wang, X. and Gotoh, O. (2009). Accurate molecular classication of cancer using simple rules. BMC Medical Genomics, 64(2)
- [Guyon and Elisseeff, 03] Guyon, I. and Elissee?, A. (2003). An introduction to variable and feature selection. Journal of Machine Learning Research, volume 3, pages 1157–1182.
- [Vasconcelos and Vasconcelos,04] Vasconcelos, N. and Vasconcelos, M. (2004). Scalable discriminant feature selection for image retrieval and recognition. In CVPR04, pages 770–775.

References (2)

- [Peng et al., 2005]Peng, H., Long, F., and Ding, C. (2005). Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy. In IEEE Trans. on PAMI, 27(8), pp.1226-1238.
- [Leonenko et al, 08] Leonenko, N., Pronzato, L., and Savani, V. (2008). A class of Rényi information estimators for multidimensional densities. The Annals of Statistics, 36(5):21532182.
- [Lozano et al, 09] Lozano M,A., Escolano, F., Bonev, B., Suau, P., Aguilar, W., Sáez, J.M., Cazorla, M. (2009). Region and constell. based categorization of images with unsupervised graph learning. Image Vision Comput. 27(7): 960–978

References (3)

- [Bonev et al.,08] Bonev, B., Escolano, F., and Cazorla, M. (2008). Feature selection, mutual information, and the classication of high-dimensional patterns. Pattern Analysis and Applications 11(3-4): 304–319.
- [Biasotti, 05] Biasotti, S. (2005). Topological coding of surfaces with boundary using Reeb graphs. Computer Graphics and Geometry, 7(1):3145.
- [Luo et al.,03 Luo, B., Wilson, R., and Hancock, E. (2003). Spectral embedding of graphs. Pattern Recognition, 36(10):22132223
- [Bonev et al.,10] Bonev, B., Escolano, F., Giorgi, D., Biasotti, S, (2010). High-dimensional Spectral Feature Selection for 3D Object Recognition based on Reeb Graphs, SSPR'2010 (accepted)