Scalable Discriminant Feature Selection for Image Retrieval and Recognition

Nuno Vasconcelos

ECE Department

University of California, San Diego

Manuela Vasconcelos

Division of Engineering and Applied Sciences

Harvard University

Introduction

- vision defines large scale-classification problems
  - large # of classes, large amounts of data per class
- discriminant feature space is a pre-requisite for success
- features are usually chosen according to intuitive, but not provably optimal/discriminant, justifications:
  - biological plausibility: Gabor, wavelet, multiresolution
  - optimality under non-classification criteria: PCA, ICA
  - perceptual relevance: edginess, color, etc.
- classification-optimal methods (search, boosting, etc)
  - do not scale well in the # of classes
  - little insight on what are the constraints for “good features”
  - large training complexity

Goals

- practical: classification-optimal FS algorithms that scale
- theoretical: the roles of discrimination and dependence
  - discriminant feature is a great asset
  - 2nd highly discriminant that does not add much info about class label (e.g. equal to 1st) is highly undesirable
  - good features balance max discrimination with min dependence
- this trade-off is not well understood
  - some solutions disregard dependencies (e.g. naïve Bayes, FS based on marginal distributions)
  - others disregard discrimination (e.g. ICA, PCA, variance-based FS methods)
  - many are “black box” solutions (e.g. boosting, forward search, …)

Optimal discrimination/dependence trade-off

- naturally formalized by information theory
  - well known relationships between independence and information
  - not-so-well known between information and discrimination
- given feature space $\mathcal{X}$ and set $Y = \{1, \ldots, M\}$ of classes, classifier is map $g^* : \mathcal{X} \rightarrow Y$ such

$$g^* = \arg\min_g P(g(\mathbf{x}) \neq y), \forall \mathbf{x}, y.$$

error lower bounded by Bayes error (BE)

$$L^* = 1 - E_{\mathbf{x}}[\max_i P(y = i|\mathbf{x})]$$

- BE depends only on the feature space, not classifier
- feature selection as the search for the BE-optimal space

Infomax principle (Linsker, Kullback)

- **classification:** M-ary problem with observations $\mathbf{Z} \in \mathcal{Z}$, best feature transformation is

$$T^* = \arg\max_T I(Y; \mathbf{X})$$

  where

$$I(Y; \mathbf{X}) = \sum_i \int p_{\mathbf{X},Y}(\mathbf{x}, i) \log \frac{p_{\mathbf{X},Y}(\mathbf{x}, i)}{p_{\mathbf{X}}(\mathbf{x}) p_Y(i)} d\mathbf{x}$$

  is the mutual information between $\mathbf{X} = T(\mathbf{Z})$ and the class label Y.

- since $I(\mathbf{X}; Y) = H(Y) - H(Y|\mathbf{X})$, this is the same as **minimizing the class-posterior entropy** (CPE)

$$T^* = \arg\min_T H(Y|\mathbf{X})$$

Properties of Infomax (NIPS'02, CVPR'03)

- **discriminant:** letting $\langle f(i) \rangle_Y = \sum_i P_Y(i) f(i)$,

$$T^* = \arg\max_T \left\langle KL \left[P_{\mathbf{X}|Y}(\mathbf{x}|i) || P_{\mathbf{X}}(\mathbf{x}) \right] \right\rangle_Y$$

  where $KL[p||q] = \int p(\mathbf{x}) \log[p(\mathbf{x})/q(\mathbf{x})] d\mathbf{x}$.

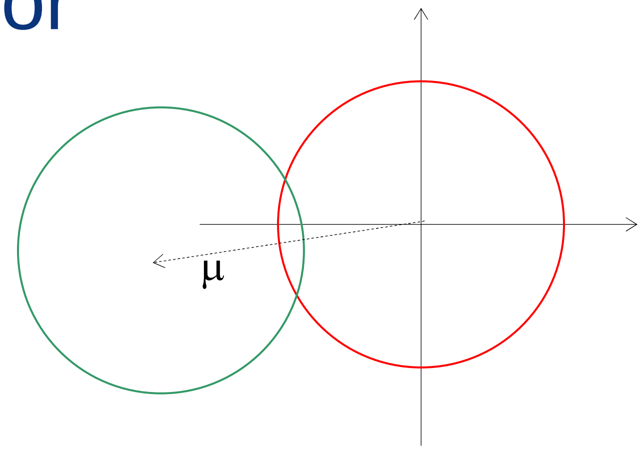
- it is possible to establish connection to Bayes error
- **Theorem:** for and M-class problem and feature space $\mathcal{X}$

$$L^*_{\mathcal{X}} \geq \frac{1}{\log M} H(Y|\mathbf{X}) - \log \frac{2M-1}{\log M} + 1$$

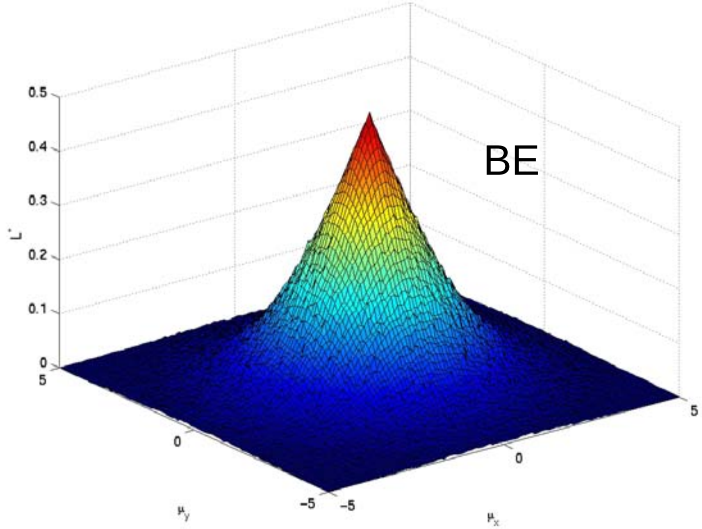
- Infomax minimizes a lower bound on BE!
- bound is tight for most problems of interest

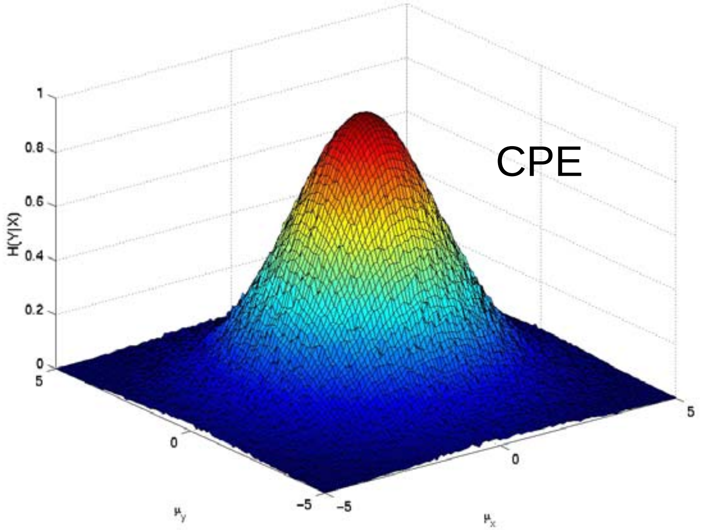
Infomax vs Bayes error

- example:
  M=2,
  $X_{|y=1} \sim N(0,I)$,
  $X_{|y=2} \sim N(\mu,I)$



- *BE and CPE as functions of μ*





- Infomax: natural formalism to analyze trade-off between discriminantion and dependencies

Discrimination vs independence

- if $\mathcal{Z}$ is n-dimensional and $\mathbf{X}^* = (X_1^*, \ldots, X_N^*)$ the optimal feature subset of size N, then

$$I(\mathbf{X}^*, Y) = \underbrace{\sum_{k=1}^{N} I(X_k^*, Y)}_{A}$$

$$- \sum_{k=2}^{N} \underbrace{[I(X_k^*; \mathbf{X}_{1,k-1}^*) - I(X_k^*; \mathbf{X}_{1,k-1}^* | Y)]}_{B}.$$

where $\mathbf{X}_{1,k-1}^* = \{X_1^*, \ldots, X_{k-1}^*\}$.

- A measures individual discriminant power of each feature B penalizes combinations that are highly informative of class label (zero when X_k and $X^*_{1,k-1}$ jointly indep of Y)

Interesting corollary

- if

$$\frac{1}{N-1}\sum_{k=2}^{N} I(X_k^*;\mathbf{X}_{1,k-1}^*) = \frac{1}{N-1}\sum_{k=2}^{N} I(X_k^*;\mathbf{X}_{1,k-1}^*|Y),$$

then

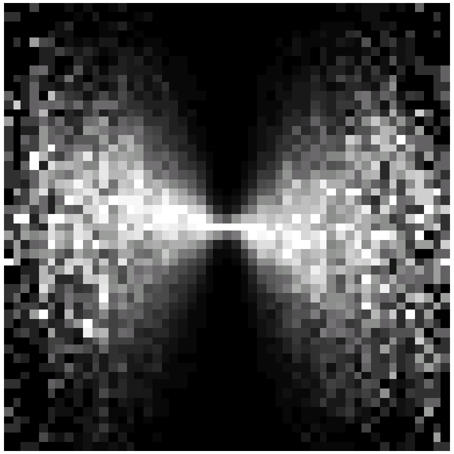
$$I(\mathbf{X}^*,Y) = \sum_{k=1}^{N} I(X_k^*,Y). \qquad (1)$$

i.e. all redundancy that does not carry information about class label can be ignored

- independent modeling of highly correlated features not necessarily sub-optimal!

Image statistics

- interesting condition: various studies reporting consistent patterns of dependence for features of biologically plausible transforms (Simoncelli et al, Mumford et al, etc.)



- although the fine details of dependence vary from class to class, the coarse structure of dependence patterns is similar for most image classes

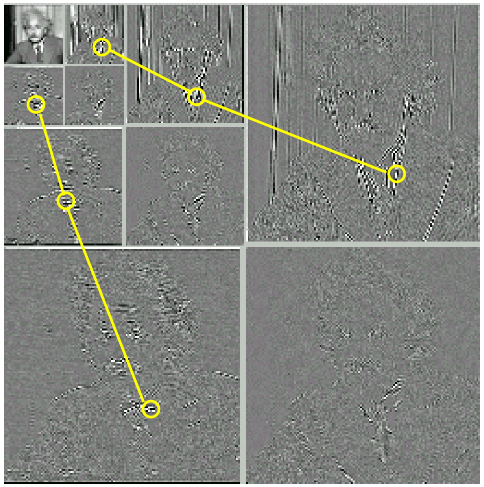
- conjecture: maximization of marginal diversity is close to optimal for visual recognition
- direct verification requires high-dimensional density estimates, problematic. We follow alternative path.

Measuring the impact of dependencies

- strategy: sequentially relax assumption that feature dependencies are not informative about class label
  - feature set grouped into exclusive subsets of l^{th} order
  - features within subsets arbitrarily dependent, no constraints
  - dependence between subsets not informative about image class
- extend (1) for each dependency order and obtain associated optimal algorithm
- interesting in two ways
  - by measuring error rate we can determine order at which dependencies do become non-informative
  - if this order is small we have an optimal FS algorithm of reduced complexity

Why should this work?

- while (1) may be too restrictive, assumption should hold for some order < full space dimension
- if the assumption of non-informative dependences holds at order *l*, we have *l-decomposability*



- e.g. dependencies between wavelet coefficients well known to be localized in both space and image scale
  - co-located coefficients of equal orientation can be arbitrarily dependent on the class
  - average dependence between such sets of coefficients does not depend on the image class (strong vertical frequencies ⇔ weak horizontal frequencies)
- even if it does not, resulting family of algorithms allows continuous trade-off between complexity and optimality

l-decomposability

- **Definition:** $\mathbf{X} = (X_1, \ldots, X_N)$ is l-decomposable iif there $\exists$ mutually exclusive subsets $\mathcal{C} = \{\mathbf{C}_1, \ldots, \mathbf{C}_{\lceil N/l \rceil}\}$

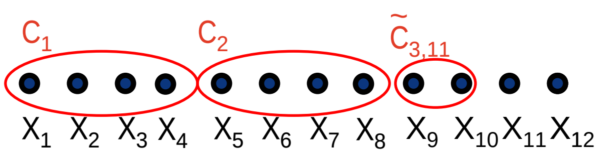
$$\mathbf{C}_i = \begin{cases} \{X_{(i-1)l+1}, \ldots, X_{il}\}, & \text{if } i < \lceil N/l \rceil, \\ \{X_{(i-1)l+1}, \ldots, X_N\}, & \text{if } i = \lceil N/l \rceil \end{cases}$$

and, for all $k \in \{2, \ldots, N\}$,

$$\sum_{i=1}^{\lceil k-1/l \rceil} \left[I(X_k; \tilde{\mathbf{C}}_{i,k} | \mathbf{C}_1, \ldots, \mathbf{C}_{i-1}) - I(X_k; \tilde{\mathbf{C}}_{i,k}) \right] = \sum_{i=1}^{\lceil k-1/l \rceil} \left[I(X_k; \tilde{\mathbf{C}}_{i,k} | \mathbf{C}_1, \ldots, \mathbf{C}_{i-1}, Y) - I(X_k; \tilde{\mathbf{C}}_{i,k} | Y) \right]$$

where $\tilde{\mathbf{C}}_{i,k} = \{\mathbf{x}_j | \mathbf{x}_j \in \mathbf{C}_i, j < k\}$.

- for example, when *N=12, l = 4, k=11*



l-decomposability

- from (A, B) jointly independent of $C \Leftrightarrow I(A,B|C) = I(A,B)$ it follows that

$$\frac{1}{\lceil k-1/l \rceil} \sum_{i=1}^{\lceil k-1/l \rceil} \left[I(X_k; \tilde{\mathbf{C}}_{i,k} | \mathbf{C}_1, \ldots, \mathbf{C}_{i-1}) - I(X_k; \tilde{\mathbf{C}}_{i,k}) \right]$$

  measures average redundancy between $\mathbf{C}_i$.

- **X** *l*-decomposable if this average redundancy is non-informative about the class label
- note that *l*-decomposability does not impose constraints on dependencies within the subsets C_i
- next we see that when arbitrary dependencies of order *l* are allowed, the optimal infomax solution only requires density estimates on subspaces of dimension *l+1*

Properties of l-decomposability

- **Theorem:** Let $\mathbf{X}^* = (X_1^*, \ldots, X_N^*)$ be the infomax-optimal set of size N. If $\mathbf{X}^*$ is l-decomposable into $\mathcal{C} = \{\mathbf{C}_1, \ldots, \mathbf{C}_{\lceil N/l \rceil}\}$ then

$$
\begin{aligned}
I(\mathbf{X}^*; Y) &= \sum_{k=1}^{N} I(\mathbf{X}_k^*; Y) && (1)\\
&- \sum_{k=2}^{N} \sum_{i=1}^{\lceil k-1/l \rceil} \left[I(X_k^*; \tilde{\mathbf{C}}_{i,k}) - I(X_k^*; \tilde{\mathbf{C}}_{i,k} | Y) \right]
\end{aligned}
$$

  where $\tilde{\mathbf{C}}_{i,k} = \{\mathbf{x}_j | \mathbf{x}_j \in \mathbf{C}_i, j < k\}$.

- this suggests a family of FS algorithms, parameterized by *l*, that trades optimality for complexity

A family of algorithms

- natural extension to traditional FS by sequential search
  - start from optimal set of cardinality 1
  - sequentially add feature that most increases the cost
- discriminant cost for selecting “next best” feature

$$C_r = I(X_r;Y) + \sum_{i=1}^{\lceil k-1/l \rceil} \left[I\left(X_r;C_{i,k} \mid Y\right) - I\left(X_r;C_{i,k}\right) \right]$$

  - O: favors features that are discriminant (large $I(X_r;Y)$)
  - O: penalizes features redundant with previously selected ($I(X_r; C_{i,k})$)
  - O: unless redundancy provides information about Y ($I(X_r; C_{i,k}|Y)$).

Feature selection algorithm

- **Algorithm 1** *Given a set of n features $\mathbf{X} = (X_1, \ldots, X_n)$, the order l, the target number of features N, and denoting the marginal diversity of X_k, $I(\mathbf{X}_k, Y)$, by md_k.*

  1. *set $\mathbf{X}^* = \mathbf{C}_1 = \{X_1^*\}$ where $X_1^* \in \mathbf{X}$ is the feature of largest marginal diversity, set $k = 2$, and $i = 1$.*

  2. *foreach $X_r \notin \mathbf{X}^*$, compute $\delta_r = \sum_{p=1}^{\lceil k-1/l \rceil} I(X_r; \tilde{\mathbf{C}}_{p,k}|Y) - I(X_r; \tilde{\mathbf{C}}_{p,k})$.*

  3. *let $r^* = \arg\max_r md_r + \delta_r$. If $k-1$ is not a multiple of l make $\mathbf{C}_i = \mathbf{C}_i \cup \mathbf{X}_{r^*}$. Else, set $i = i+1$, and let $\mathbf{C}_i = \mathbf{X}_{r^*}$. In both cases make $\mathbf{X}^* = \cup_i \mathbf{C}_i$, $k = k+1$, and go to 2 if $k < N$.*

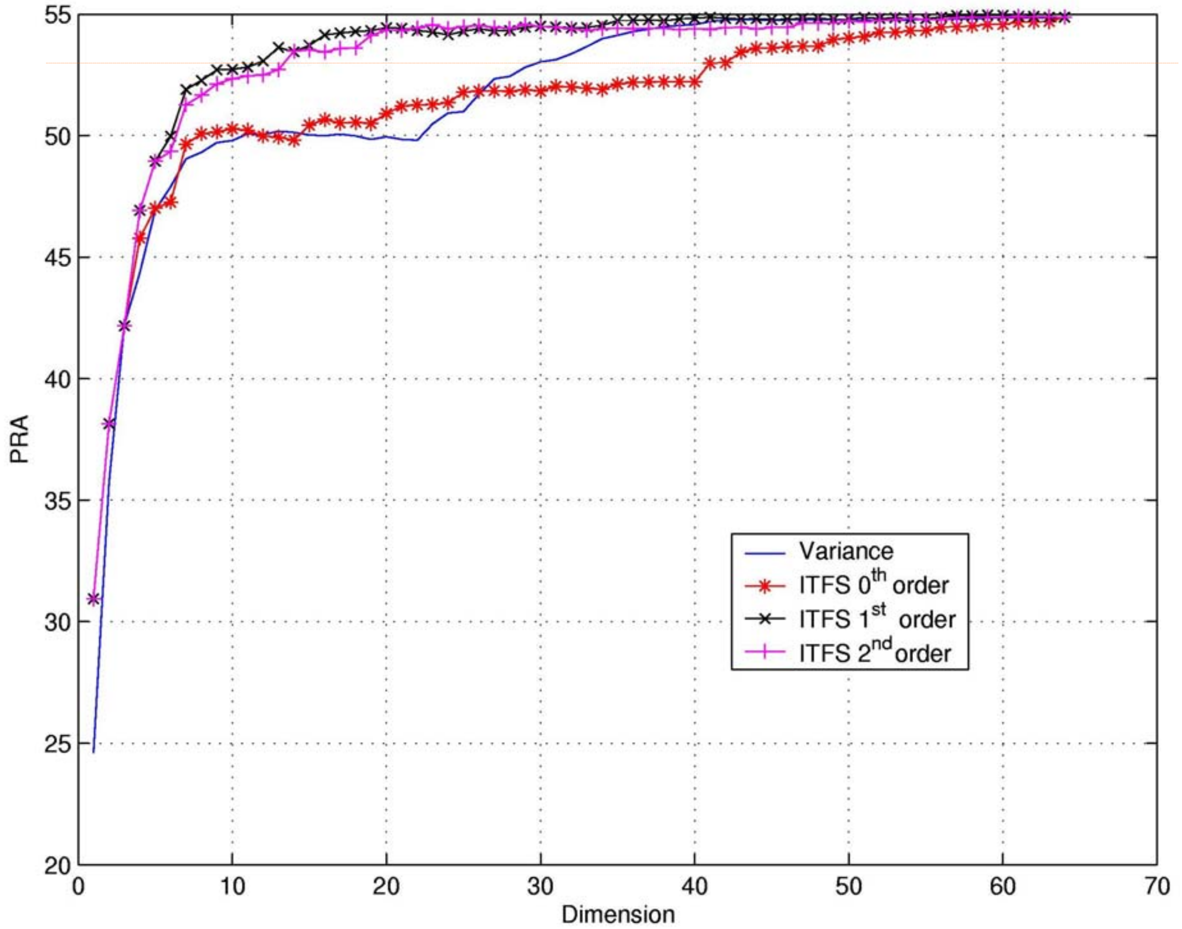
- what *l* is needed to capture all significant dependencies?

Experimental set-up

- Two databases
  - Brodatz: texture, 112 classes, 1008 images
  - Corel: natural images, 15 classes, 1500 images
- recognition: 20% testing, 80% training
  - training images as DB, test images as queries
  - precision/recall measured for each query, averaged over all queries
  - PR curve summarized by its integral PR Area (PRA)
  - 8x8 image neighborhoods, GMM classifier
  - various feature transforms: DCT, wavelet, PCA, and ICA
- Evaluation: PRA vs number of selected features

Results

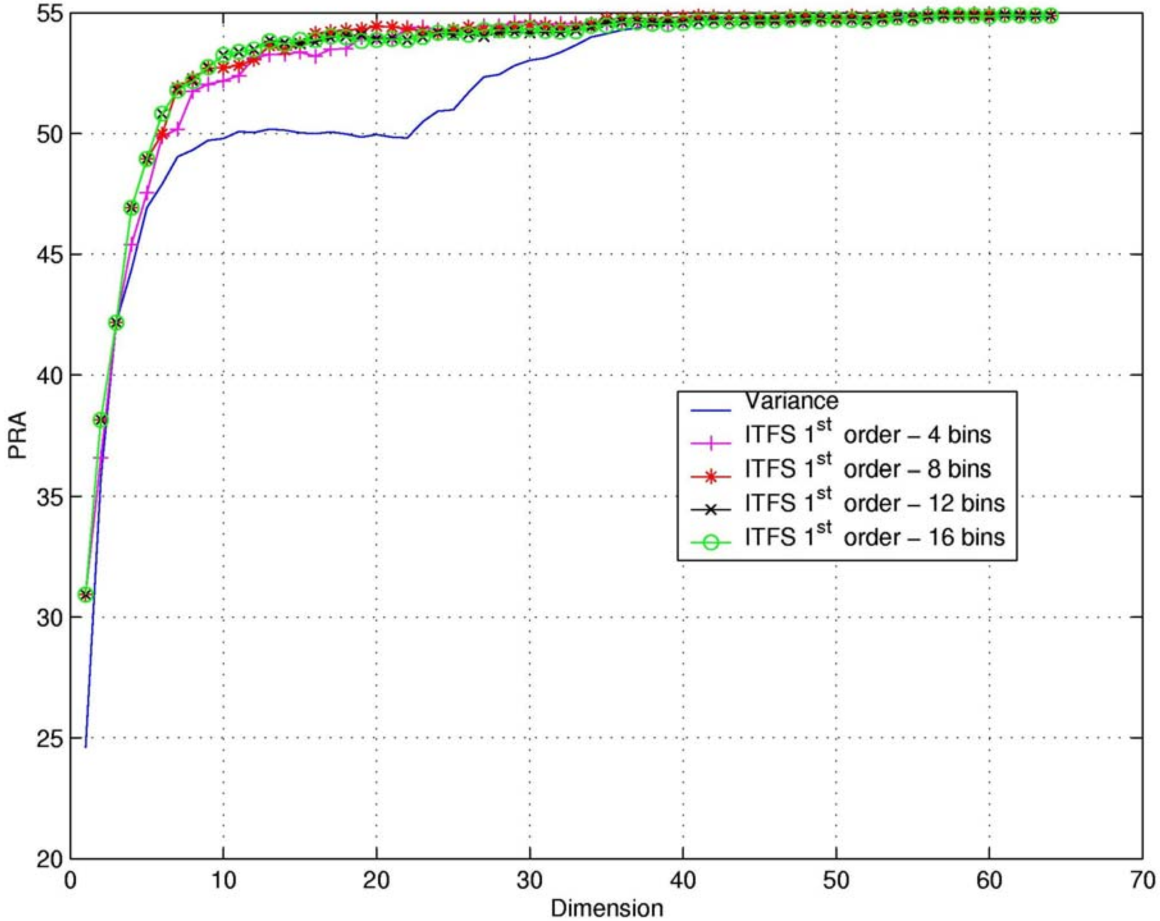
- question: how large does *l* have to be?
- compared ITFS with *l in {0,1,2}* and variance compaction
- PRA shown for Corel and DCT features
- similar results on Brodatz & with other feature sets



- Main observations:
  - ITFS can significantly outperform variance-based methods (10 vs 30 features for equivalent PRA)
  - for ITFS there is no noticeable gain for *l > 1*!

Results

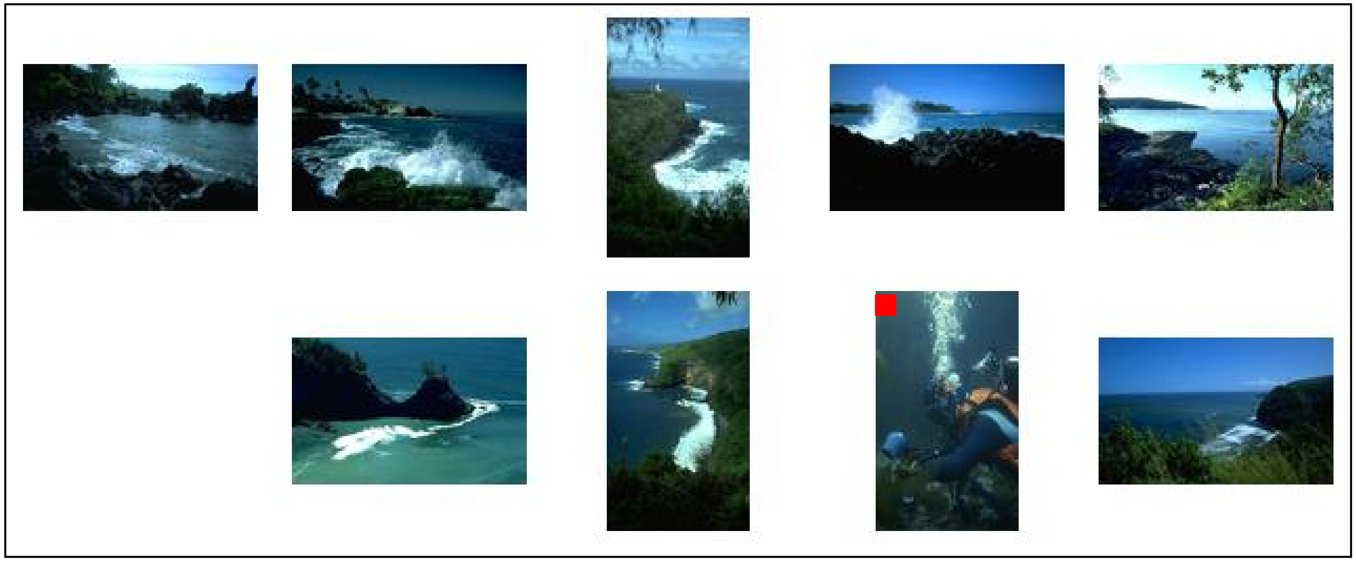
- question: how accurate do the density estimates have to be?
- compared ITFS with *l =1 and various histogram sizes*
- PRA shown for Corel and DCT features
- similar results on Brodatz & with other feature sets



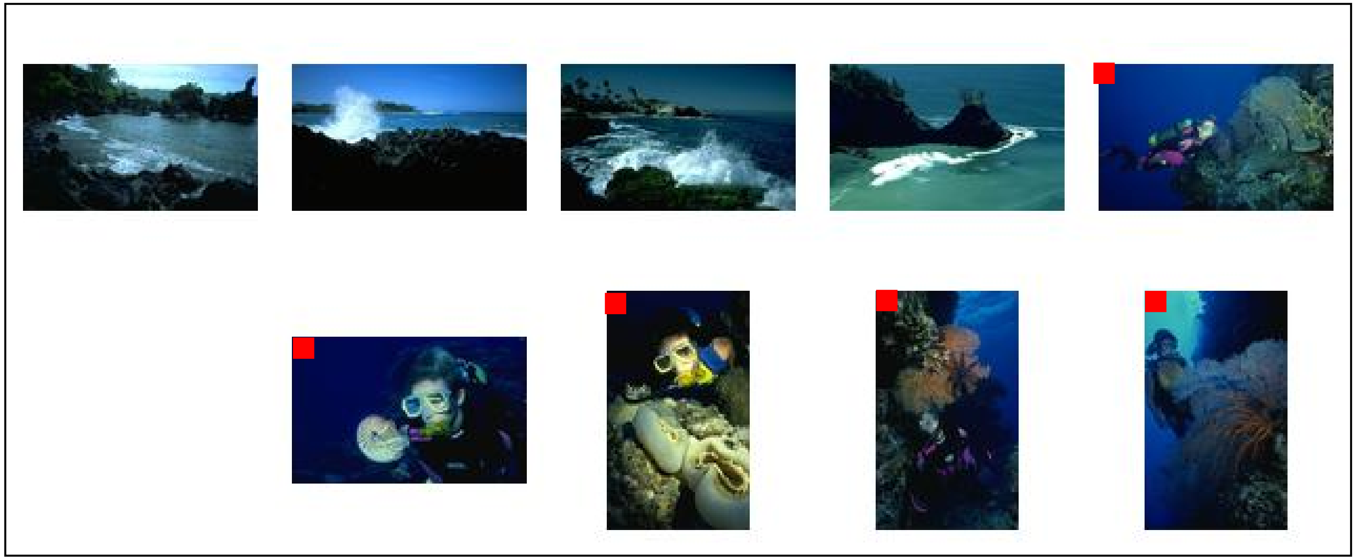
- Main observations:
  - ITFS is quite insensitive to the quality of the estimates (no noticeable variation above 8 bins per axis, small degradation for 4)
  - always significantly better than variance

ITFS vs variance

ITFS:
(l=1)



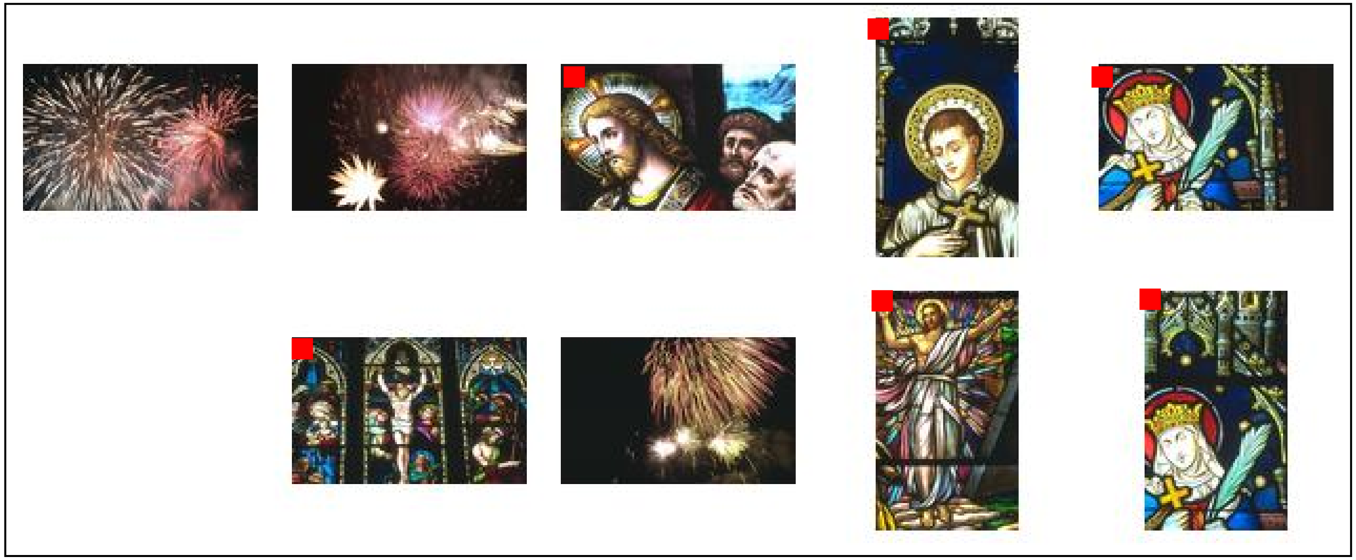
Var:



ITFS vs variance

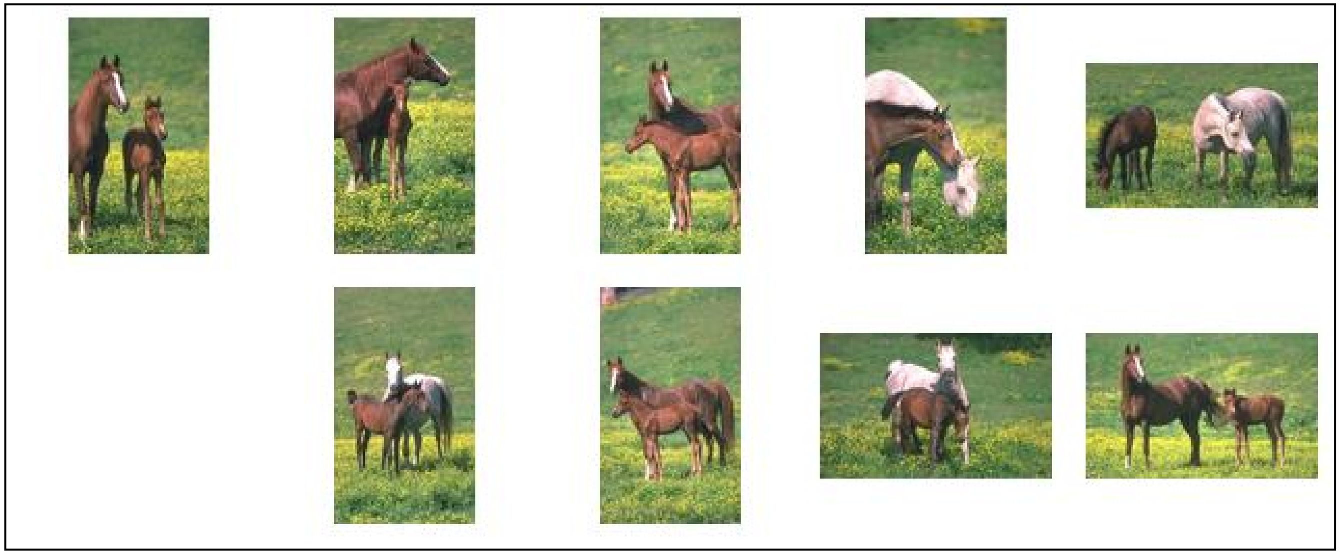
ITFS:
(l=1)

Var:

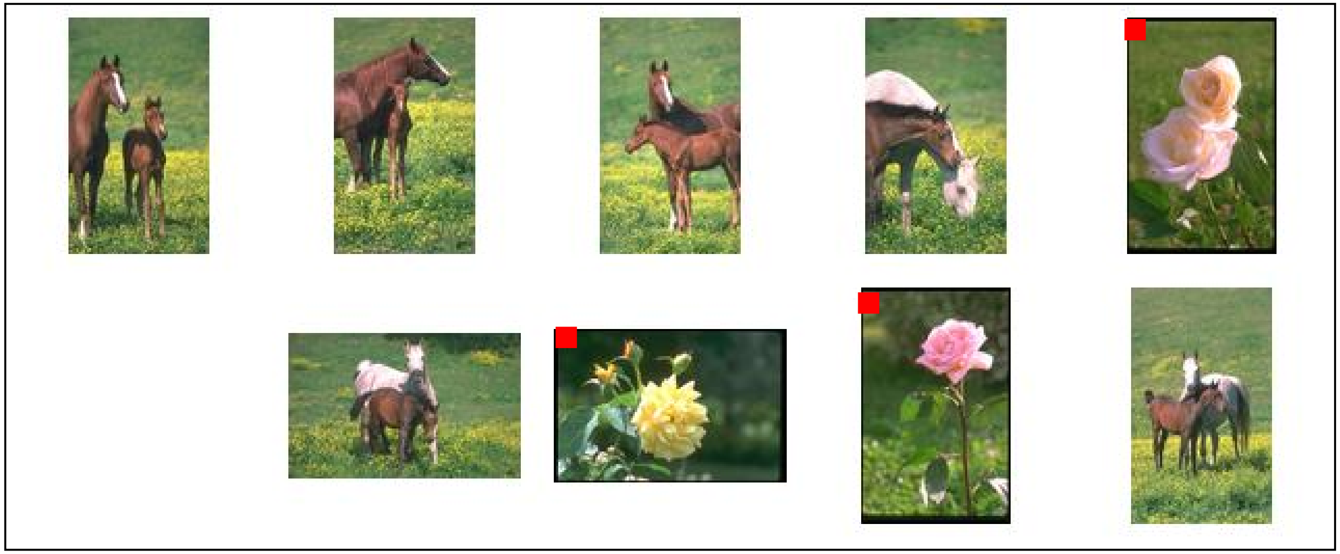


ITFS vs variance

ITFS:
(l=1)



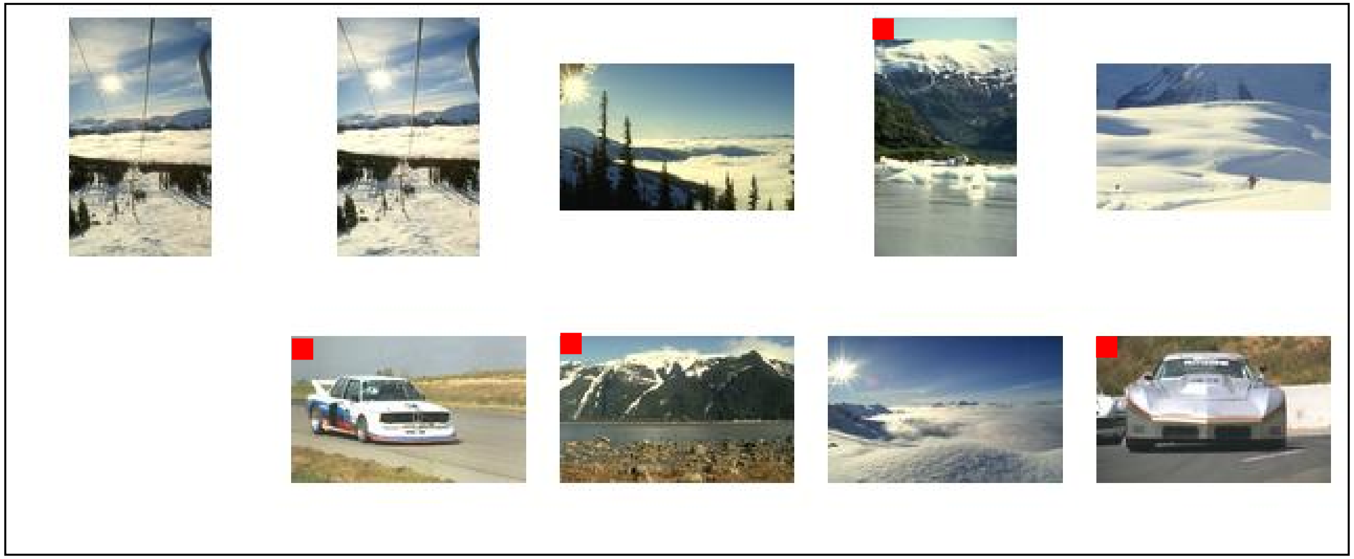
Var:



ITFS vs variance

ITFS:
(l=1)

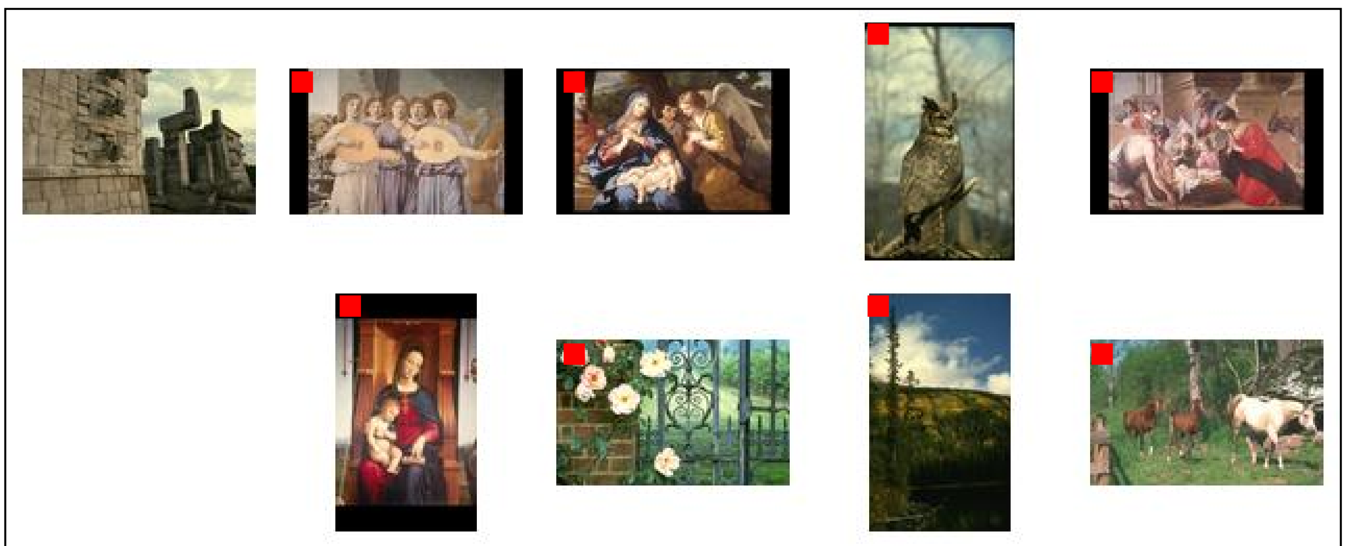
Var:



ITFS vs variance

ITFS:
(l=1)

Var:



Conclusions

- feature selection: search for the Bayes error-optimal space of a given classification problem
- relationships between BE and infomax, make latter natural formalism to understand trade-off between dependence and discrimination
- introduced the concept of *l*-decomposability
- family of FS algorithms that trade-off infomax optimality for complexity
- second-order dependencies seem to be sufficient to achieve near-optimal performance
- optimal/discriminant FS with reduced complexity