Bridging the Gap between Structural and Statistical Pattern Recognition

Horst Bunke

Melchor Visiting Professor
Department of Computer Science and Engineering
University of Notre Dame

and

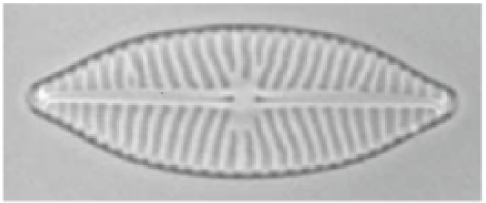
Institute of Computer Science and Applied Mathematics
University of Bern, Switzerland

bunke@iam.unibe.ch
http://www.iam.unibe.ch/fki/staff/prof.-dr.-horst-bunke

Contents

Introduction

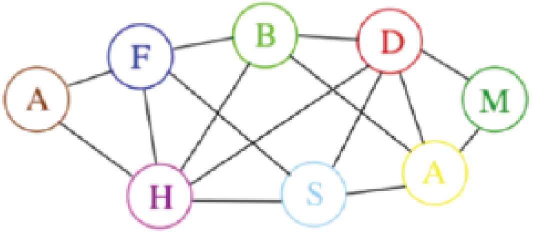
Traditional subdivision of pattern recognition:



Statistical description

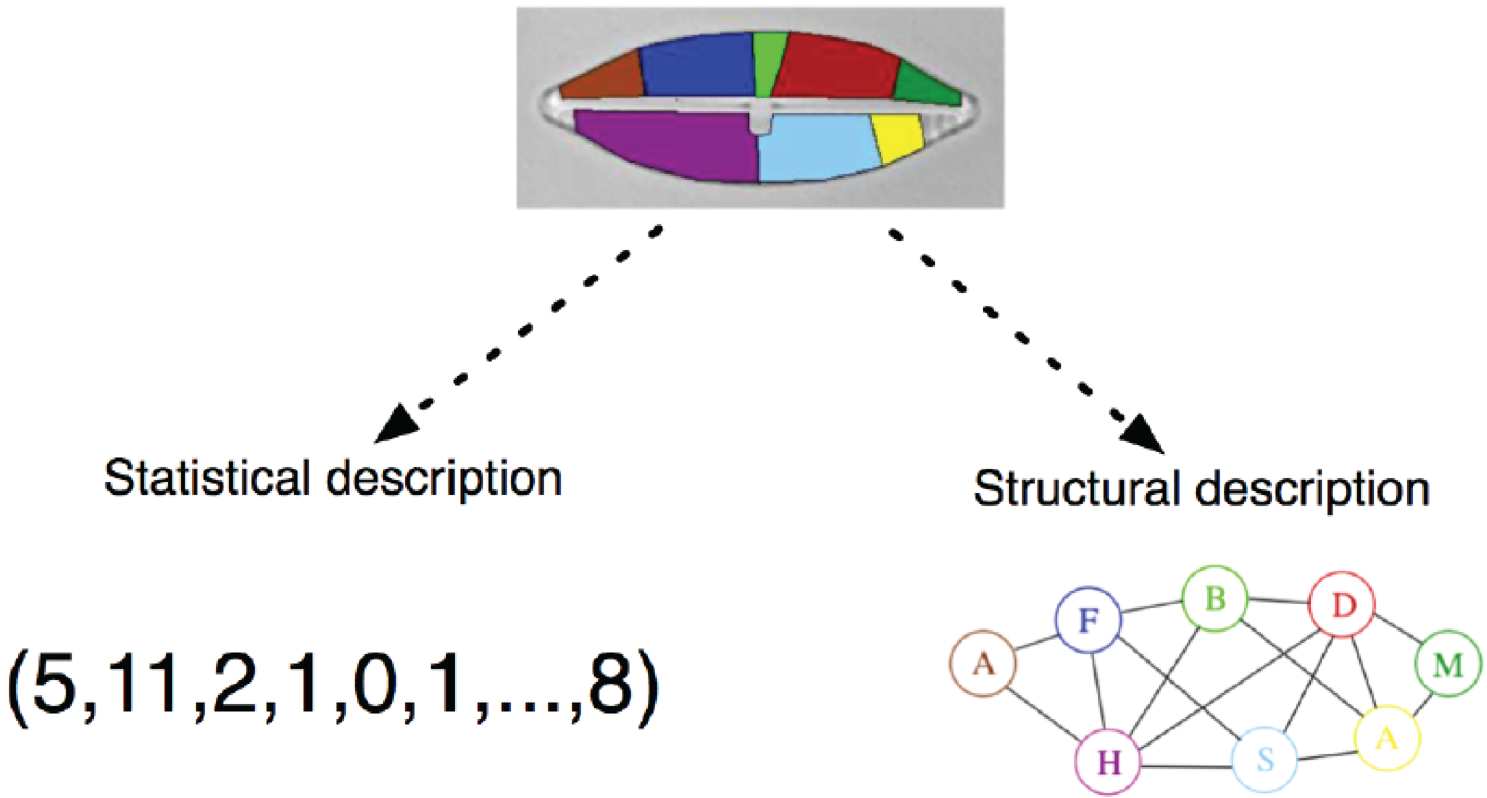
(5,11,2,1,0,1,...,8)

Structural description



Introduction

Traditional subdivision of pattern recognition:



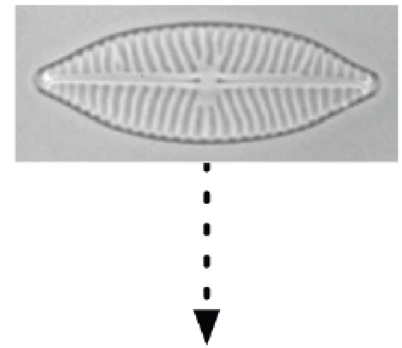
Statistical Approach

Advantages:

- Theoretically well founded
- Many powerful algorithms available

Disadvantages:

- Dimension of feature vectors fixed
- Only unary feature values, but no relations can be modelled



Statistical description

$$(5,11,2,1,0,1,\ldots,8)$$

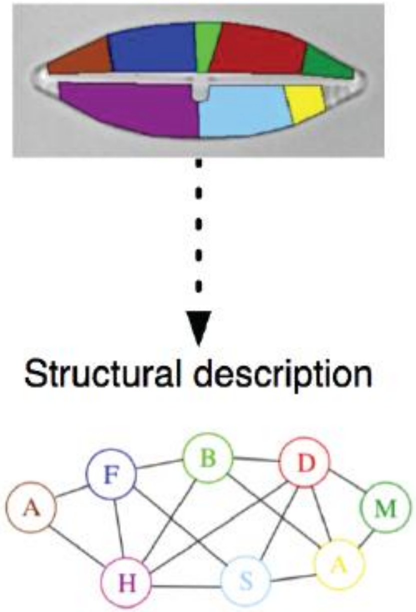
Structural Approach

Advantages:

- Representation size is variable
- Higher representational power (structural relationships)

Disadvantages:

- Lack of mathematical structure in the graph domain
- Lack of algorithmic tools



- Overview

	vectors	graphs
representational power	-	+
available tools	+	-

- Overcoming the limitations:
  - Graph kernels
  - Graph embedding

Illustration of the *kernel trick*:

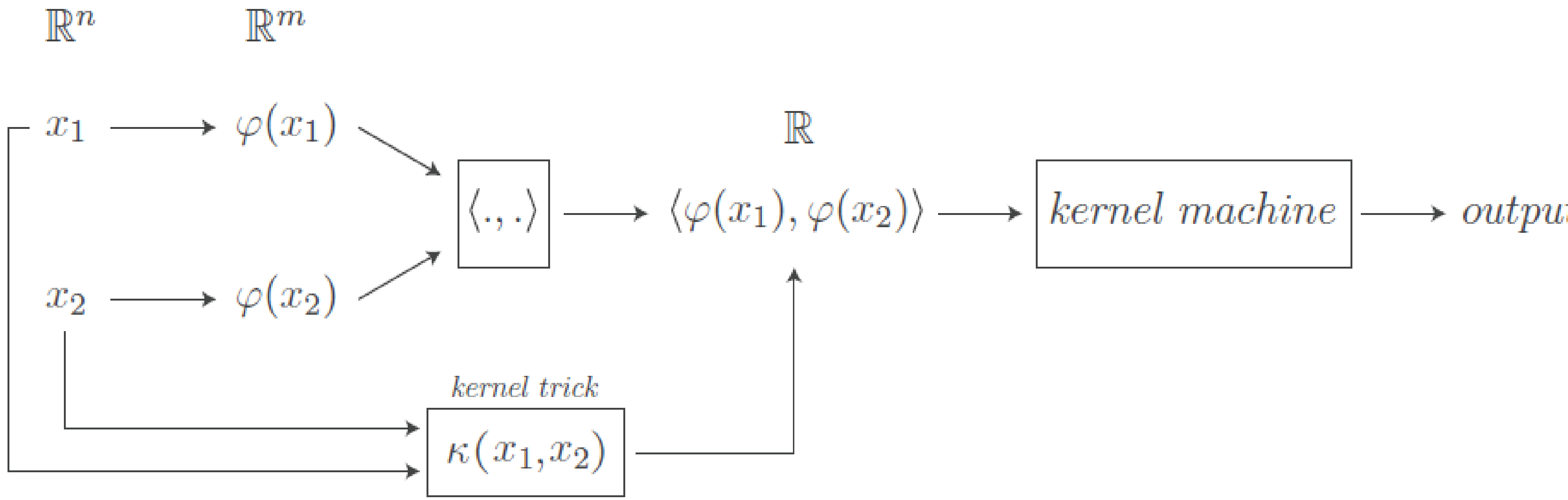
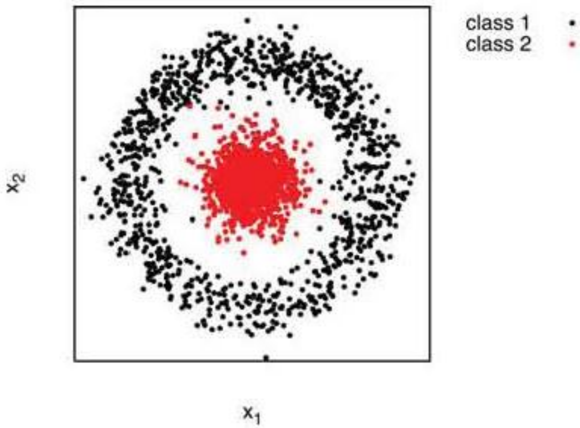


Illustration of a problem that becomes linearly separable after transformation into a new feature space:



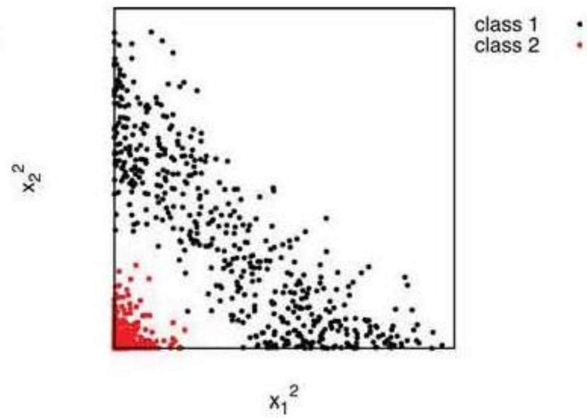
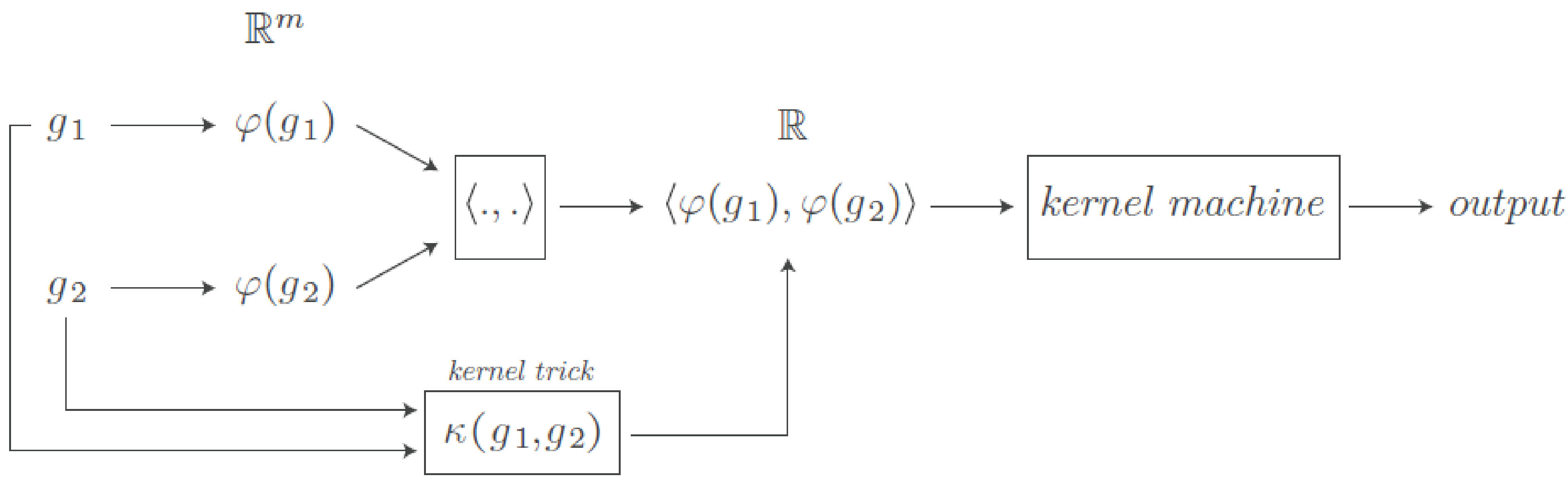


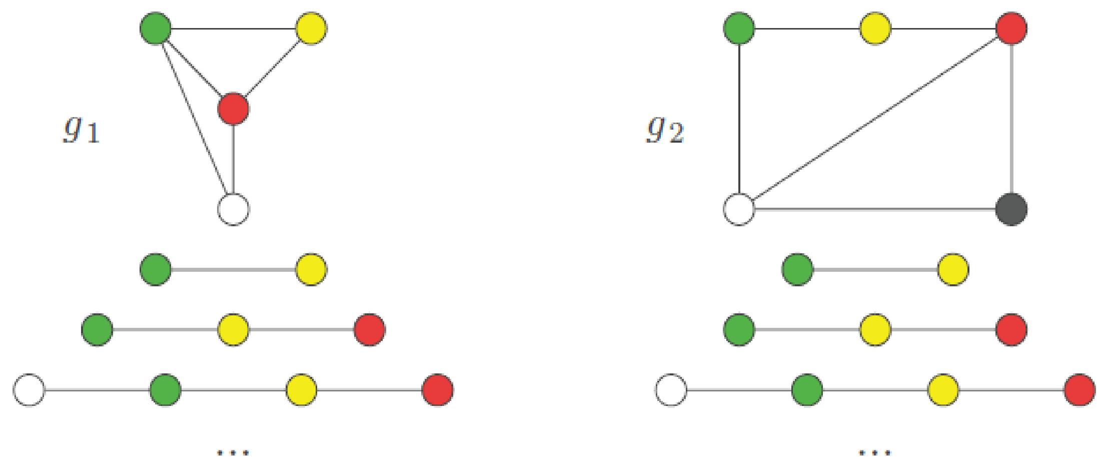
Illustration of the *kernel trick* applied to graphs:



consequences: all kernel machines that have been developed for feature vectors become instantly applicable to graphs

Illustration: Random Walk Kernel

Random walk kernel: compute number of pairs of common walks with identical label sequences (of arbitrary lengths)



$$\kappa(g_1, g_2) = \sum_{i,j=1}^{|V_x|} \left[\sum_{n=0}^{\infty} \lambda_n E_x^n \right]_{i,j}$$ where E_x is the adjacency-matrix of the product graph

for suitable weights $\lambda_n = \gamma^n$ the sum exists: $\lim_{i \to \infty} \sum_{n=0}^{i} \gamma^n E^n = (I - \gamma E_x)^{-1}$

Graph Embedding

- With graph kernels we are still confined to using only kernel machines
- Graph embedding maps graphs to points in $\mathbb{R}^n$:

  Definition: Let G be a set of graphs. A *graph embedding* is a function $\varphi : G \to \mathbb{R}^n$ mapping graphs to n-dimensional vectors, i.e.,

  $$\varphi(g) = (x_1, \ldots, x_n)'$$

- Consequently, graph embedding gives us access to non-kernelizable algorithms as well

- Previous work:
  - Fingerprints in chemo-informatics, graphlets
  - Topological features from complex network research
  - Various features based on eigen-decomposition, Ihara coefficients, etc.

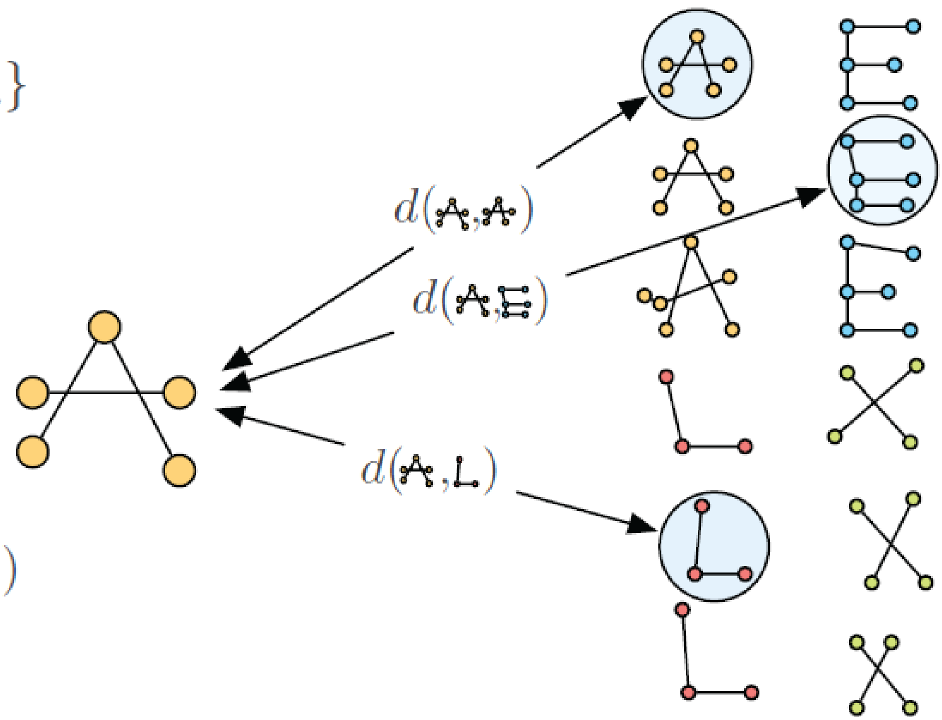
Graph Embedding in Dissimilarity Space

- Graph Set: $G = \{g_1, \ldots, g_t\}$
- Graph edit distance: $d(g_1, g_j)$
- Prototype set: $P = \{p_1, \ldots, p_n\}$
- The mapping

$$\varphi_n^P : G \to \mathbb{R}^n$$

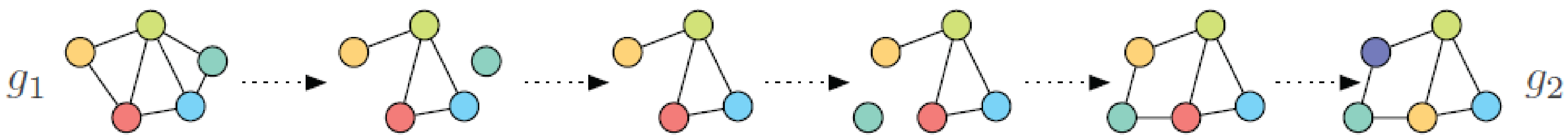
is defined as the function

$$\varphi_n^P(g) \mapsto (d(g, p_1), \ldots, d(g, p_n))$$

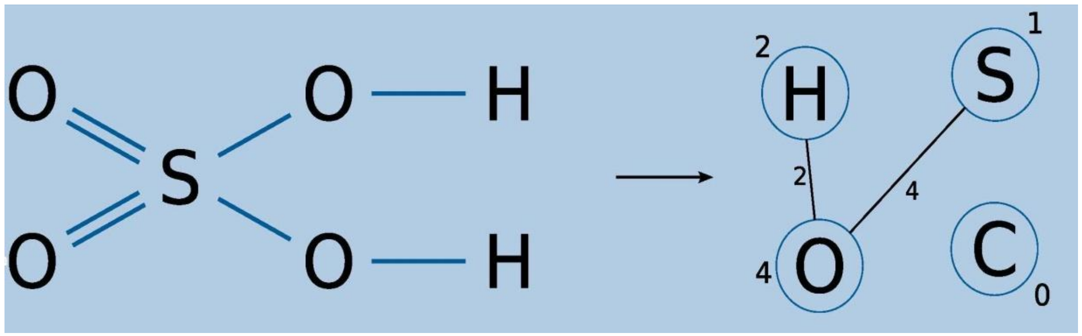


Graph edit distance $d(g_1,g_2)$

- Measures the distance (dissimilarity) of given graphs g_1 and g_2
- Is based in the idea of editing g_1 into g_2
- Common edit operations are deletion, insertion and substitution of nodes and edges
- Can be used with a cost function
- Is computationally expensive, but approximate solutions with complexity $O(n^3)$ exist



Graph Embedding by 1st and 2nd Order Node Label Statistics



$$\varphi(g)=(\underbrace{2,4,0,1}_{\text{nodes}},\underbrace{0,2,0,0,0,0,4,0,0,0}_{\text{edges}})$$

- Equivalent to counting the number of nodes with a certain label, and the number of edges between pairs of nodes with given labels
- Only O(n)+O(e) time complexity
- Extensions:
  - Continuous (non-discrete) node labels
  - Edges labels
  - Experimental results comparable with dissimilarity space embedding

Application 1: Automatic Transcription of Handwritten Medieval Text

A. Fischer. Handwriting Recognition and Historical Documents. Phd Thesis, University of Bern, 2012

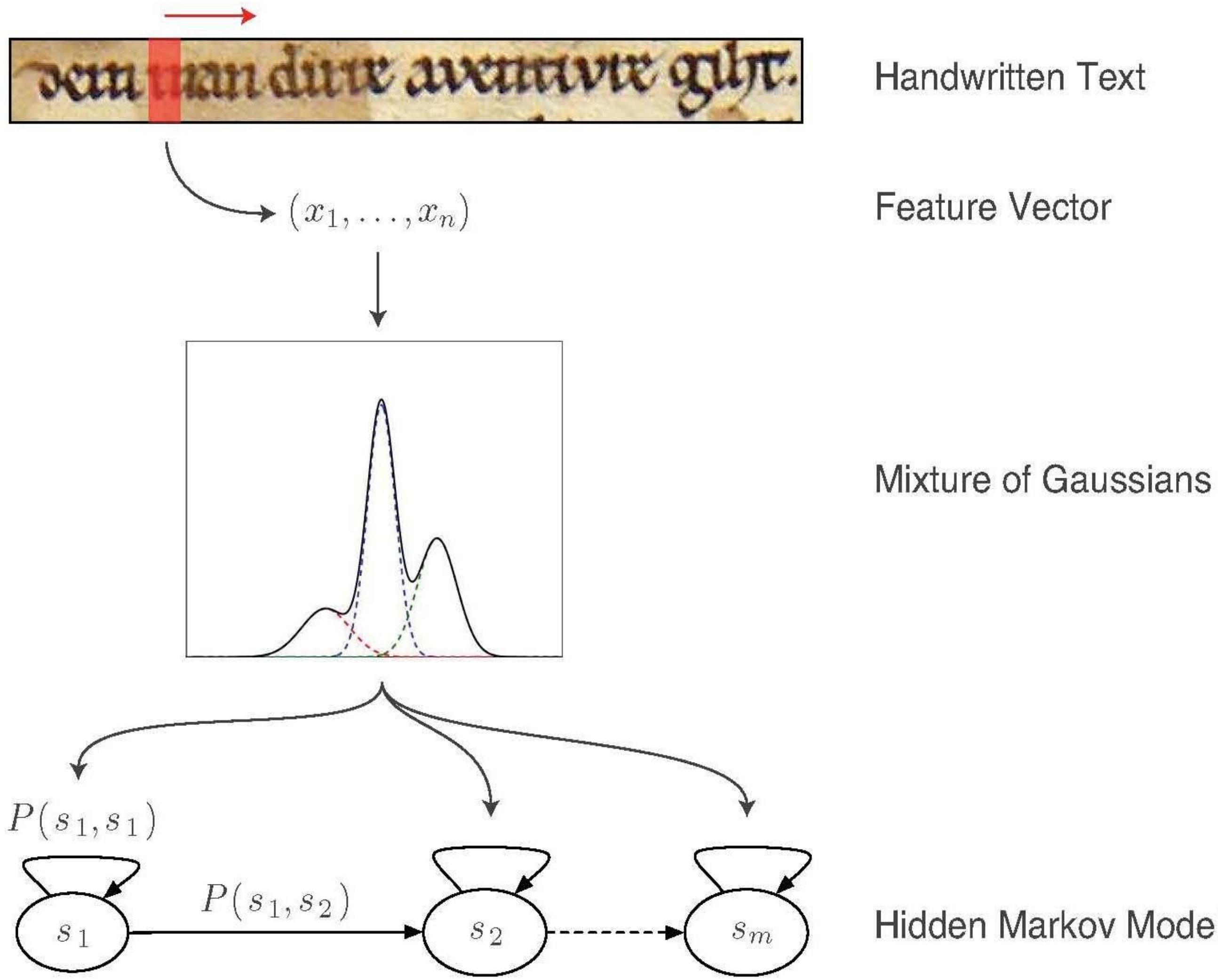
- Digitization of historical documents has become a focus of intensive research
- Objective is to maintain cultural heritage and make vast amounts of historical material available on the internet
- Not only digitization, but also transcription is needed


beginch.
bvche

Challenges in the Transcription of Handwritten Historical Documents

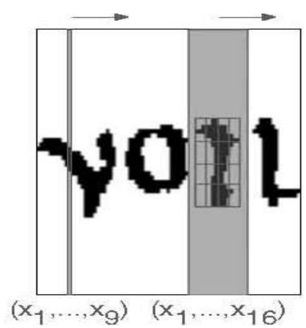
- Layout analysis and extraction of text
  - Decorations
  - Decay of paper or parchment
  - Faded ink
  - Bleed through
  - Various other artifacts
- Acquisition of training samples for recognition costly and difficult (language often known only to experts, special letters)
- Lack of language model, etc.

Conventional Approach



Conventional Features

- Based on a sliding window, e.g. features by
  - Marti et al.: 9 features extracted from a window of 1 pixel width
  - Vinciarelli et al.: 16 windows of size 4 x 4 pixel; fraction of black pixels in each window; result: 16 features



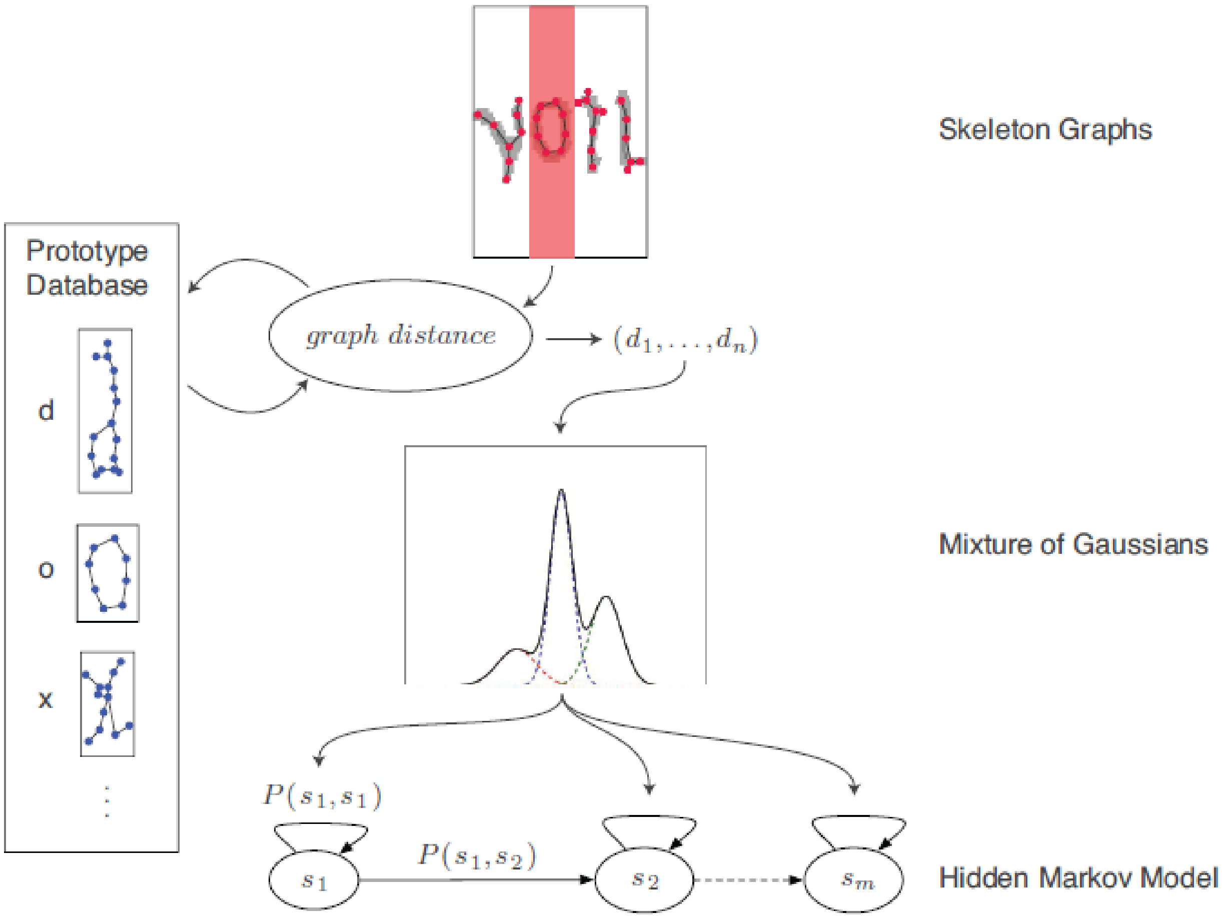
- **Potential problem with conventional approach:**
  - Two-dimensional shape of characters is not adequately modeled; no structural relations

- **Possible solution:**
  - Use skeletons to represent the handwriting by a graph
  - Transform the graph of a handwritten text into a sequence of feature vectors
  - Apply HMMs or RNN to sequence of feature vectors

Graph Extraction

- Apply a thinning operator to generate the skeleton of the image
- Nodes:
  - Key points: crossings, junctions, end points, left-most points of circular arcs
  - Secondary points: equidistant points on the skeleton between key points; distance d is a parameter
- Edges:
  - Nodes that are neighbors on the skeleton are connected by edges
  - However, in the experiments it turned out that the performance without edges is comparable to that with edges if parameter d is chosen appropriately; therefore, no edges were used

General Idea of Graph Based Approach

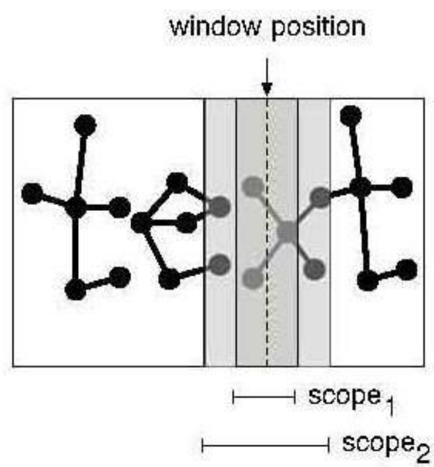


Prototype Selection

- One prototype per class manually selected
- Prototypes automatically selected from automatically extracted characters

Sliding Window

- Width of window is dynamically adapted to width of prototype

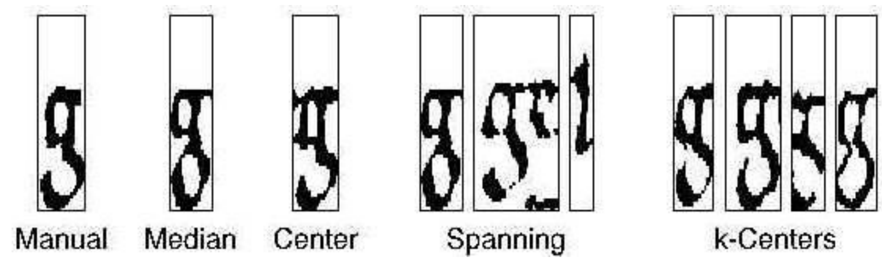


Experimental Results

Features	Prototype Selector	Recognition rate (single word rec.)
Marti		88.69
Vinciarelli		90.49
Graph	manual	94.00
	median	94.07
	center	94.31
	spanning	94.14
	k-centers	94.51*

- Stat. sign. (t-test, α=0.05)

Selected Prototypes



- Number of prototypes for Spanning and k-Centers was determined from the interval [1,5] on a validation set

Comments

- In this application, graph-matching based feature extraction could reduce the error rate by about 50% compared to a standard set of features
- Because the graphs are rather small, the additional computational cost is moderate (compared to HMM decoding)
- Combining different feature sets or different classifiers with each other could be an interesting topic for further studies
- Recent experiments with alternative graph distance measures have given promising results

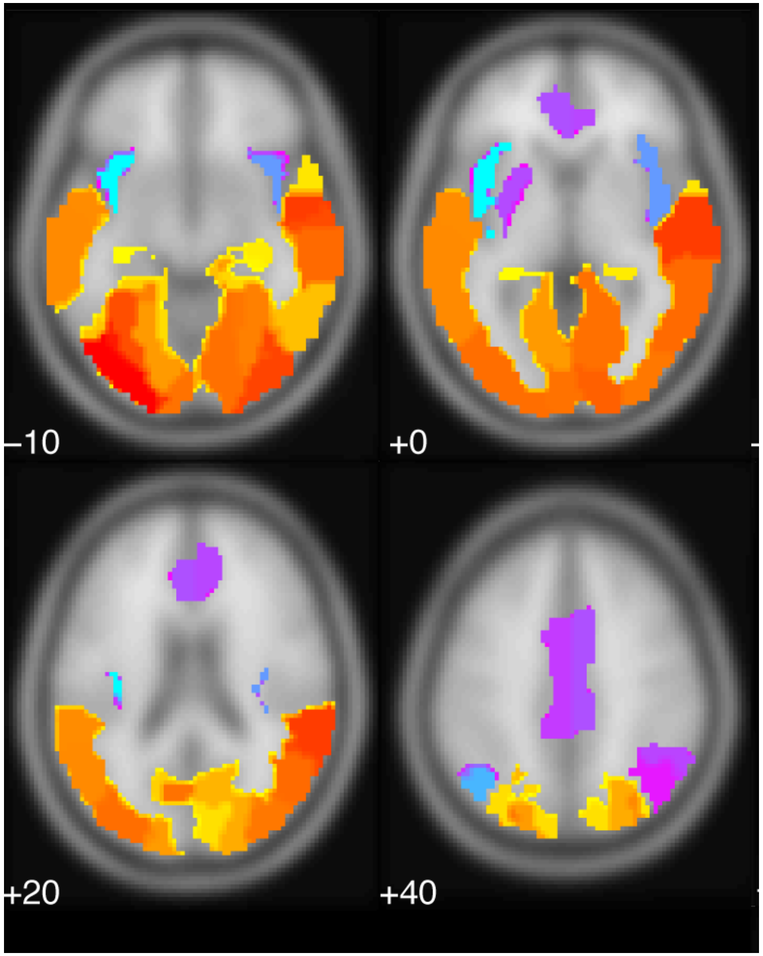
Application 2: Brain State Decoding Using Functional Magnetic Resonance Imaging (fMRI)

J. Richiardi, D. Van De Ville, K. Riesen, and H. Bunke. Vector space embedding of undirected graphs with fixed-cardinality vertex sequences for classification. In Proc. 20th Int. Conference on Pattern Recognition, pages 902–905. IEEE Computer Society Press, 2010.

J. Richiardi, S. Achard, H. Bunke, D. Van De Ville, D.: Machine learning with brain graphs, IEEE Signal Processing Magazine, 2013 to appear

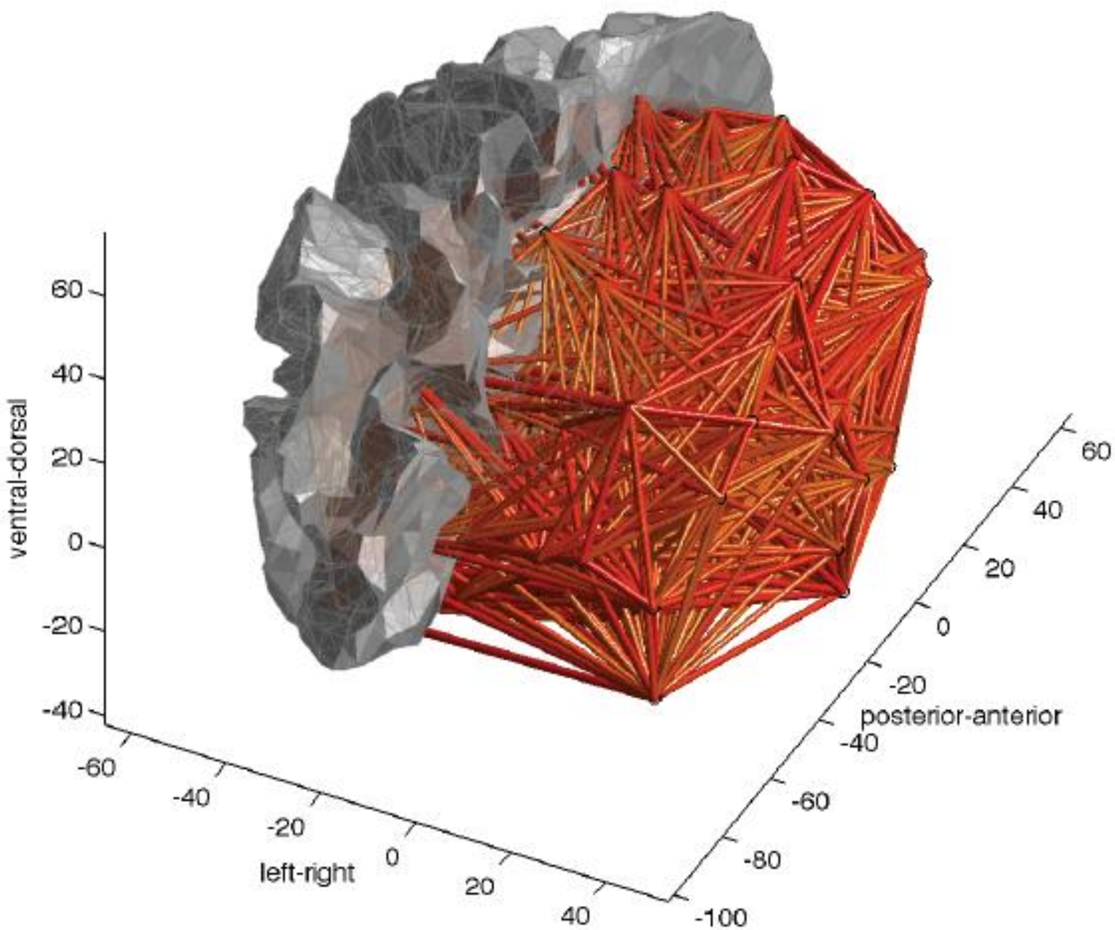
- Partners: University of Geneva, EPFL Lausanne, University of Bern
- Task: from fMRI data, decide whether a person is resting or watching a movie
- Perspective in the long range:
  - “Mind reading”
  - Better understanding of the brain
  - Clinical use (better diagnostic and therapeutic procedures)

fMRI is a technique for measuring brain activity. It works by detecting the changes in blood oxygenation and flow that occur in response to neural activity. When a brain area is more active it consumes more oxygen and to meet this increased demand blood flow increases to the active area. Hence, fMRI can be used to produce activation maps showing which parts of the brain are involved in a particular mental process.



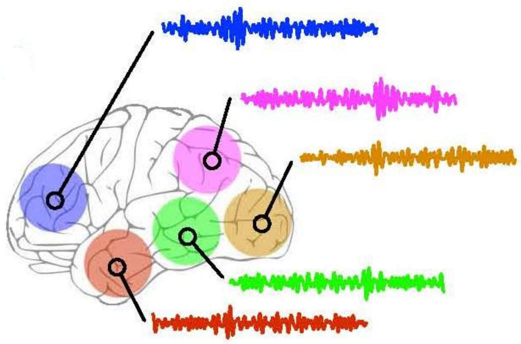
−10
+0
+20
+40

Basic model/understanding in this work: brain is a graph



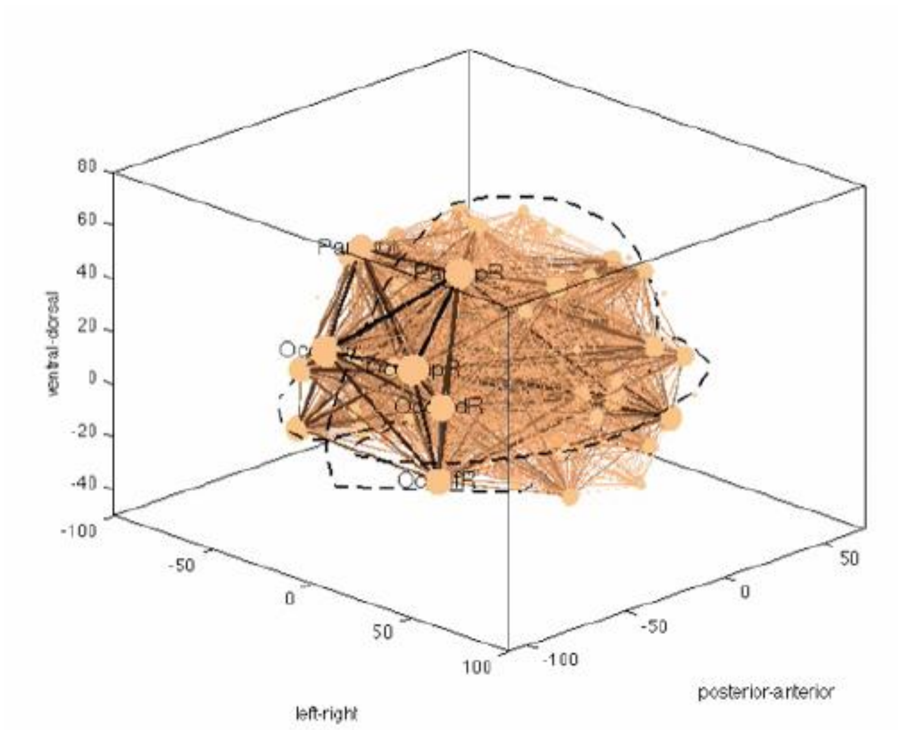
Data Acquisition

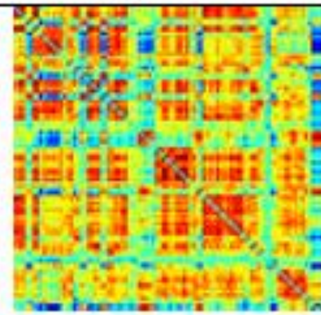
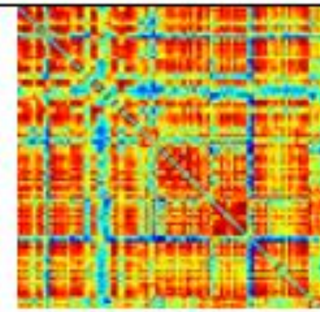
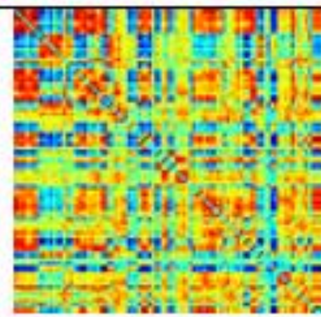
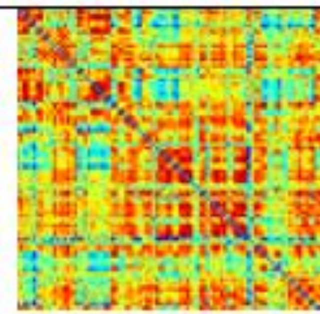
- fMRI images from 15 subjects (4-D data)
- Spatio-temporal resolution 3.75 x 3.75 x 4.2 mm^3 x 1.1 s
- 9 alternating blocks of resting (90 s) and watching movie (50s), concatenated to one sequence for each activity
- All voxels are mapped to a *brain atlas* that contains 90 regions
- As a result, one gets two time sequence $x_1,x_2,\ldots,x_n$ and $y_1,y_2,\ldots,y_m$ for each region, one for each activity
- These time series are filtered into four sub-bands using orthogonal discrete wavelet transform
- Finally, four times series are obtained for each region and each activity



Graph Generation

- Nodes: each region of the brain atlas is represented by a node; for each node we have four times series (for each of the two activities)
- Edges: the graph is completely connected and has weighted edges; the edge weight is the correlation coefficient r ∈ [-1,1] between two time series of the same sub-band and the same activity



	Resting	Movie
Subband 1		
Subband2		

Classification Experiment

- Graphs were transformed to features vectors (graph embedding)
  - Concatenate upper right diagonal of adjacency matrix into one long vector (d=4005)
  - Apply dissimilarity space embedding, using all graphs from the training set as prototypes (d=29)
- Three standard classifiers were applied (all from WEKA):
  - SVM with linear kernel
  - Decision forest
  - Multilayer perceptron (only for dissimilarity space embedding)
- Leave-one-out protocol because of small data set (15 graphs per class and subband)

Experimental Results

Subband	Classifer	DE	DISSE
1	SVM DF MLP	53% 53% -	53% 57% 53%
2	SVM DF MLP	87% 80% -	60% 60% 63%
3	SVM DF MLP	93% 93% -	83% 77% 87%
4	SVM DF MLP	97% 83% -	83% 67% 83%

Comments

- Direct embedding yields feature vectors of very high dimensionality
- Dissimilarity space embedding yields feature vectors of rather low dimensionality (due to small data set)
- A solution in between could lead to even better results
  - Apply feature reduction methods after direct embedding
  - Extend data set to obtain more prototypes (i.e. dimensions) for dissimilarity embedding
- A combination of several, or all, sub-bands could be beneficial as well

Summary, Discussion, and Conclusions

- Structural PR allows us to represent objects in terms of their parts and relations between them, which is an advantage over statistical PR
- On the other hand, statistical PR offers a wealth of mathematical tools for classification, clustering, and similar tasks
- Graph kernels and graph embedding allow us to get the best from both worlds
- In addition to introducing graph kernels and graph embedding in this talk, we have reviewed two applications where these concepts were successfully applied
- There remain a number of challenges for future research:
  - Make methods faster (like linear time embedding)
  - Make them able to deal with graphs consisting of millions of nodes
  - Develop software tools and make them available on the web

- Graph based methods have emerged in various fields:
  - Pattern Recognition and Computer Vision
    - *X. Bai, J. Cheng, E. Hancock (eds.): Graph-Based Methods in Computer Vision, IGI Global, 2013*
    - *O. Lezoray, L. Grady (eds.): Image Processing with Graphs: Theory and Practice, CRC Press, 2012*
    - *K. Riesen and H. Bunke: Graph Classification and Clustering Based on Vector Space Embedding, World Scientific, 2010*
  - Machine Learning
    - *T. Gärtner: Kernels for Structured Data. World Scientific, 2008*
  - Data Mining
    - *D. Chakrabarti, C. Faloutsos: Graph Mining – Laws, Tools, and Case Studies, Morgan & Claypool, 2012*
    - *D. Cook and L. Holder (eds.): Mining Graph Data. Wiley-Interscience, 2007*
  - Complex Network Research
    - *M. Newman: Networks - An Introduction, Oxford University Press, 2010*
    - *E. Estrada: The Structure of Complex Networks, Oxford University Press, 2011*
- Only weak links between the corresponding communities
- But there is a lot that one can learn from another

Bridging the gap between these fields is another great challenge for the future – as hard or even harder than bridging the gap between Structural and Statistical Pattern Recognition

Acknowledgments

- Former students at University of Bern:
  Stefan Fankhauser, Michel Neuhaus, Kaspar Riesen, Andreas Fischer
- Collaborators at EPFL, Lausanne and University of Geneva:
  Jonas Richiardi, Dimitri Van De Ville
- Collaborators at CVC and UPC, Barcelona:
  Jaume Gibert, Ernest Valveny, Miquel Ferrer
- Swiss National Science Foundation
- University of Bern
- University of Notre Dame
- Bob Duin and Ela Pekalska
- Collaborators at DSTO, Edinburgh, Australia:
  Peter Dickinson, Miro Kraetzl
- Collaborators at University of Technology, Sidney:
  Ehsan Zare Borzeshi, Massimo Piccardi