Hybrid Generative-Discriminative Nucleus Classification of Renal Cell Carcinoma

Aydın Ulaş, Peter Schüffler, Manuele Bicego
Umberto Castellani, Vittorio Murino

SIMBAD WORKSHOP 2011
Venice, September 28th, 2011

ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Outline

1. The Problem & The Data Set
2. Probabilistic Latent Semantic Analysis
3. Methodology and Results
4. Discussion and Future Work

Outline

1. The Problem & The Data Set
2. Probabilistic Latent Semantic Analysis
3. Methodology and Results
4. Discussion and Future Work

The Problem

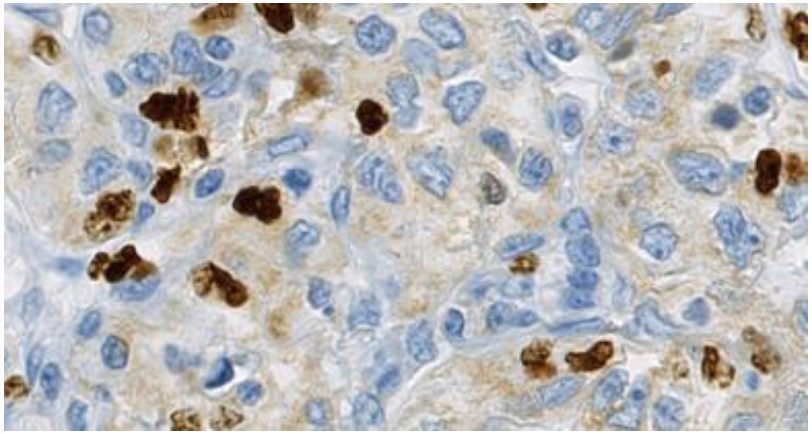
- Renal Clear Cell Carcinoma (RCC) is a very common human cancer
- Cancer cells begin to divide uncontrolled
- Four stages of the cancer are known
- stage I (limited tumor of max 7cm diameter) to stage IV (involvement of distant lymphnodes and metastases)
- Different therapies in different stages are known (from medical treatments to surgery)

The Problem

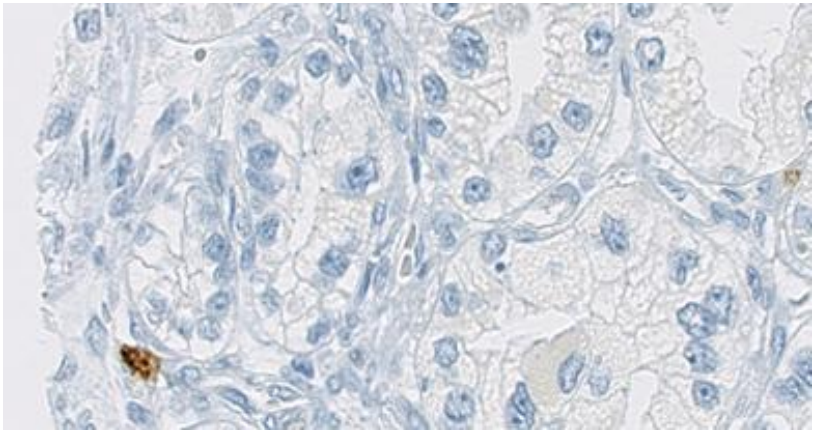
- For the staging, the grading of different protein expression levels might be relevant.
  - E.g. the protein expression level of MIB-1, a proliferation protein. Also other cancer-marker proteins are possible.
  - High grade on MIB-1 means proliferating cancer cells
- Biomarkers help to identify the staging of the disease
- Search for new protein markers that have distinct grading patterns in distinct cancer stages. Research done on **Tissue Micro Arrays.**
- The aim is to automate the process

Examples of MIB-1 grading

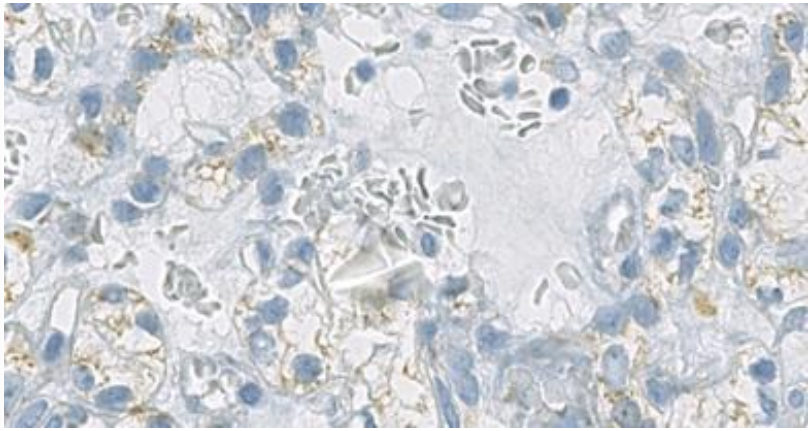
High grade (brown = MIB-1 positive)



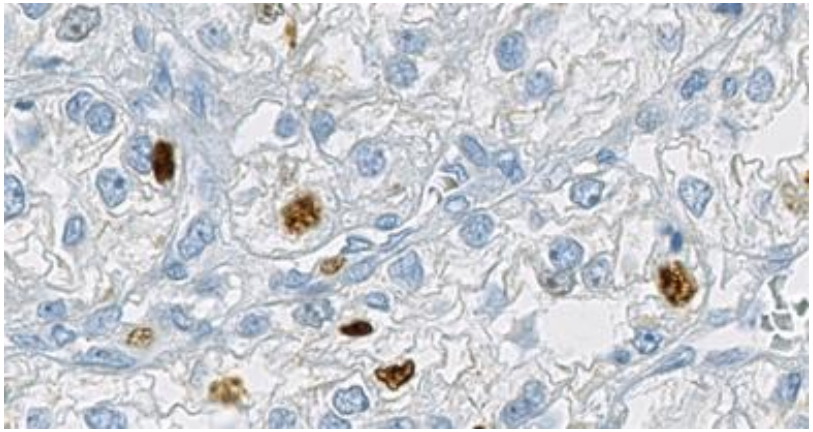
Low grade (blue = cell nucleus)



Zero grade



Medium grade



ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

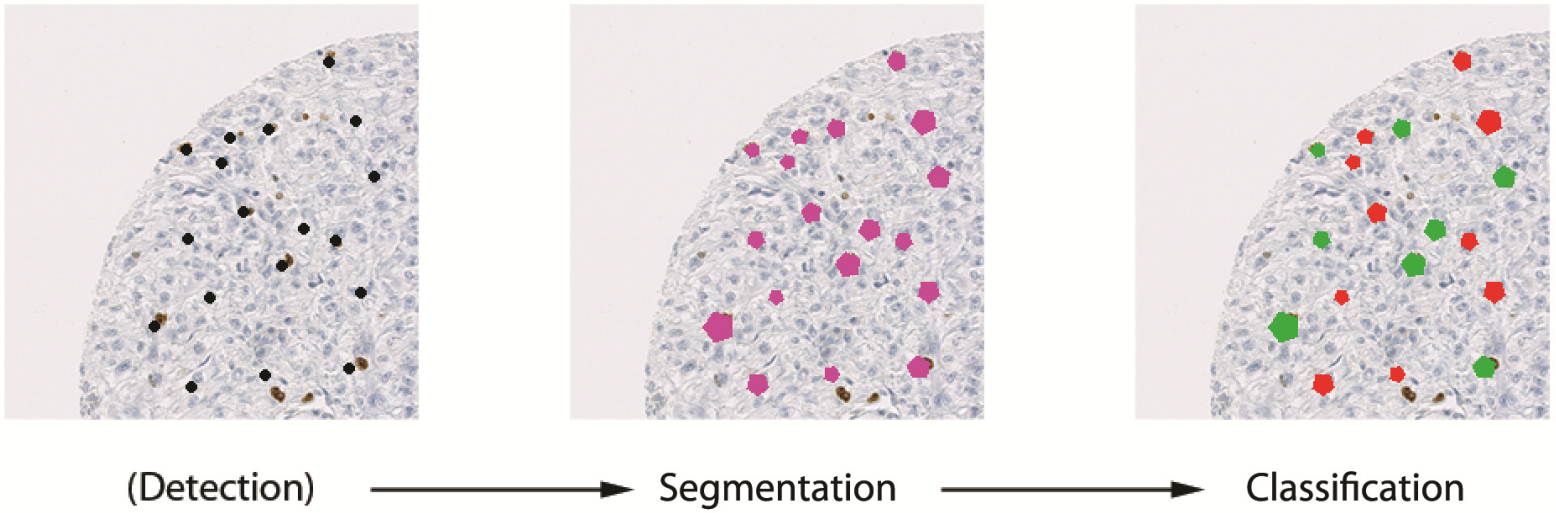
SIMBAD

Problems of MIB-1 Grading

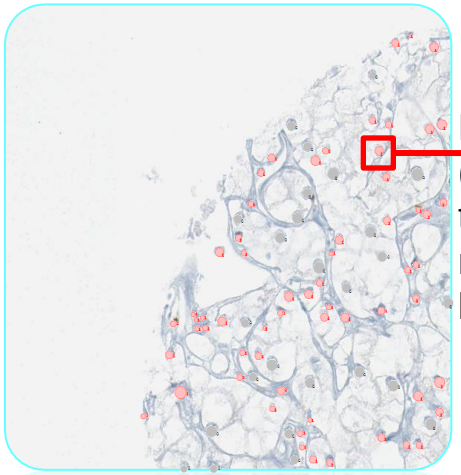
- **Grade = percentage of MIB-1 stained cancerous nuclei** (= #brown nuclei among cancerous)
- Human grading is possible
  - Time consuming (especially in large data sets)
  - Difficult and subjective (high variance among humans)
  - Fuzzy: **«no», «low», «medium», «high»** instead of percentage.
- Manual rating and assessment under microscope is inconsistent
  - High variability of cancerous tissue
  - Subjective experience of humans

TMA – Analysis Pipeline

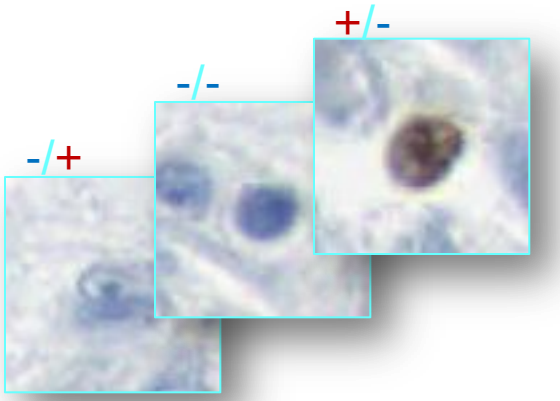
- Computer based approach for Tissue Micro Array grading



Database Design



Nuclei Extraction (detection by two pathologists, nuclei locations known)



8 TMA images
from 8 patients

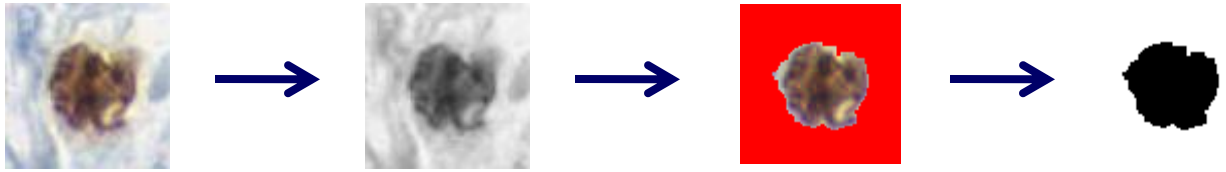
1633 patches (80x80px)
Binary labels from **two pathologists** for each patch
1273 patches with equal label (890 vs. 383)

ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

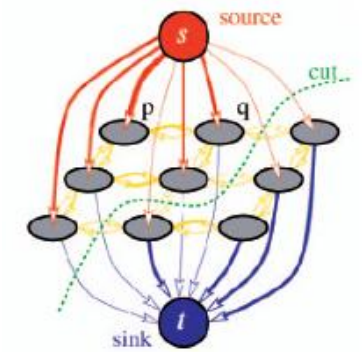
SIMBAD

Nucleus Segmentation

- Segmentation via Graph Cut (Boykov, Veksler)



- Source: Midpoint of patch: Nucleus
- Roundish shape favored

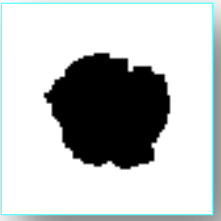


ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

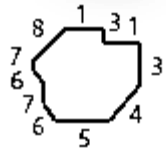
SIMBAD

Feature Extraction

- General features
  - Histogram foreground
  - Histogram background
  - Pyramid Histograms of Oriented Gradients (PHOG, Bosch et al.)



- Shape features
  - Histogram of Freeman Chain Code
  - Histogram of 1D-signature
  - Region properties (area, diameter, …)

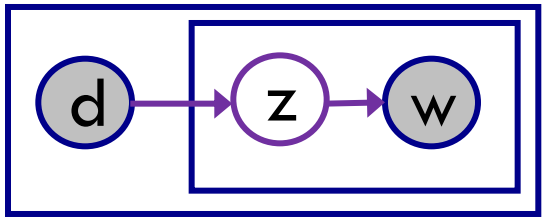


Outline

1. The Problem & The Data Set
2. Probabilistic Latent Semantic Analysis
3. Methodology and Results
4. Discussion and Future Work

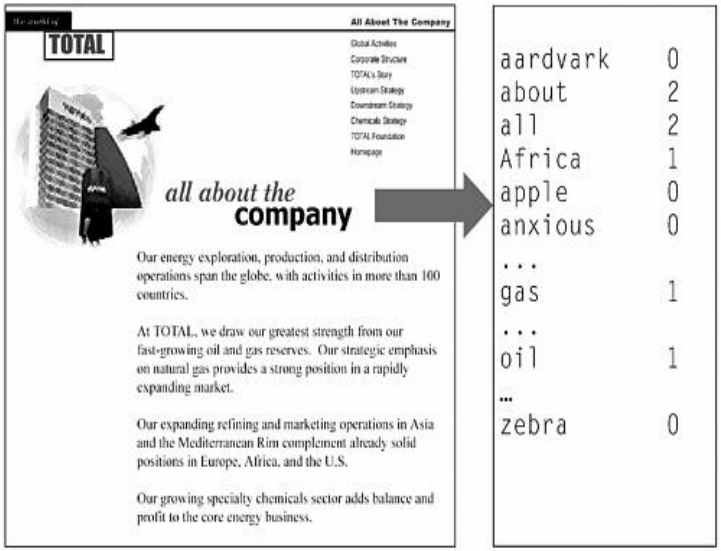
Topic Models

- They are probabilistic tools widely used in text analysis and computer vision communities
- They can model a dataset in terms of hidden topics (processes)



Topic Models (Text Analysis)

- Extension of the Bag of Words (BoW) approach
  - A document is seen as an unordered collection of words
- *Not interested in the position but the number of occurrences of words*
- documents are characterized by word occurrences (histograms)



Topic Models

- Problem: the same word can have different meanings depending on the context

"Home"	"sports"	"space"	"computers"	"weather"
Kitchen	Team	Space	Drive	Rain
Door	Game	Sun	Windows	Snow
Garden	Play	Research	Card	Sun
Windows	Year	Center	DOS	Season
Bedroom	Games	Earth	SCSI	Weekend
Space	Season	NASA	Sun	Cloudy

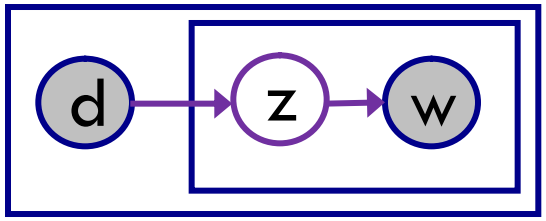
Topic Models

- Words can be disambiguated by looking at the context
- Topic models introduce an intermediate level, based on the concept of **topic**
  - it represents “what we are talking about”
  - the topics are extracted looking at co-occurrence of words in documents
- Every document is characterized by the presence of one or more topics (e.g. sport, finance, politics)

  .... which may induce the presence of some words

Topic Models: PLSA

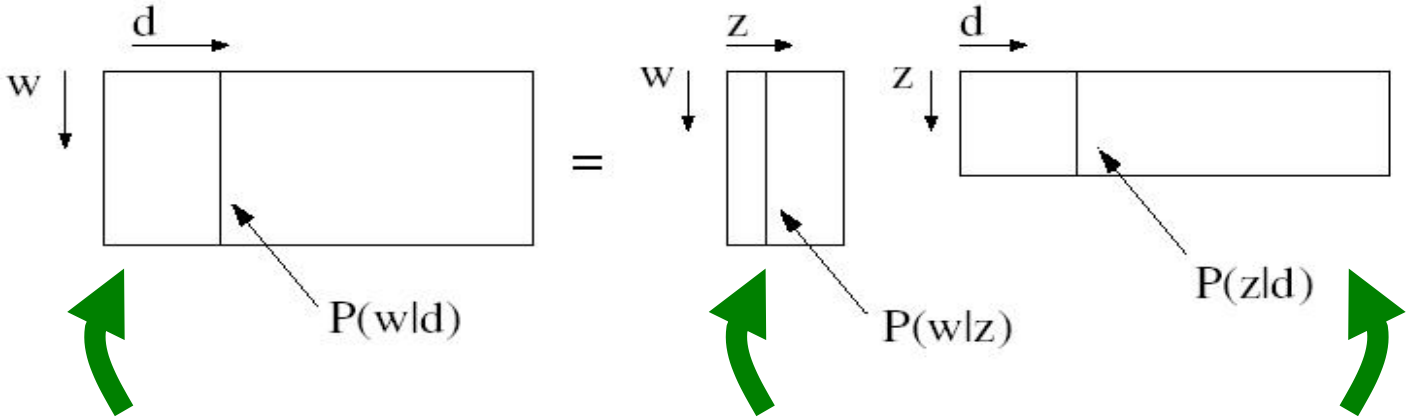
- Here we employ the Probabilistic Latent Semantic Analysis (PLSA)
- Given the counting matrix **n(w,d)** (number of occurrences of the word **w** in the document **d**)
  - the PLSA permits to decompose the probability of a word in a document through the topics distributions



ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

pLSA

$$p(w_i \mid d_j) = \sum_{k=1}^{K} p(w_i \mid z_k)\, p(z_k \mid d_j)$$



Observed word distributions

word distributions per topic

Topic distributions per document

Training: estimation of **p(w|z)** and **p(z|d)**

Outline

1. The Problem & The Data Set
2. Probabilistic Latent Semantic Analysis
3. Methodology and Results
4. Discussion and Future Work

Methodology

- Consider the frequencies as word counts in a document
- Apply pLSA and train to find **p(z|d)**
- On the new space apply classification algorithms

Methodology

- 3 patient subset (474 nuclei patches) selected preserving benign/malignant ratio
- 321 (67 %) benign, 153 (33 %) malignant
- 10-fold stratified CV
- Eight representations (ALL, BG, COL, FCC, FG, LBP, PHOG)
- Number of topics chosen by CV
- Compare results using the original feature space and using the space created by the topic distributions (**p(z|d)**)

Methodology

- Support Vector Machines
  - svl (linear kernel)
  - svp (polynomial kernel, p = 2)
  - svr (radial basis function kernel)
- Classification algorithms
  - ldc: linear discriminant classifier
  - qdc: quadratic discriminant classifier
  - knn: k-nearest neighbor classifier
  - tree: decision tree
- Implemented using PRTools [http://www.prtools.org/]

Results: SVMs

	svl		*svp*		*svr*	
	ORIG	PLSA	ORIG	PLSA	ORIG	PLSA
ALL	68.36	**74.26**	65.40	**75.06**	74.47	75.11
BG	72.88	70.82	66.79	**71.50**	**74.22**	71.92
COL	66.90	69.03	56.93	**70.32**	68.98	68.82
FCC	67.30	67.72	66.89	67.92	67.95	68.57
FG	70.68	71.97	64.12	**72.62**	70.49	71.09
LBP	68.61	69.43	42.36	**70.70**	68.79	70.47
PHOG	75.45	**79.67**	63.92	**79.22**	76.55	76.80
SIG	67.72	**68.34**	58.64	**67.69**	67.72	67.72

ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Results: Other Classifiers

	ldc		*qdc*		*knn*		*tree*	
	ORIG	PLSA	ORIG	PLSA	ORIG	PLSA	ORIG	PLSA
ALL	71.71	70.21	69.55	69.01	72.35	**73.44**	71.97*	70.30
BG	**70.79**	68.31	68.48	67.52	**74.25**	71.29	62.25	**67.29**
COL	69.42	69.86	67.55	67.94	69.41	68.62	60.62	62.44
FCC	66.68	65.25	60.76	**65.19**	66.66	67.71		
FG	70.24	70.70	68.59	68.78	69.79	70.48	63.07	63.46
LBP	71.55	71.98	70.71	68.37	71.13	70.29	60.14	63.97
PHOG	75.29	77.57*	67.93	**74.62***	70.71	**74.69***	63.51	66.49
SIG	67.73	66.87	64.74	**68.95**	63.50	67.72	58.04	61.85

ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Outline

1. The Problem & The Data Set
2. Probabilistic Latent Semantic Analysis
3. Methodology
4. Discussion and Future Work

Discussion

- We proposed to use the generative abilities of pLSA to project our data into a new space
- Using pLSA and applying the topic model idea from NLP, we can project our data into another space and achieve higher classification accuracies
- Except for some specific cases (SVM with radial basis kernels and decision trees), the space created using pLSA is superior to the original space
- The best results are obtained on the new space

Future Work

- Outputs of pLSA (p(z|d) and p(w|z)) are probability density functions
- Kernels can be directly computed from p(z|d) and used in kernel based classification
- Other score spaces based on pLSA can also be used:
  - FESS (free energy score space)
  - Fisher Score
  - PD (posterior divergence) …
- Renal Cancer Cell Classification Using Generative Embeddings and Information Theoretic Kernels (PRIB 2011)

Acknowledgements

- SIMBAD project (EU FP7, contract 213250) for the financial support
- J.J. Verbeek for the implementation of pLSA
- PRTools for the implementation of classification algorithms

Questions?

Why not LDA

- LDA (Latent Dirichlet allocation)
- Topics are assumed to have a Dirichlet prior
- In pLSA, you have to estimate the number of topics
- Theoretically LDA is a better tool
- Our experiments have shown that accuracies are comparable
- You can also estimate the number of topics in pLSA using information theoretic measures such as:
  - BIC (Bayesian Information Criterion)
  - AIC (Akaike's Information Criterion) …