

Learning Image Classification and Retrieval Models

Thomas Mensink

26 October 2012

Jury members

Prof. Frédéric Jurie	Président
Prof. Christoph Lampert	Rapporteur
Dr. Barbara Caputo	Rapporteur
Dr. Cordelia Schmid	Advisor
Dr. Jakob Verbeek	Advisor
Dr. Gabriela Csurka	Advisor

Image Understanding



Loved the colour of the sky and water against the dark tree reflections.



A long way to go to the top... Just above my house, close to Grenoble, in front of Belledonne mountains¹.

Bridge the semantic gap: the relation between low-level image features and semantic interpretation [2]

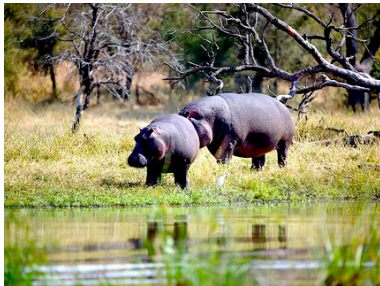
1. Images from the Stony Brook University Captioned Photo data set
2. Smeulders *et al.*, Content-based image retrieval at the end of the early years, PAMI 2000

Image Classification



Giant Panda

fish; group; bush; claws.



Hippopotamus

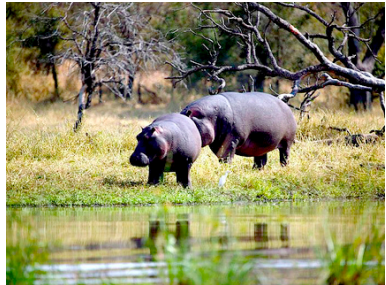
fish; solitary; jungle; walks.

-
1. Images from Animals with Attributes data set

Image Classification






Giant Panda
fish; group; bush; claws.



Hippopotamus
fish; solitary; jungle; walks.

-
1. Images from Animals with Attributes data set

Image Retrieval

Query	Top Ranked Retrieved Images
<i>Paris</i>	
	

1. Query-by-text using Google Image Search
2. Query-by-example using the BigImbaz Visual Copy Detection Search

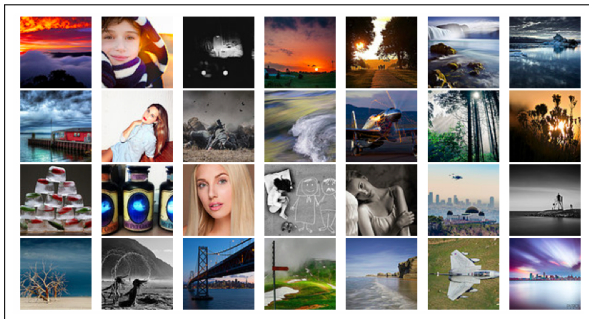
Relation between Classification and Retrieval

- Retrieval as classification
 - Rank images according to confidence score from classifier
- Classification as retrieval
 - Obtain similar images and propagate labels

Goals

1. Scaling to large data sets
2. Adapting to novel classes
3. Leverage user interaction
4. Modeling label dependencies
5. Exploiting multi-modal data

Goal 1 - Scaling to large data sets



Flickr hosts over 6 billion photos

- Large volumes of images
- Large number of potential labels
- Efficient methods for representation, learning and testing

Goal 2 - Adapting to novel classes



300 million photos are uploaded per day to Facebook

- Relevance of categories, classes or labels changes over time
 - new images, products and creations

Goal 3 - Leverage user interaction



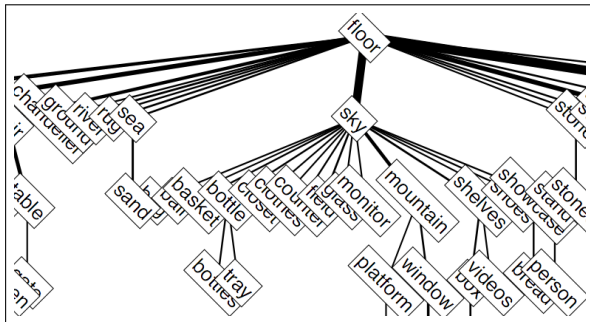
Rock **true**

Sea **true**

Beach **???**

- Incorporate the user set labels into predictions of other labels
- Find the middle ground between
 - automatic image labeling
 - manual image labeling

Goal 4 - Modeling label dependencies



A tree structure defined over image labels

- Labels have an intrinsic structure or dependence
- To benefit from user input, structure is required

Goal 5 - Exploiting multi-modal data



President Barack Obama greets people on the tarmac after arriving at John F. Kennedy International Airport, Monday, July 30, 2012, in New York.

Photo by Jason DeCrow (AP)



The Netherlands, US, Australia and Canada compete in the Men's Pair Heat 2 on Day 1 of the London 2012 Olympic Games at Eton Dorney on July 28, 2012 in Windsor, England.

Photo by Streeter Lecka

News photos from Yahoo! News

- Data is often multi-modal
 - Image is accompanied by title, place and textual description
- Exploit complementarities of visual and textual information to improve image retrieval and annotation



Outline

1. Introduction

- Motivation and Goals
- Image Representations

2. Large Scale Classification and Adapting to Novel Classes

3. Leverage User Interaction using Label Dependencies

4. Exploiting Multi-Modal Data

5. Conclusion and Discussion

Bag of Visual Words

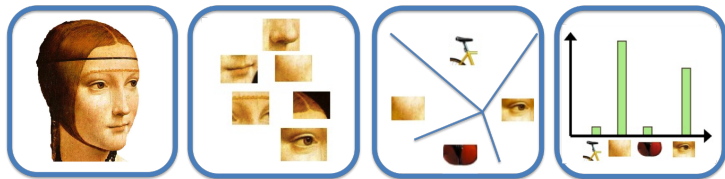
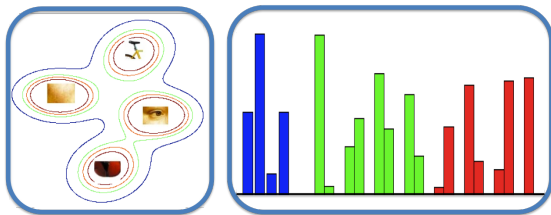


Image courtesy of Li Fei-Fei.

- Successfully used for retrieval [1] and classification [2]
- Bag-of-Visual Words pipeline:
 1. Consider an image as a unordered set of patches
 2. Represent each patch with a descriptor, e.g. SIFT or LCS
 3. Assign each patch to “visual-word” from “visual-dictionary”
 4. Count the frequency of each visual-word
- Visual-dictionary: k-means clustering on large set of patches

1. Sivic and Zisserman, Video Google: A text retrieval approach to object matching, ICCV'03
2. Csurka *et al.*, Visual categorization with bags of keypoints, ECCV'04

Fisher Vector



- Fisher vectors [1] for Mixture of Gaussians [2-3]
 - Visual-dictionary: MoG in feature space $p(\mathbf{x}|\theta)$
 - Take the gradient for each patch $\frac{1}{T} \nabla_{\theta} \ln p(\mathbf{x}_t; \theta)$
- Encodes more information per visual word:
 - frequency, mean and variance
- Power and ℓ_2 normalization:
 - *Improved Fisher Kernel for Large-Scale Image Classification*
Perronnin, Sánchez & Mensink, ECCV'10

1. Jaakkola and Haussler, Exploiting generative models in discriminative classifiers, NIPS'99
2. Perronnin and Dance, Fisher kernels for image categorization. CVPR'07



Outline

1. Introduction
2. Large Scale Classification and Adapting to Novel Classes
3. Leverage User Interaction using Label Dependencies
4. Exploiting Multi-Modal Data
5. Conclusion and Discussion

Motivation

- Real-life data sets are evolving over time:
 - new images or items are added every second
 - new labels, tags and products are incorporated over time
 - for example, Flickr, Twitter, Facebook, Amazon. . .
- Need to index, retrieve, search and categorize these classes
- Related publications
 - *Metric learning for large scale image classification: Generalizing to new classes at near-zero cost*, Mensink, Verbeek, Perronnin & Csurka, ECCV 2012
 - *Large scale metric learning for distance-based image classification*, Mensink, Verbeek, Perronnin & Csurka, submitted

Related Work

- Recent focus on large-scale image classification
 - ImageNet data set [1,2]
 - ▶ currently over 14 million images
 - ▶ multi-class with 20 thousand classes
- Good performance is usually obtained by using:
 - High dim. features: Super Vector [3] & Fisher Vector [4]
 - Linear 1-vs-Rest SVM classifiers
 - Stochastic Gradient Descent training

1. Deng *et al.*, ImageNet: A large-scale hierarchical image database, CVPR'09
2. Deng *et al.*, What does classifying 10,000 image categories tell us?, ECCV'10
3. Lin *et al.*, Large-scale image classification: Fast feature extraction, CVPR'11
4. Perronnin *et al.*, Good practice in large-scale image classification, CVPR'12

Adapting to new images and classes

- Limitations of 1-vs-Rest SVM for open-ended data sets:
 - Continued training when **new images** become available
 - For a **new class** training starts from scratch

Our approach:

- Distance based classifiers:
 - k-Nearest Neighbors
 - Nearest Class Mean Classification
- Trivial addition of new images or new classes
- Critically depends on the distance function
 - Introduce new metric learning approach for NCM

Nearest Class Mean Classifier



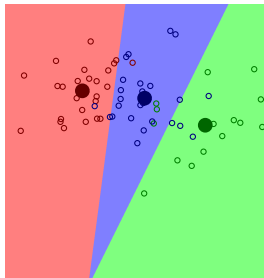
Nearest Class Mean Classifier

- Represent each class by its mean

$$\mu_c = \frac{1}{N_c} \sum_{i:y_i=c} \mathbf{x}_i$$

- Assign an image i to the class with the closest class mean

$$c^* = \underset{c}{\operatorname{argmin}} d(\mathbf{x}, \mu_c)$$



- ✓ Very fast at test time: linear model
- ✓ Easy to integrate new images
- ✓ Easy to integrate new classes
- ✗ Class only represented with mean, less expressive than k-NN

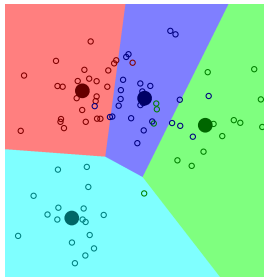
Nearest Class Mean Classifier

- Represent each class by its mean

$$\mu_c = \frac{1}{N_c} \sum_{i:y_i=c} \mathbf{x}_i$$

- Assign an image i to the class with the closest class mean

$$c^* = \operatorname{argmin}_c d(\mathbf{x}, \mu_c)$$



- ✓ Very fast at test time: linear model
- ✓ Easy to integrate new images
- ✓ Easy to integrate new classes
- ✗ Class only represented with mean, less expressive than k-NN

NCM – Probabilistic Interpretation

- Define multinomial probability distribution over classes
- We use the soft-min formulation

$$p(c|\mathbf{x}) = \frac{\exp -d(\mathbf{x}, \boldsymbol{\mu}_c)}{\sum_{c'=1}^C \exp -d(\mathbf{x}, \boldsymbol{\mu}_{c'})}$$

- Corresponds to class posterior in generative model
 - $p(\mathbf{x}|c) = \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_c, \Sigma)$, with shared covariance matrix

NCM – Metric Learning

- Replace the Euclidian distance $d(\mathbf{x}, \mu_c)$
- Use Mahalanobis distance, parametrized by W

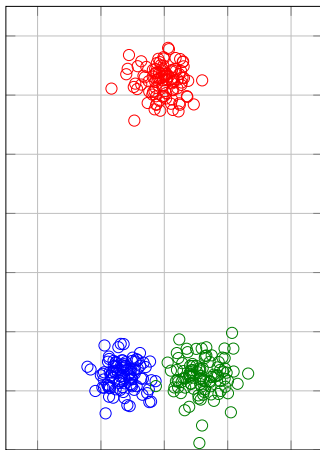
$$d_W(\mathbf{x}, \mu_c) = (\mathbf{x} - \mu_c)^\top W^\top W(\mathbf{x} - \mu_c)$$

- Learn low-rank projection matrix $W : m \times D$
- Discriminative maximum likelihood training:

$$\mathcal{L}(W) = \sum_{i=1}^N \ln p(c_i | \mathbf{x}_i)$$

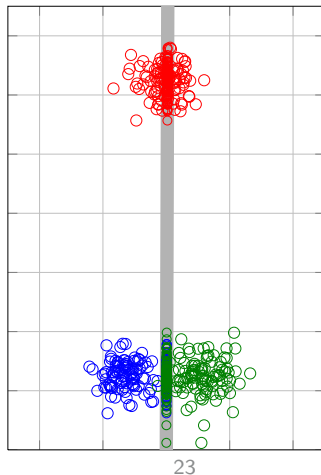
Comparison to FDA

- Three not linearly separable classes
 - Find best projection in 1 dimension



Comparison to FDA

- Classical Fisher Discriminant
 - maximizes variance between all class means



Comparison to FDA

- Our proposed metric learning approach
 - maximizes variance between nearby class means

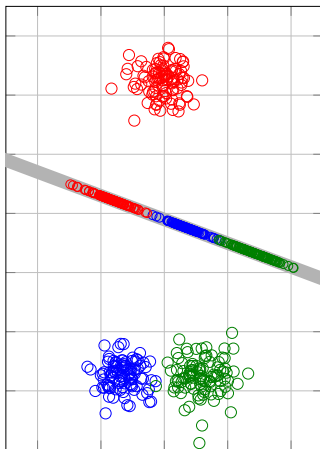


Illustration of Learned Distances



L2								
	crane		stupa		roller coaster		bell cote	
Mah.								
	cabbage tree		pine		pandanus		iron tree	

Relation to other linear classifiers

$$f_c(\mathbf{x}) = b_c + \mathbf{w}_c^\top \mathbf{x}$$

■ Linear SVM

- Learn $\{b_c, \mathbf{w}_c\}$ per class

■ WSABIE [1]

- $\mathbf{w}_c = \mathbf{v}_c W$ $W \in \mathbf{R}^{m \times D}$
- Learn $\{\mathbf{v}_c\}$ per class and shared W

■ Nearest Class Mean

- $b_c = \|W\boldsymbol{\mu}_c\|_2^2$, $\mathbf{w}_c = -2(\boldsymbol{\mu}_c^\top W^\top W)$
- Learn shared W

1. Weston *et al.*, Scaling up to large vocabulary image annotation, IJCAI'11

Experimental Evaluation



Experimental Evaluation

■ Data sets:

- ILSVRC'10: 1.2M training images, 1,000 classes
- ImageNet-10K: 4.5M training images, 10K classes

■ Image features:

- 4K and 64K dimensional Fisher Vectors
- PQ Compression on 64K features [1]
 - ▶ Reduces memory usage from 320GB to 10GB

■ Training:

- Stochastic Gradient Descent
- Learning rate and early stopping set by validation set

1. Jégou *et al.*, Product quantization for nearest neighbor search, PAMI'11

ILSVRC'10 - Top 5 Accuracy

- k-NN & NCM improve with metric learning
- NCM outperforms more flexible k-NN

	Eucl	Mahalanobis		
Dimensionality	4K	256	512	1024
k-NN, LMNN [1] - dynamic	44.1	61.0	60.9	59.6
NCM, learned metric	32.0	62.6	63.0	63.0

1. Weinberger & Saul, Distance Metric Learning for LMNN Classification, JMLR'09

ILSVRC'10 - Top 5 Accuracy

- k-NN & NCM improve with metric learning
- NCM outperforms more flexible k-NN
- Distance based classifiers competitive with SVMs

	Eucl	Mahalanobis		
Dimensionality	4K	256	512	1024
k-NN, LMNN [1] - dynamic	44.1	61.0	60.9	59.6
NCM, learned metric	32.0	62.6	63.0	63.0
WSABIE [2]		60.6	61.3	61.5

- Baseline: 1-vs-Rest SVM **61.8**

1. Weinberger & Saul, Distance Metric Learning for LMNN Classification, JMLR'09
2. Weston *et al.*, Scaling up to large vocabulary image annotation, IJCAI'11

Generalization to Novel Classes



Generalization to Novel Classes

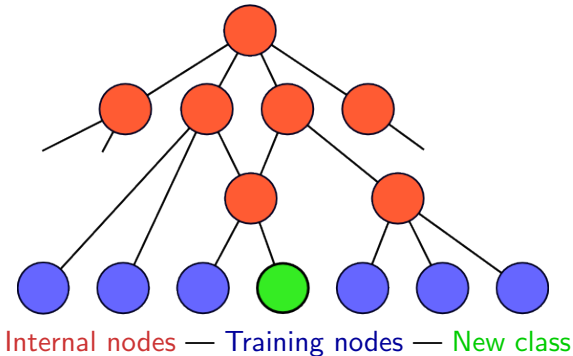
- Nearest Class Mean Classifier
 - Compute means of ImageNet-10K classes: \pm **1 CPU hour**
 - Re-use metric learned on ILSVRC'10
- 1-vs-Rest SVM baseline
 - Train 10K SVM classifiers: \pm **280 CPU days**
- NCM is faster by a factor of **8500!**

Feat. dim.	64K		21K	128K	128K
Method	NCM	SVM	SVM [1]	SVM [2]	SVM [3]
Top-1	13.9	21.9	6.4	16.7	18.1

1. Deng *et al.*, What does classifying 10,000 image categories tell us?, ECCV'10
2. Sánchez and Perronnin, High-dimensional signature compression, CVPR'11
3. Perronnin *et al.*, Good practice in large-scale image classification, CVPR'12

Transfer Learning - Zero-Shot Prior

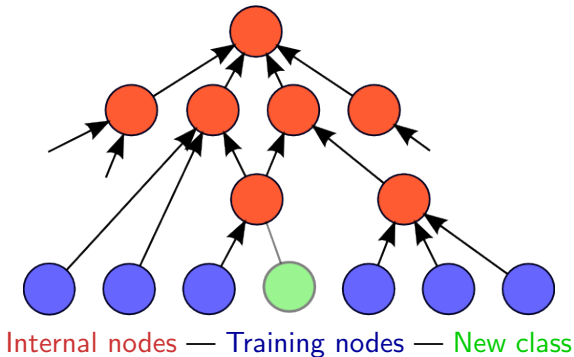
- Use ImageNet class hierarchy to estimate mean of new class [1]



1. Rohrbach *et al.*, Knowledge transfer and zero-shot learning in large-scale, CVPR'11

Transfer Learning - Zero-Shot Prior

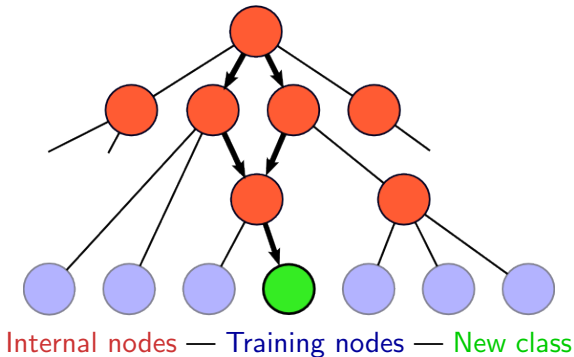
- Use ImageNet class hierarchy to estimate mean of new class [1]



1. Rohrbach *et al.*, Knowledge transfer and zero-shot learning in large-scale, CVPR'11

Transfer Learning - Zero-Shot Prior

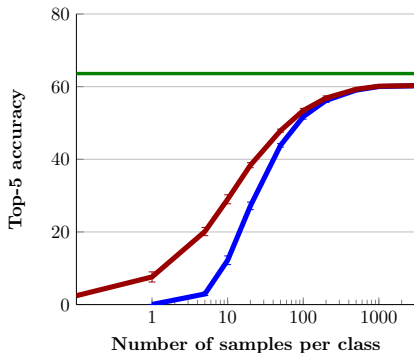
- Use ImageNet class hierarchy to estimate mean of new class [1]



1. Rohrbach *et al.*, Knowledge transfer and zero-shot learning in large-scale, CVPR'11

Transfer Learning - Results ILSVRC'10

- **Step 1** Metric learning on 800 classes
- **Step 2** Estimate means for remaining 200 for evaluation:
 - Data mean per class
 - Zero-Shot prior + data mean per class
 - Baseline — trained on all 1000 classes



Conclusion

- Nearest Class Mean Classification
 - We proposed metric learning by maximum-likelihood
 - Outperforms more flexible k-NN, on par with SVM
- Advantages of NCM over 1-vs-Rest SVMs
 - Allows adding new images and classes at near zero cost
 - Shows competitive results on unseen classes
 - Can benefit from class priors for small sample sizes
- More details in thesis
 - Extension using multiple class centroids
 - Different learning objectives to speed up training
 - Analysis on convergence of low-rank formulation



Outline

1. Introduction
2. Large Scale Classification and Adapting to Novel Classes
3. Leverage User Interaction using Label Dependencies
4. Exploiting Multi-Modal Data
5. Conclusion and Discussion

Motivation

- Multi-label image classification
- Interactive annotation to trade-off between
 - Fully automatic annotation – cheap but low accuracy
 - Fully manual annotation – expensive and high accuracy
- Need a structure to benefit from labels set by the user

- Related publications
 - *Learning structured prediction models for interactive image labeling*, Mensink, Verbeek & Csurka, CVPR 2011
 - *Tree-structured CRF models for interactive image labeling*, Mensink, Verbeek & Csurka, PAMI 2012
 - *Learning to rank and quadratic assignment*, Mensink, Verbeek & Caetano, DISCML'12

Related work

- Automatic image annotation
 - 1-vs-Rest SVM Classifiers [1]
 - Nearest Neighbors approaches [2]
 - Image ranking [3,4]
- No explicit modeling of label dependencies
- Interactive multi-class classification [5]

Our approach:

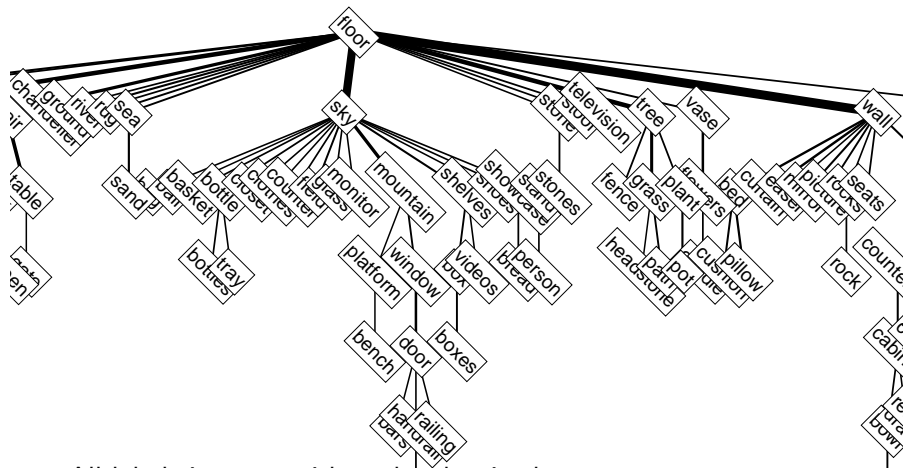
- Model explicitly structure in labels
- Interactive labeling scenario

1. Everingham *et al.*, The PASCAL Visual Object Classes Challenge 2007-2011
2. Guillaumin *et al.*, TagProp: metric learning in nearest neighbor models, ICCV'09
3. Grangier and Bengio, Kernel-based model to rank images from text queries, PAMI'08
4. Weston *et al.*, Learning to rank with joint word-image embeddings, ECML'10
5. Branson *et al.*, Visual Recognition with Humans in the Loop, ECCV'10

Tree structure over class labels



Tree structure over class labels



- All labels interact with each other in the structure
- Allows for efficient and exact inference

Tree structured model

- Vector of (binary) labels: $\mathbf{y} = \{y_1, \dots, y_L\}$
- Energy between a specific labeling and the image

$$E(\mathbf{y}, \mathbf{x}; \mathbf{w}, \mathbf{v}) = \sum_{i=1}^L \psi_i(y_i, \mathbf{x}; \mathbf{w}) + \sum_{(i,j) \in \mathcal{E}} \psi_{ij}(y_i, y_j; \mathbf{v})$$

- Gibbs distribution for a specific configuration \mathbf{y} :

$$p(\mathbf{y}|\mathbf{x}) = \frac{1}{Z(\mathbf{x})} \exp -E(\mathbf{y}, \mathbf{x}; \mathbf{w}, \mathbf{v})$$

- Belief Propagation for label prediction, elicitation and parameter learning

Learning

- Learning $\{\mathbf{w}, v\}$, using log-likelihood:

$$\mathcal{L} = \sum_{n=1}^N \ln p(\mathbf{y}_n | \mathbf{x}_n).$$

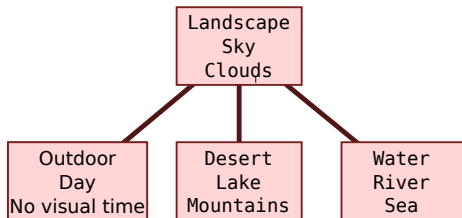
- Energy is linear in parameters \rightarrow log-likelihood concave
 - Maximize log-likelihood with gradient ascent
-
- Obtaining the tree structure
 - Finding optimal tree for conditional models is intractable [1]
 - Optimal for generative models: the Chow-Liu algorithm [2]

-
1. Bradley and Guestrin, Learning tree conditional random fields, ICML'10
 2. Chow and Liu, Approximating probability distributions with trees, IT'68

Extensions

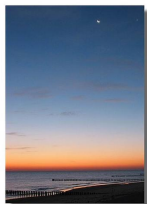


Extension 1 — Trees over groups of labels



- Increase the discriminative power by using nodes with k labels
- Every state in a node is modeled explicitly
 - Each node has 2^k states
 - Label marginals read-off from state marginal table
- Trade-off: model expressiveness vs computational complexity

Extension 1 — Trees over groups of labels



State	Marginal	Landscape	Sky	Clouds
1	3.4 %	0	0	0
2	0.0 %	0	0	1
3	9.8 %	0	1	0
4	59.9 %	0	1	1
5	0.4 %	1	0	0
6	0.0 %	1	0	1
7	2.6 %	1	1	0
8	23.9 %	1	1	1
Label marginal		26.9%	96.2%	83.8%

- Increase the discriminative power by using nodes with k labels
- Every state in a node is modeled explicitly
 - Each node has 2^k states
 - Label marginals read-off from state marginal table
- Trade-off: model expressiveness vs computational complexity

Extension 2 — Mixture-of-trees

- Learn multiple trees
 - different group sizes k
 - different structures over a fixed set of nodes
- Define a mixture of T trees as


$$p(\mathbf{y}|\mathbf{x}) = \sum_{t=1}^T \pi_t p_t(\mathbf{y}|\mathbf{x})$$

- Learning a mixture-of-trees
 - Each tree is learned independently
 - Mixing weights are set uniformly

Interactive image labeling

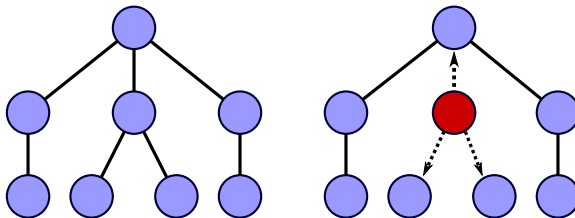


Interactive image labeling

29 labels	Before	Questions	After
	Fast		Toughskin
	Active		Swims
	Smart		Arctic
	Meatteeth	Toughskin ✓	Water
	Newworld	Paws ✗	Fish
	Agility	Swims ✓	Ocean
	Tail	Mountains ✗	Fast
	Meat	Arctic ✓	Active
	Strong		Strong
	Chewteeth		Smart

- Ask the user at **test** time to set a few labels
 - To improve the prediction performance
- Iterative strategy: ask one label at the time

Label elicitation



- Determine which label should be set by the user
 - Objective: select label y_i to minimize expected uncertainty of the remaining labels $H(\mathbf{y}_{\setminus i} | y_i, \mathbf{x})$
- After label is set, update predictions on other labels
 - Information propagated in tree is now combination of visual information and user-provided information

Experimental evaluation



Experimental evaluation

■ Data sets

- ImageClef VCDT 2010 Challenge (imageClef) [1]
- Scene Understanding (SUN'09) [2]
- Animals with Attributes (AwA) [3]

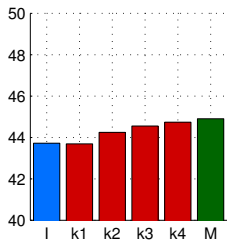
■ Performance: mean average precision (MAP)

- retrieval performance per label

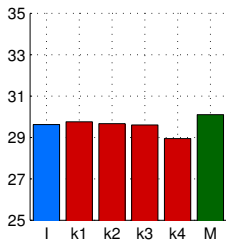
-
1. Nowak and Huiskes, New strategies for image annotation, ImageCLEF'10
 2. Choi *et al.*, Hierarchical context on a large database of object categories, CVPR'10
 3. Lampert *et al.*, Detect unseen object classes by between-class attribute transfer, CVPR'09

Results - Fully Automatic Labeling

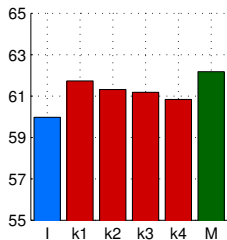
- Baseline is state-of-the-art for ImageClef'10 and SUN'09.



ImageCLEF'10



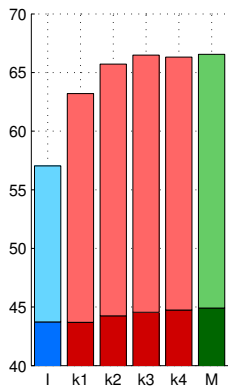
SUN09



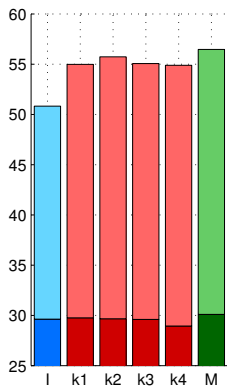
AwA

Results - Interactive Labeling

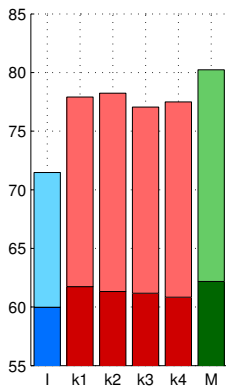
- 10 labels are asked and set by the user



ImageCLEF'10



SUN09



AwA

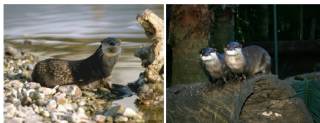
Attribute-based image classification



Attribute-based image classification

otter

black: yes
white: no
brown: yes
stripes: no
water: yes
eats fish: yes



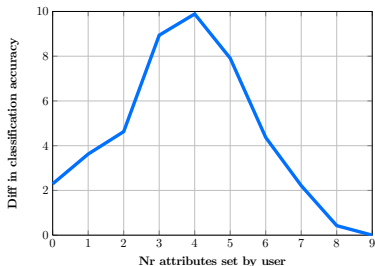
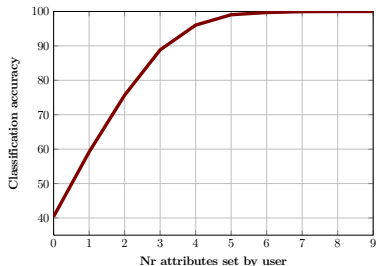
polar bear

black: no
white: yes
brown: no
stripes: no
water: yes
eats fish: yes



- Classes are defined by a given set of attributes
- Zero-shot: Images of 10 test classes not used for training
- Use structured models for attribute prediction
- Ask for attribute values to improve class prediction

Results Attribute-based image classification



- Any useful question eliminates at least 1 class
 - Never more than 9 questions needed

Conclusions

- Structured models for predicting image labels
 - Image annotation, and
 - Attribute-based classification.
- Mixture-of-trees is a powerful yet tractable structured model
 - Efficiently transfers knowledge of labels set by the user
 - Allows to ask relevant labels to set by user
- More details in thesis
 - Comparison of joint and stage learning of unary potentials
 - Alternatives to obtain tree structures
 - Learning to rank with pairwise label interactions



Outline

1. Introduction
2. Large Scale Classification and Adapting to Novel Classes
3. Leverage User Interaction using Label Dependencies
4. Exploiting Multi-Modal Data
5. Conclusion and Discussion

Motivation

- Different modalities as weak form of supervision
 - Multiple modalities relatively cheap to obtain
 - Combining text and image retrieval improves performance [1]
-
- Related publications
 - *Transmedia Relevance Feedback for Image Autoannotation*, Mensink, Verbeek & Csurka, BMVC 2010
 - *Weighted Transmedia Relevance Feedback for Image Retrieval and Autoannotation*, Mensink, Verbeek & Csurka, Tech Report 2011

1. ImageCLEF Photo Retrieval Challenge 2006-2010

Related Work

- Late Fusion - Combine mono-modal similarities
 - ✓ Well studied problems, and well engineered solutions [1,2]
 - ✗ Unable to exploit the correlations between different modalities

- Early Fusion - Joint representation of different modalities
 - ✓ Exploit the correlations
 - ✗ Representation should allow for heterogeneity of the modalities
 - variations in semantic meaning (words vs. low level)
 - histogram concatenation, or topic models [3]

- Intermediate Fusion
 - Transmedia relevance feedback

-
1. Manning *et al.*, Introduction to information retrieval, CUP'08
 2. Datta *et al.*, Image retrieval: Ideas, influences, and trends of the new age, ACM'08
 3. Barnard *et al.*, Matching words and pictures, JMLR'03.

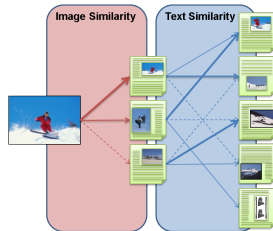
Transmedia Relevance Feedback

■ Pseudo-Relevance Feedback [1]

- Textual query
- Extract keywords from top k documents retrieved
- New query: query + extracted keywords

■ Transmedia Relevance Feedback [2,3]

- Visual query
- Rank using visual similarity
- Swap modality
- New query: textual description from top k documents



1. Salton and Buckley, Improving retrieval performance by relevance feedback, ASIS'90
2. Chang and Chen, Using a word-image ontology for image retrieval, CLEF'06
3. Clinchant *et al.*, XRCEs participation to ImageCLEFphoto 2007, CLEF'07

Weighted Relevance Feedback

- Equal Weighted [1]

$$s(q, d) = \sum_{k=1}^K s_1(q, d_k) s_2(d_k, d)$$

- Linear Weighted

$$s(q, d; \gamma) = \sum_{k=1}^K \gamma_k s_1(q, d_k) s_2(d_k, d)$$

- Constrain γ_k to be positive and ordered

- Softmax Weighted

$$s(q, d; \gamma) = \sum_{k=1}^K \tilde{s}_1(q, d_k; \gamma) s_2(d_k, d)$$

- with $\tilde{s}_1(q, d_k; \gamma) \propto \exp(\gamma s_1(d_k, q))$
- Positive and ordered by construction

1. Ah-Pine *et al.*, Leveraging image, text and cross-media similarities, Springer 2010.

Learning Retrieval Functions

- Combine a set of (visual, textual, and transmedia) distances

$$f(q, d) = \sum_i w_i s_i(q, d; \gamma_i)$$

- Learn parameters $\{\mathbf{w}, \gamma\}$ using comparative classification [1]
 - score relevant document higher than negative document

- Correcting for inter-query variations

$$f'(q, d) = \alpha_q f(q, d) + \beta_q$$

- Difference in distribution scores
- Ranking is independent of $\{\alpha_q, \beta_q\}$

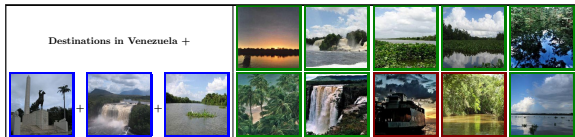
1. Joachims, SVMs for multivariate performance measures, ICML'05

Experimental evaluation



Experimental evaluation – Image retrieval

■ ImageClef'08 Retrieval Challenge



■ Comparison to participants

Method	MAP	P@20
AVEIR	31.8	43.5
UP-GPLSI	33.0	43.1
DCU	35.1	47.6
XRCE	41.1	57.3
Ours - 2 comp	42.7	59.7
Ours - 6 comp	43.1	59.9

- 2-Comp: combination of text and image-to-text distances

Experimental evaluation – Image annotation

- TagProp [1] is a weighted nearest neighbor labeling approach
 - Learns a weighting of different visual distances
 - Nearest neighbors also based on transmedia distance
 - Transmedia parameters learning integrated in TagProp
- Performance measured in MAP and iMAP
 - MAP measures keyword based retrieval performance
 - iMAP measures annotation performance
- Annotation results

Dataset	Method	MAP	iMAP
Corel-5K	TagProp	36.0	54.2
	TagProp + Transmedia	38.1	55.6
IAPR-TC12	TagProp	35.4	47.0
	TagProp + Transmedia	35.9	48.0

1. Guillaumin *et al.*, TagProp: metric learning in nearest neighbor models, ICCV'09

Conclusions

- Transmedia relevance for combining modalities
 - Defines true multi-modal distance, *e.g.* from visual to text
 - Parametrized version allow to learn parameters from data
- Multi-modal image retrieval
 - Query dependent variables to learn better parameters
- Image Annotation
 - Transmedia distance incorporated into TagProp
- More details in thesis:
 - Comparison of learning objectives for retrieval
 - Comparison of different modalities for annotation



Outline

1. Introduction
2. Large Scale Classification and Adapting to Novel Classes
3. Leverage User Interaction using Label Dependencies
4. Exploiting Multi-Modal Data
5. Conclusion and Discussion

Conclusions

Goals

1. Scaling to large data sets
2. Adapting to novel classes
3. Leverage user interaction
4. Modeling label dependencies
5. Exploiting multi-modal data

Conclusions

Goals

1. Scaling to large data sets
2. Adapting to novel classes

Contributions

- Metric learning approach for Nearest Class Mean classifier
 - On par with state-of-the-art linear SVMs
 - Generalizes well to unseen classes
 - Proven scalability to data sets with millions of images

Conclusions

Goals

3. Leverage user interaction
4. Modeling label dependencies

Contributions

- Mixtures-of-trees to model label dependencies
 - Moderate improvement in fully automatic setting
 - Efficiently leverages user interaction
 - Shown to work in different scenarios: image annotation and attribute-based classification

Conclusions

Goals

5. Exploiting multi-modal data

Contributions

- Parametrized transmedia relevance feedback
 - Effective way to combine multiple modalities
 - Learned parameters outperform manually set ones
 - Validated on multi-modal image retrieval and image annotation

Future work



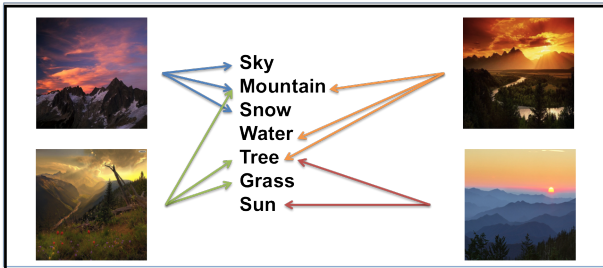
Future work

Active and interactive learning of annotation models

- Learn classifiers and annotate large evolving set of images
- Balance the accuracy versus the annotation cost
 - For a label: select informative images to learn classifier
 - For a image: select informative labels
- Research questions
 1. Select between active and interactive
 2. Model label dependencies in an evolving set of user labels
 3. Incentives for high quality user input

Future work

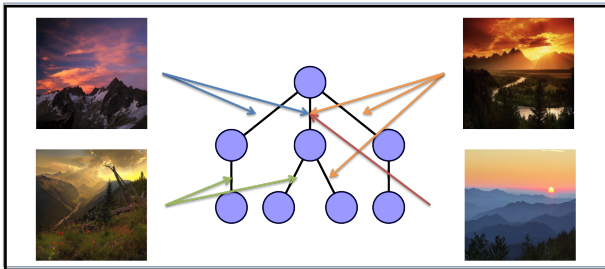
Structured-prediction in non-parametric models



- Combine kNN with structured models
 - Propagate pairwise marginals to the test image
- Research questions
 1. How to define structure over labels
 2. How to locally adapt the structure

Future work

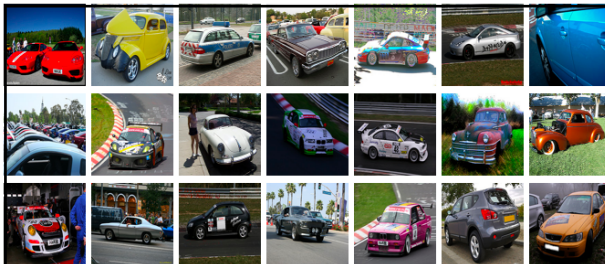
Structured-prediction in non-parametric models



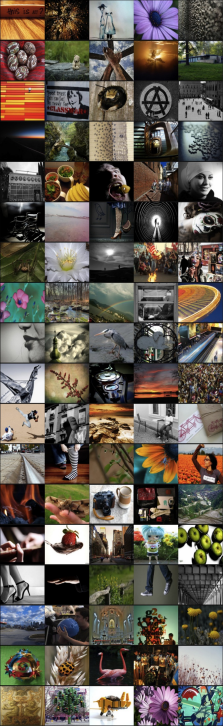
- Combine kNN with structured models
 - Propagate pairwise marginals to the test image
- Research questions
 1. How to define structure over labels
 2. How to locally adapt the structure

Future work

Address visual diversity in image classification



- Assumption: semantic class is a single coherent visual concept
- Richer class representations
 - using unsupervised discovery, e.g. Latent-SVM
 - using ideas of “visual phrases”



Learning Image Classification and Retrieval Models

Thomas Mensink

26 October 2012