Tree structured CRF models for interactive image labeling

Thomas Mensink[1,2] Gabriela Csurka[1] **Jakob Verbeek**[2]

[1]Xerox Research Centre Europe, Grenoble, France

[2]INRIA, Grenoble, France

To appear at CVPR 2011

Outline

1. Introduction
2. Structured image annotation models
3. Label Elicitation
4. Experimental Evaluation
5. Attribute-based image classification

Interactive Image labeling

Interactive Image labeling

- Sky, Tree, Building, Sea, Plant, Ground, Rock, Person, Windows, Sand, Water.

Interactive Image labeling

- Sky, Tree, Building, Sea, Plant, Ground, Rock, Person, Windows, Sand, Water.
- Ask the user: Building (**false**), Rock (**true**), Sea (**true**), ...

Interactive Image labeling

- Sky, Tree, Building, Sea, Plant, Ground, Rock, Person, Windows, Sand, Water.
- Ask the user: Building (**false**), Rock (**true**), Sea (**true**), ...
- Update the ranked list of keywords based on this information

Introduction - 1

- Image labeling problem, a.k.a. classification, annotation, attribute prediction,. . .
- Used for *e.g.*: keyword based retrieval, indexing , clustering, . . .
- State of the art: **train binary SVMs per label** using fancy features (SIFT, Bow, Fisher Kernels, spatial pyramids, **...**)

Introduction - 1

- Image labeling problem, a.k.a. classification, annotation, attribute prediction,. . .
- Used for *e.g.*: keyword based retrieval, indexing , clustering, . . .
- State of the art: **train binary SVMs per label** using fancy features (SIFT, Bow, Fisher Kernels, spatial pyramids, ...)
- **Problem 1**: it ignores structure in output, correlation between labels (*e.g.* car & indoor).
- **Problem 2**: how to incorporate user input

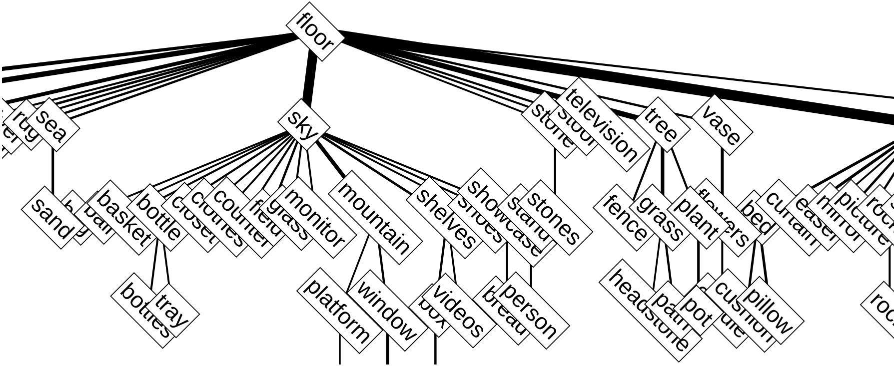
Introduction - 2

- How to obtain a (tractable) structure?
- How to learn the parameters of this structure?
- How to select labels to ask the user?
- How does it perform?

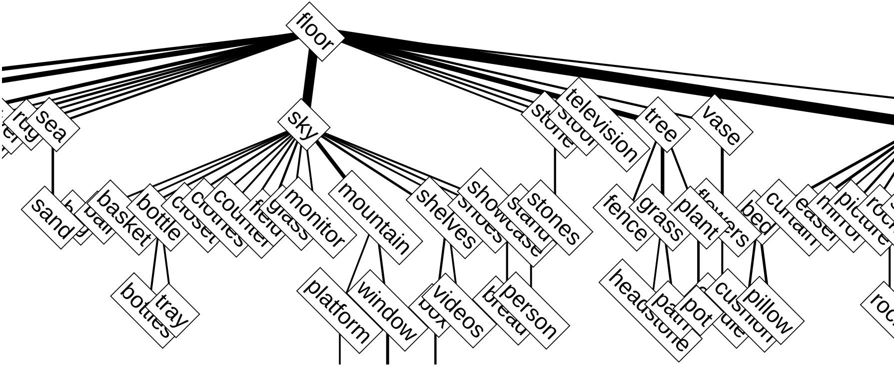
Outline

1. Introduction
2. Structured image annotation models
3. Label Elicitation
4. Experimental Evaluation
5. Attribute-based image classification

Tree Structures

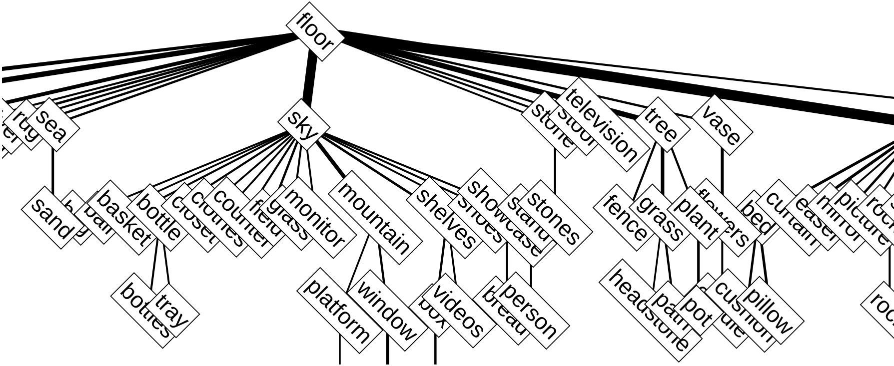


Tree Structures



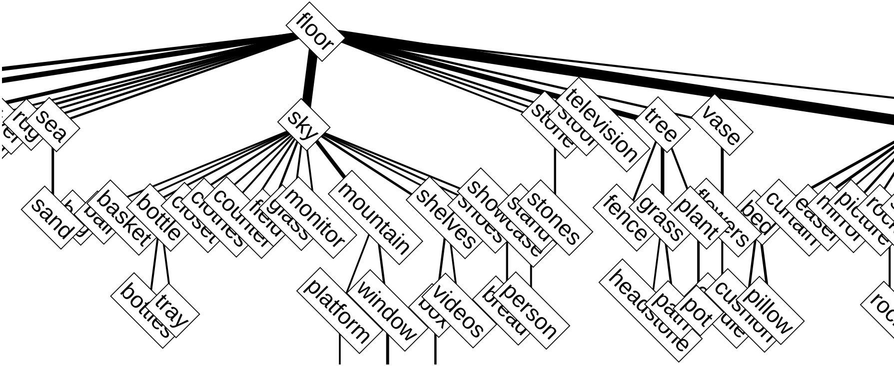
- Nodes are (class/category/attributes) labels.

Tree Structures



- Nodes are (class/category/attributes) labels.
- Learn weights between nodes to encode co-occurrence.

Tree Structures



- Nodes are (class/category/attributes) labels.
- Learn weights between nodes to encode co-occurrence.
- Exact inference in tree structure is tractable (using BP).
- Inference is used for learning, label prediction and label elicitation.

Tree structured model on image labels

- Each node presents a label in the tree.
- Vector of (binary) labels: $\boldsymbol{y} = \{y_1, \ldots, y_L\}$.
- Edges (L-1) are (somehow) given: $\mathcal{E} = \{e_1, \ldots, e_{L-1}\}$.

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

$$p(\boldsymbol{y}|\boldsymbol{x}) = \frac{1}{Z(\boldsymbol{x})} \exp -E(\boldsymbol{y}, \boldsymbol{x}), \tag{2}$$

$$Z(\boldsymbol{x}) = \sum_{\boldsymbol{y}\in\{0,1\}^L} \exp -E(\boldsymbol{y}, \boldsymbol{x}) \tag{3}$$

Unary Potentials

$$E(\boldsymbol{y}, \boldsymbol{x}) = \underbrace{\sum_{i=1}^{L} \psi_i(y_i, \boldsymbol{x})}_{\text{Unary Potentials}} + \sum_{(i,j) \in \mathcal{E}} \psi_{ij}(y_i, y_j)$$

- y_i is a label Rock, Sea, City, People,. . .
- $\psi_i(y_i = l, \boldsymbol{x}) = [\phi_i(\boldsymbol{x}), 1]^\top \boldsymbol{w}_i^l$
- $\phi_i(\boldsymbol{x})$: Pre-trained SVM score for label i

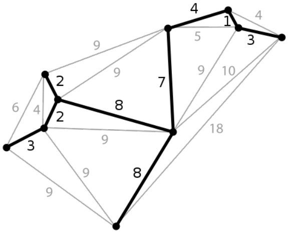
Pairwise Potentials

$$E(\boldsymbol{y}, \boldsymbol{x}) = \sum_{i=1}^{L} \psi_i(y_i, \boldsymbol{x}) + \underbrace{\sum_{(i,j)\in\mathcal{E}} \psi_{ij}(y_i, y_j)}_{\text{Pairwise Potentials}}$$

- $y_i =$ Sand, and $y_j =$ City
- Independent of image input
- $\psi_{ij}(y_i = s, y_j = t) = v_{ij}^{st}$

Defining the Tree

- Optimal tree structure for conditional models is intractable
- For generative models use the Chow-Liu algorithm



- Fully connected graph
- Edge weight = Mutual Information
- Maximum Spanning Tree

Learning

- Learning $\boldsymbol{w}$ and v in unary and pairwise potentials
- Using Log-likelihood (concave):

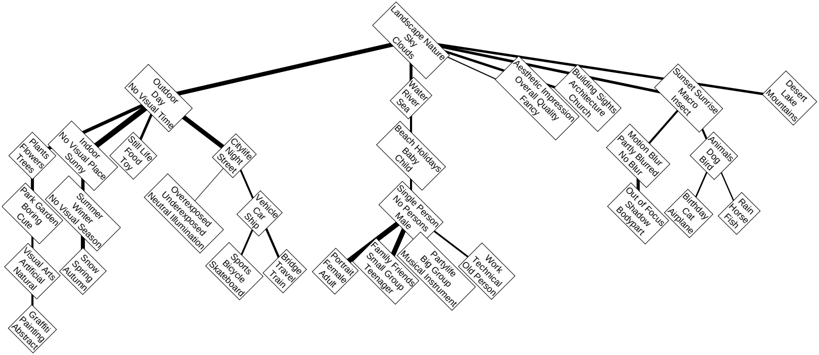
$$\mathcal{L} = \sum_{n=1}^{N} \mathcal{L}_n = \sum_{n=1}^{N} \ln p(\boldsymbol{y}_n | \boldsymbol{x}_n).$$

- Gradients:

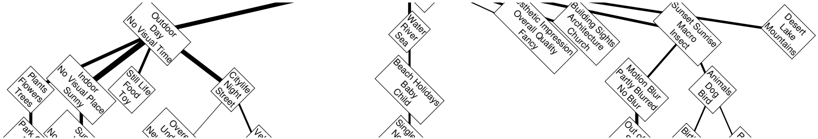
$$\frac{\partial \mathcal{L}_n}{\partial \boldsymbol{w}_i^l} = \Big(p(y_i = l | \boldsymbol{x}_n) - [\![y_{in} = l]\!] \Big) \phi_i(\boldsymbol{x}_n), \tag{4}$$

$$\frac{\partial \mathcal{L}_n}{\partial v_{ij}^{st}} = p(y_i = s, y_j = t | \boldsymbol{x}_n) - [\![y_{in} = s, y_{jn} = t]\!], \tag{5}$$

Trees over groups of labels



Trees over groups of labels



- To allow more dependencies between labels
- A node is a group of fully connected labels.
- Every state modeled explicitly, a node has 2^k states.

- To define a tree-structure
  - Agglomerative clustering of labels,
  - Chow-Liu algorithm on these clusters.

Compund Node

State	Marginal	Landscape/Nature	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
Marginal on label = true		26.9%	96.2%	83.8%

- BP gives us node marginals,
- read-off label marginals $p(y_i|\boldsymbol{x})$.
- message passing: $O(2^{2k})$

Outline

1. Introduction
2. Structured image annotation models
3. Label Elicitation
4. Experimental Evaluation
5. Attribute-based image classification

Label Elicitation

SUN 09 - 5 labels	Before	Questions
	01 Sky 02 Tree 03 Building 04 Sea 05 Rocks 06 Plant 07 Ground 08 Rock 09 Person 10 Window	

Label Elicitation

SUN 09 - 5 labels

Before	Questions
01 Sky	
02 Tree	
03 Building	
04 Sea	Building
05 Rocks	
06 Plant	
07 Ground	
08 Rock	
09 Person	
10 Window	

Label Elicitation

SUN 09 - 5 labels

Before	Questions
01 Sky	
02 Tree	
03 Building	
04 Sea	Building
05 Rocks	Tree
06 Plant	
07 Ground	
08 Rock	
09 Person	
10 Window	

Label Elicitation

SUN 09 - 5 labels	Before	Questions
	01 Sky	
	02 Tree	
	03 Building	
	04 Sea	Building
	05 Rocks	Tree
	06 Plant	Sea
	07 Ground	
	08 Rock	
	09 Person	
	10 Window	

Label Elicitation

SUN 09 - 5 labels	Before	Questions
	01 Sky	
	02 Tree	
	03 Building	
	04 Sea	Building
	05 Rocks	Tree
	06 Plant	Sea
	07 Ground	Rocks
	08 Rock	
	09 Person	
	10 Window	

Label Elicitation

SUN 09 - 5 labels

Before	**Questions**
01 Sky	
02 Tree	
03 Building	
04 Sea	Building
05 Rocks	Tree
06 Plant	Sea
07 Ground	Rocks
08 Rock	Rock
09 Person	
10 Window	

Label Elicitation

SUN 09 - 5 labels

Before	**Questions**	**After**
01 Sky		01 Rock
02 Tree		02 Rocks
03 Building		03 Sea
04 Sea	Building	04 Sky
05 Rocks	Tree	05 Sand
06 Plant	Sea	06 Ground
07 Ground	Rocks	07 Plant
08 Rock	Rock	08 Person
09 Person		09 Window
10 Window		10 Water

Label Elicitation

- **interactive setting:** Ask the user at *test* time to set some of many labels for a single example.

- **active learning:** Ask the user at *train* time for class label of some of many examples.

Label Elicitation

- Select a label i such that expected uncertainty in remaining labels is minimized:

$$H(\boldsymbol{y}_{\setminus i}|y_i, \boldsymbol{x}) = \sum_l p(y_i = l|\boldsymbol{x}) H(\boldsymbol{y}_{\setminus i}|y_i = l, \boldsymbol{x}).$$

- Entropy Identity:

$$H(\boldsymbol{y}|\boldsymbol{x}) = H(y_i|\boldsymbol{x}) + H(\boldsymbol{y}_{\setminus i}|y_i, \boldsymbol{x})$$

- Equals to select label i with highest entropy $H(y_i|\boldsymbol{x})$.

Outline

1. Introduction
2. Structured image annotation models
3. Label Elicitation
4. Experimental Evaluation
5. Attribute-based image classification

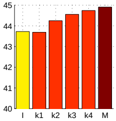
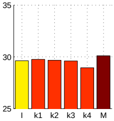
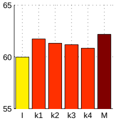
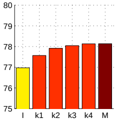
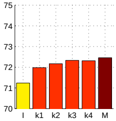
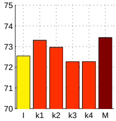
Databases

Table: Basic statistics of the three data sets.

		ImageCLEF	SUN'09	Animals w.A.
# Train images		6400	4367	24295
# Test images		1600	4317	6180
# Labels		93	107	85
Train img/label		833	219	8812
Train label/img		12.1	5.34	30.8
Nr of parameters for	k = 1	± 740	852	676
trees with	k = 2	±1284	1480	1172
group size	k = 3	±2912	3340	2644
	k = 4	±7508	8640	6836

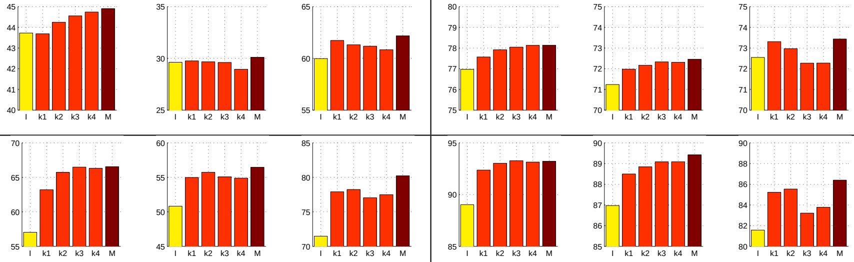
- Performance evaluated using:
  - MAP: retrieval performance per label,
  - iMAP: annotation performance per image.

Results 1

MAP			iMAP		
ImageCLEF'10	SUN09	AwA	ImageCLEF'10	SUN09	AwA
					

Results 1

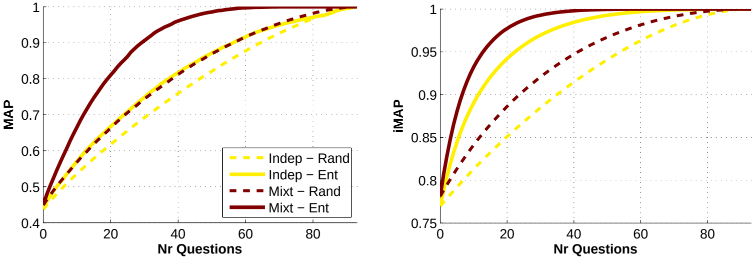
MAP			iMAP		
ImageCLEF’10	SUN09	AwA	ImageCLEF’10	SUN09	AwA



LEAR

xerox

Results 2



- Interactive image annotation performance as a function of the amount of user input, ImageCLEF dataset

Outline

1. Introduction

2. Structured image annotation models

3. Label Elicitation

4. Experimental Evaluation

5. Attribute-based image classification

Attribute-based image classification

otter

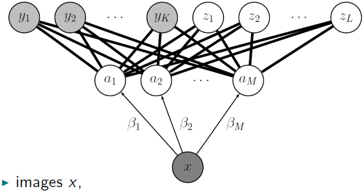
black:	yes
white:	no
brown:	yes
stripes:	no
water:	yes
eats fish:	yes

polar bear

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

zebra

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



- images x,
- class labels $y_1, \ldots, y_K \in \mathcal{Y}$ (at training time)
- class labels $z_1, \ldots, z_L \in \mathcal{Z}$ (at test time)
- attributes $a_1, \ldots, a_M \in \{0,1\}^M$ (encode description)

Attribute-based image classification - 2

- Predict attributes with our tree-structured models.

Attribute-based image classification - 2

- Predict attributes with our tree-structured models.
- Deterministic mapping between attributes and classes.

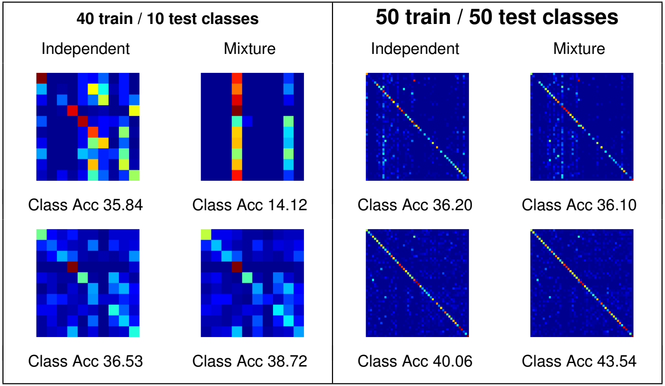
$$p(z=c|\boldsymbol{x}) = \frac{p(\boldsymbol{y}_c|\boldsymbol{x})}{\sum_{c'=1}^{C} p(\boldsymbol{y}_{c'}|\boldsymbol{x})} = \frac{\exp -E(\boldsymbol{y}_c, \boldsymbol{x})}{\sum_{c'=1}^{C} \exp -E(\boldsymbol{y}_{c'}, \boldsymbol{x})}. \quad (6)$$

- Note: does not require belief-propagation, it suffices to evaluate $E(\boldsymbol{y}_c, \boldsymbol{x})$ for the C attribute configurations.

Correction Term

- Observation: some classes are over-predicted:

$$p(z = c|\boldsymbol{x}) \propto \exp\left(-E(\boldsymbol{y}_c, \boldsymbol{x}) - u_c\right), \tag{7}$$



Label Elicitation for classification

- Label Elicitation on Attribute Level

Label Elicitation for classification

- Label Elicitation on Attribute Level
- Goal to minimize uncertainty on class label
- Any informative question rules out at least 1 class.

Label Elicitation for classification

- Label Elicitation on Attribute Level
- Goal to minimize uncertainty on class label
- Any informative question rules out at least 1 class.

- Results (again) in attribute i with highest entropy $H(y_i|\boldsymbol{x})$.
- But $p(y_i|\boldsymbol{x})$ is defined differently:

$$p(y_i = 1|\boldsymbol{x}) = \sum_c p(z = c|\boldsymbol{x}) y_{ic}, \tag{8}$$

LEAR xerox

Results Classification

	Init	1	2	3	4	5	6	7	8
Indep	36.5	53.1	68.5	77.8	85.1	90.6	94.5	97.7	99.4
Mixt	38.7	55.3	72.3	84.8	92.4	96.9	99.0	99.8	100.0

- classification accuracy of the independent and mixture of trees models.
- Initial results, and after user input for one up to eight selected attributes.

Conclusions

- Tree-structured CRF models for interactive
  - Image annotation, and
  - Attribute-based classification

- Improves moderately over independent models
- Real power in interactive setting: (i) propagate user input, (ii) ask more informative questions

Tree structured CRF models for interactive image labeling

Questions?!?