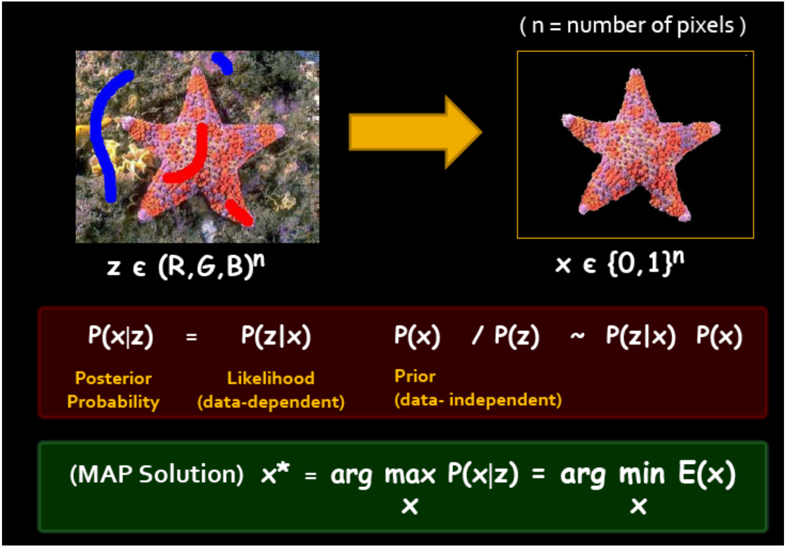
Visual Recognition: Examples of Graphical Models

Raquel Urtasun

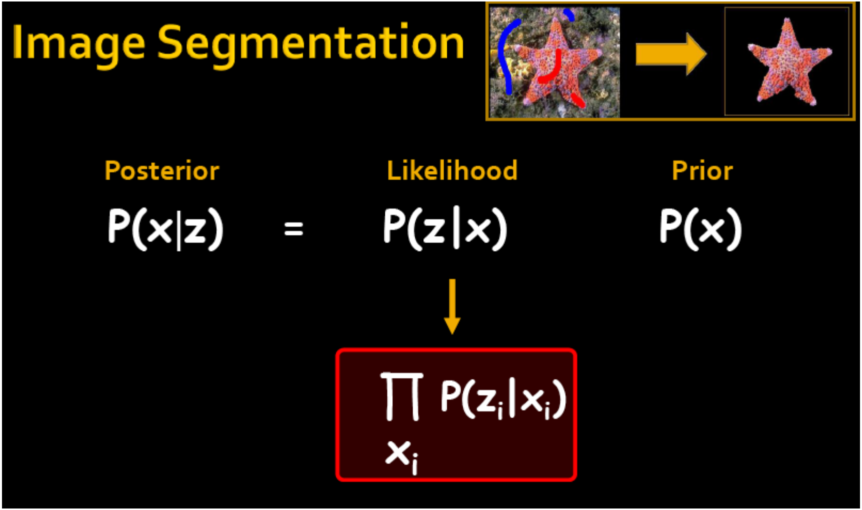
TTI Chicago

March 8, 2012

Example: Segmentation from Scribles

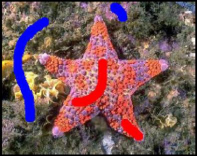


[Source: P. Kohli]

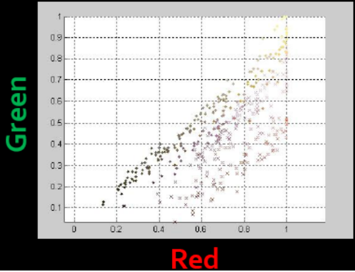


[Source: P. Kohli]

Likelihood $P(x|z) \sim P(z|x)\,P(x)$

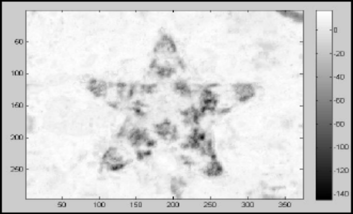


$$P(z|x) = F_{GMM}(z,x)$$

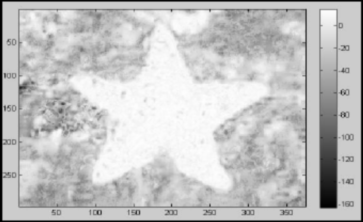


[Source: P. Kohli]

Likelihood $P(x|z) \sim P(z|x)\, P(x)$



Log $P(z_i|x_i=0)$



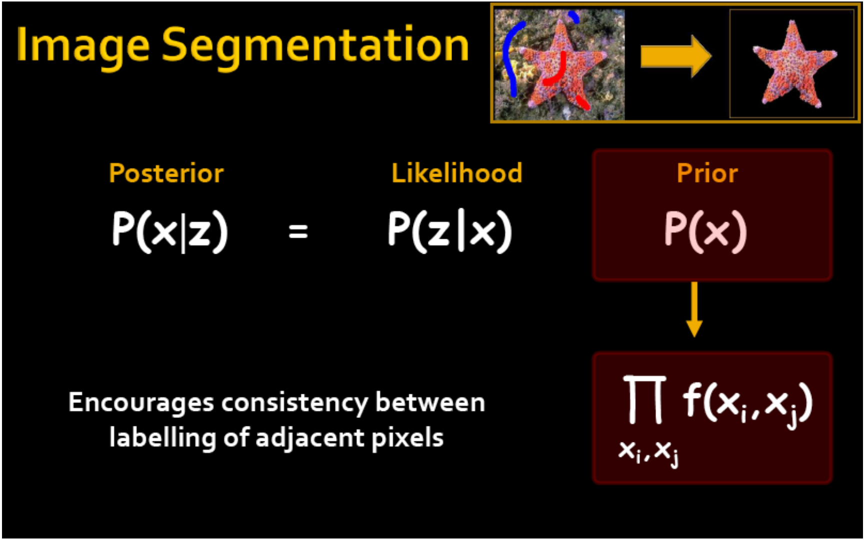
$P(z_i|x_i=1)$

MAP Solution

$$x^* = \underset{x}{\operatorname{argmax}}\, P(z|x)$$

$$= \underset{x}{\operatorname{argmax}} \prod_{x_i} P(z_i|x_i)$$

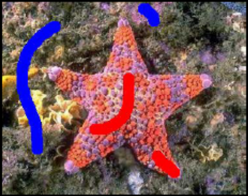
[Source: P. Kohli]

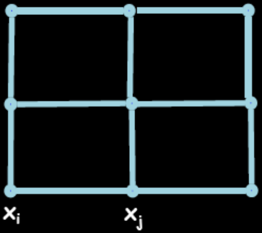


[Source: P. Kohli]

Prior

$$P(x|z) \sim P(z|x)\, P(x)$$





$$P(x) = \prod_{i,j \in N} f_{ij}(x_i, x_j)$$

$$= \prod_{i,j \in N} \exp\{-|x_i - x_j|\}$$

"MRF Ising prior"

[Source: P. Kohli]

Posterior and Energy Functions

$$P(\mathbf{x}|\mathbf{z}) = \boxed{\prod_{x_i} P(z_i|x_i)} \; \boxed{\prod_{x_i,x_j} P(x_i,x_j)}$$

Posterior Probability

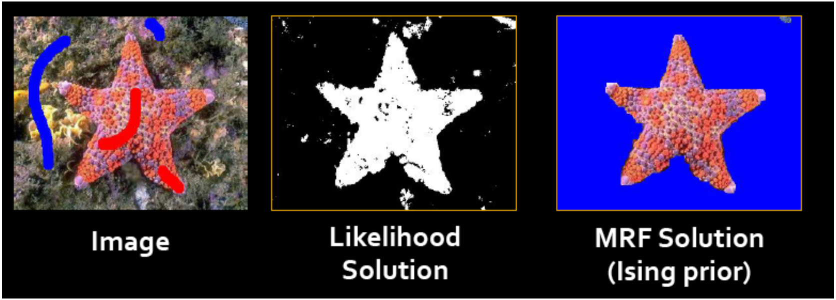
$$\Big\downarrow \text{-ve log}$$

$$\boxed{E(\mathbf{x},\mathbf{z},\mathbf{w}) = \sum_i \theta_i(x_i,z_i) + w \sum_{i,j} \theta_{ij}(x_i,x_j,z_i,z_j)}$$

Energy

[Source: P. Kohli]

Results of the Ising Model



[Source: P. Kohli]

Conditional Random Fields

$$P(x|z) = \prod_{x_i} P(z_i|x_i) \quad \boxed{\prod_{x_i,x_j} P(x_i,x_j,z_i,z_j)}$$

$$\Big\downarrow \text{-ve log}$$

$$\boxed{E(x,z,w) = \sum_i \theta_i (x_i,z_i) + w \sum_{i,j} \theta_{ij} (x_i,x_j,z_i,z_j)}$$

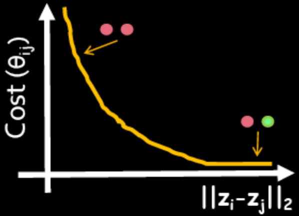
[Boykov and Jolly '01] [Blake et al. '04] [Rother, Kolmogorov and Blake '04]

[Source: P. Kohli]

Conditional Random Fields

$$E(x,z,w) = \sum_i \theta_i(x_i, z_i) + w \sum_{i,j} \theta_{ij}(x_i, x_j, z_i, z_j)$$

Pairwise Cost



[Boykov and Jolly '01] [Blake et al. '04] [Rother, Kolmogorov and Blake '04]

[Source: P. Kohli]

Conditional Random Fields

$$E(x,z,w) = \sum_{i} \theta_i (x_i, z_i) + w \sum_{i,j} \theta_{ij} (x_i, x_j, z_i, z_j)$$

Pairwise Cost

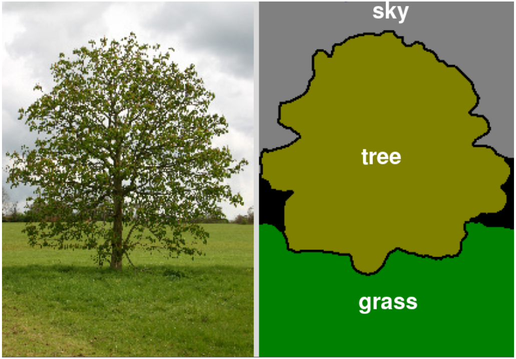
Global Minimum (x*)

[Boykov and Jolly ' 01] [Blake et al. '04] [Rother, Kolmogorov and Blake `04]

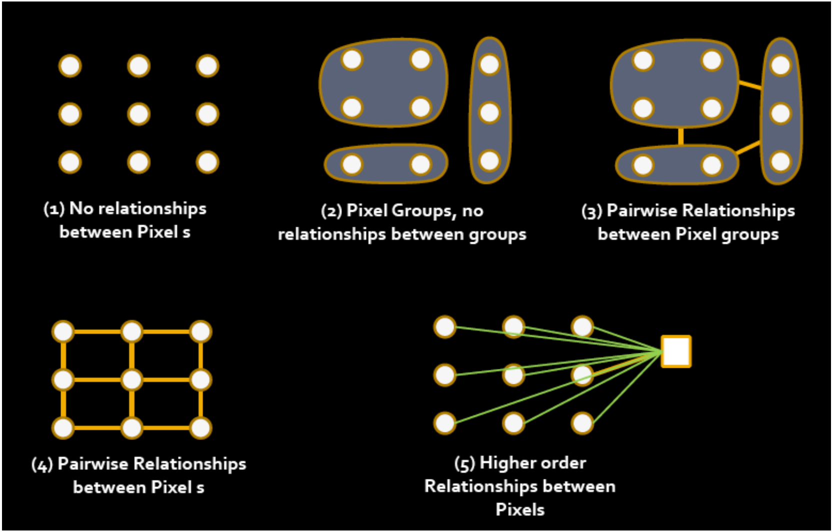
[Source: P. Kohli]

Example: Supervised Semantic Segmentation

- Assign a label to every pixel

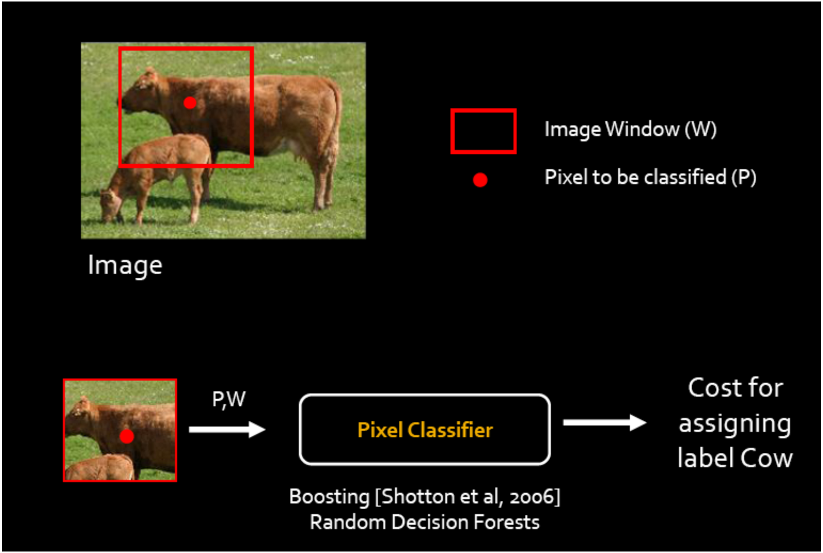


Different Approaches



[Source: P. Kohli]

Building Unitary Potentials



[Source: P. Kohli]

Image Segmentation

n = number of pixels

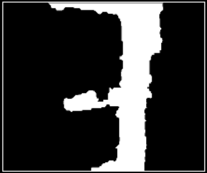
$E: \{0,1\}^n \rightarrow R$

$0 \rightarrow fg, \; 1 \rightarrow bg$

$$E(X) = \sum_i c_i x_i + \sum_{i,j} d_{ij} |x_i - x_j|$$

Image

Unary Cost

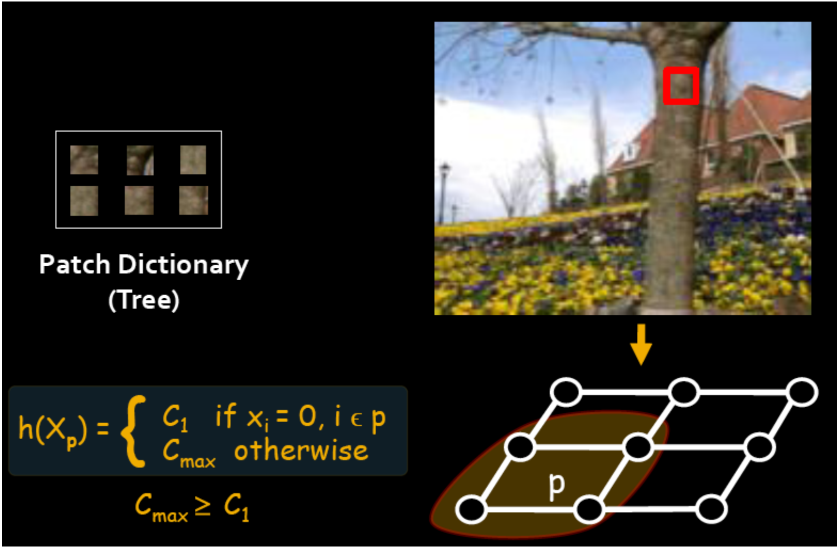


Segmentation

[Boykov and Jolly '01] [Blake et al. '04] [Rother, Kolmogorov and Blake `04]

[Source: P. Kohli]

High order patch potentials



[Source: P. Kohli]

Image Segmentation

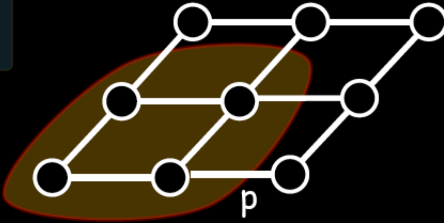
n = number of pixels

$E: \{0,1\}^n \rightarrow R$

$0 \rightarrow fg, \; 1 \rightarrow bg$

$$E(X) = \sum_i c_i x_i + \sum_{i,j} d_{ij} |x_i - x_j| + \sum_p h_p(X_p)$$

$$h(X_p) = \begin{cases} C_1 & \text{if } x_i = 0, i \in p \\ C_{max} & \text{otherwise} \end{cases}$$



[Kohli *et al.* '07]

[Source: P. Kohli]

Image Segmentation

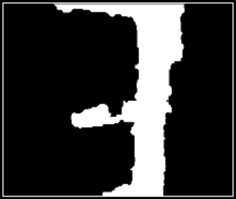
n = number of pixels

$E: \{0,1\}^n \rightarrow R$

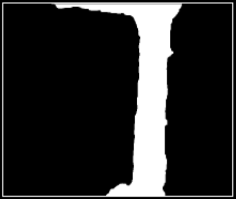
$0 \rightarrow fg, \; 1 \rightarrow bg$

$$E(X) = \sum_i c_i x_i + \sum_{i,j} d_{ij} |x_i - x_j| + \sum_p h_p(X_p)$$

Image



Pairwise Segmentation

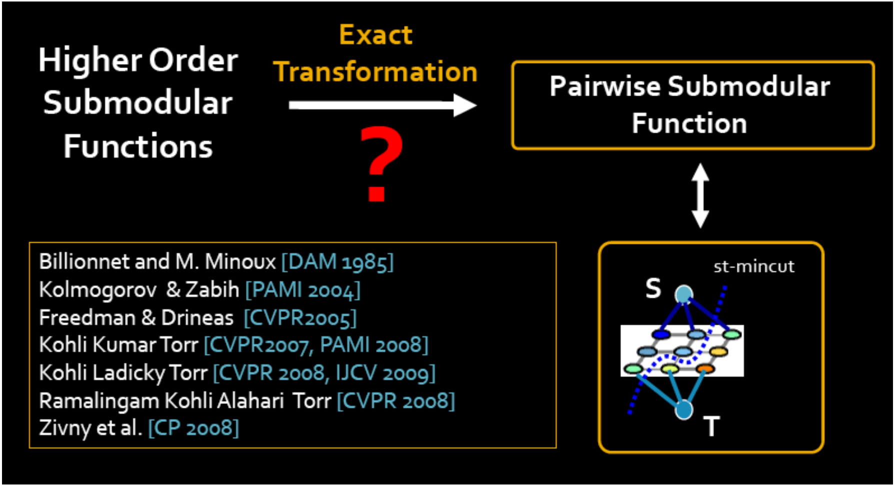


Final Segmentation

[Kohli *et al.* '07]

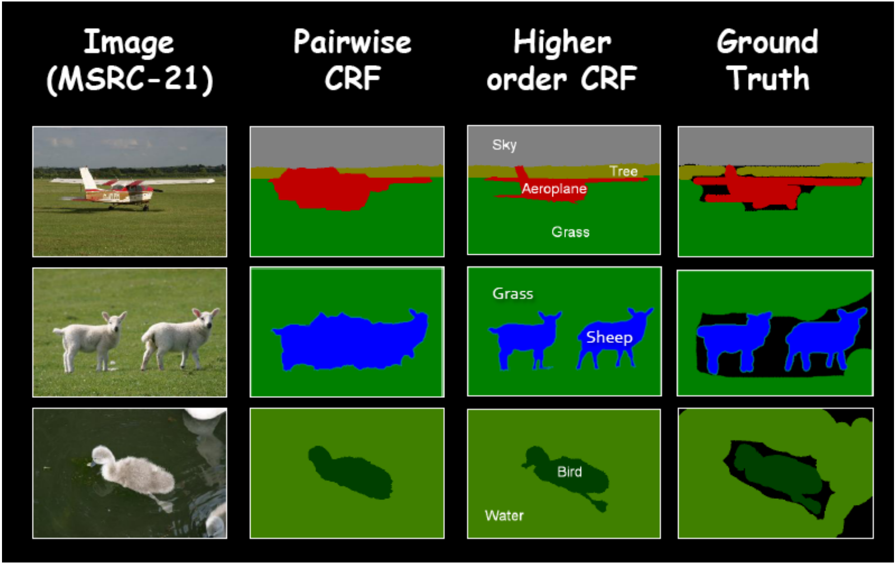
[Source: P. Kohli]

Minimizing higher order terms



[Source: P. Kohli]

Qualitative Results

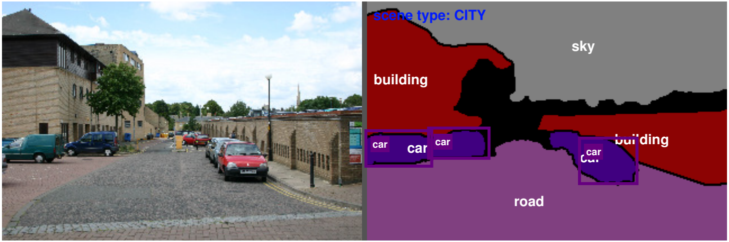


[Source: P. Kohli]

Example: Holistic Scene Understanding

For an image we would like to reason about:

- **Objects**: which class, where, how many?
- **Segmentation**: which semantic label does each pixel take?
- **Scene classification**: which scene am I looking at?



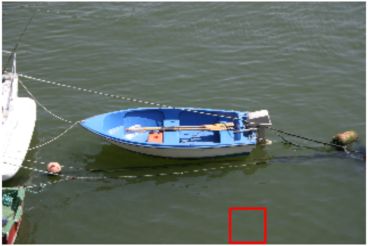
Why Holistic?

Let's use a classifier for each task independently. What's in the patch?

- detector: *bird*
- seg classif.: *water*
- scene: *boat*

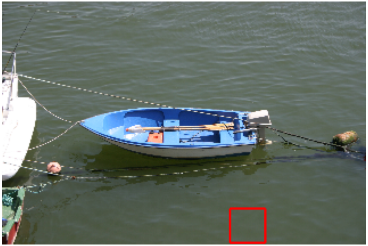
Why Holistic?

Let's use a classifier for each task independently. What's in the patch?

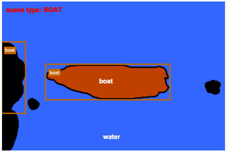


- detector: *bird*
- seg classif.: *water*
- scene: *boat*

Let's use a classifier for each task independently. What's in the patch?

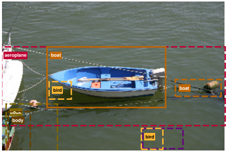


- detector: *bird*
- seg classif.: *water*
- scene: *boat*



groundtruth

segmentation only



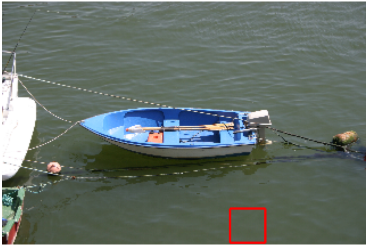
detection only

boat

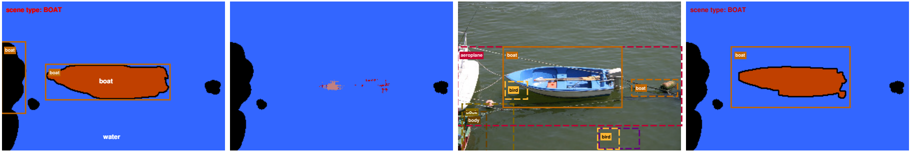
scene only

Why Holistic?

Let's use a classifier for each task independently. What's in the patch?



- detector: *bird*
- seg classif.: *water*
- scene: *boat*

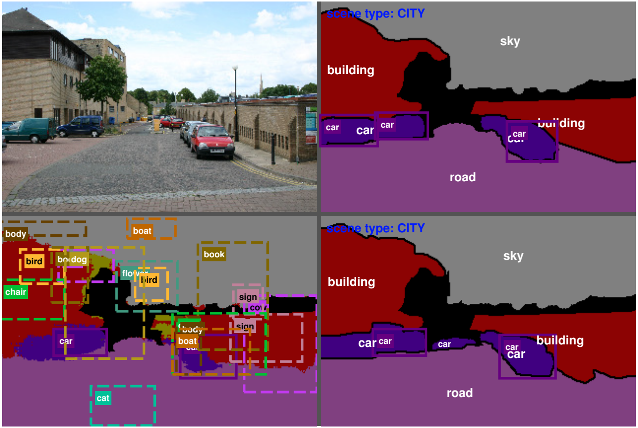


groundtruth | segmentation only | detection only | joint

Holistic Scene Understanding

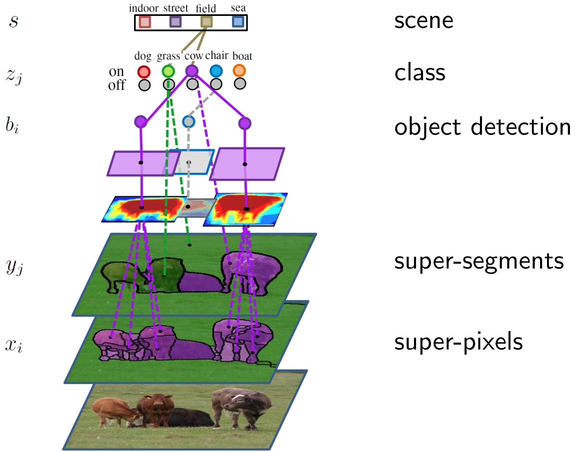
We want to reason about the scene as a **whole**.

- Joint inference of scene type, 2D objects and semantic segmentation
- Efficient learning and inference with structure prediction



Compact Holistic Model

- Define the problem as hierarchical CRF
- Compatibility potentials + evidence + shape prior



Compact Holistic Model

We define the problem as a *holistic conditional random field*

$$p(\mathbf{a}) = p(\mathbf{x}, \mathbf{y}, \mathbf{z}, \mathbf{b}, \mathbf{s}) = \frac{1}{Z} \prod_i \psi_i(\mathbf{a}_i) \prod_\alpha \psi_\alpha(\mathbf{a}_\alpha)$$

where $\mathbf{a} = (\mathbf{x}, \mathbf{y}, \mathbf{z}, \mathbf{b}, \mathbf{s})$ represents the set of all random variables

- $x_i \in \{1, \dots, \mathcal{C}\}$: class label of the i-th super-pixel (first layer of the hierarchy)
- $y_i \in \{1, \dots, \mathcal{C}\}$: class label of the i-th super-segment (second layer)
- $b_i \in \{0, 1\}$: binary variable indicating whether an object detection is *on* or *off*
- $z_i \in \{0, 1\}$: binary variable indicating the presence of class i in the image
- $s \in \{1, \dots, \mathcal{S}\}$: scene type label

Compact Holistic Model

- **Learning** the weights w_i, where $w_i \phi_i = \log(\psi_i)$, is done with primal-dual approximated learning algorithm

- Joint **inference** is performed by computing the MAP estimate:

$$\max_{\mathbf{x},\mathbf{y},\mathbf{z},\mathbf{b},\mathbf{s}} \frac{1}{Z} \prod_i \psi_i(\mathbf{a}_i) \prod_\alpha \psi_\alpha(\mathbf{a}_\alpha)$$

  We use a convergent message-passing algorithm without restriction to submodularity and potential specific moves

Unitary Potentials

- Super-pixel and super-segment:

  $\phi_i(x_i)$ and $\phi_j(y_j)$: average of TextonBoost pixel potentials inside each region

- Object detection:

$$\phi_l^{BBox}(b_i) = \begin{cases} \sigma(r_i - \lambda_l) & \text{if } b_i = 1 \wedge c_i = l \\ 0 & \text{otherwise.} \end{cases}$$

  Here r_i is the score from Felzenswalb et al. detector, λ_l is the threshold of the detector for that class, c_i is the detector class, and $\sigma(x) = 1/(1 + \exp(-1.5\,x))$ is a logistic function that converts the classifier score into probability.

- Scene:

$$\phi^{Scene}(s = k) = \sigma(t_k)$$

  where t_k denotes the classifier score for scene class k

Pairwise potentials

- Super-pixel – Super-segment: we use the P^n potentials by Kohli et al.,CVPR'07:

$$\phi_{i,j}(x_i, y_j) = \begin{cases} -\infty & \text{if } x_i \neq y_j \\ 0 & \text{otherwise.} \end{cases}$$

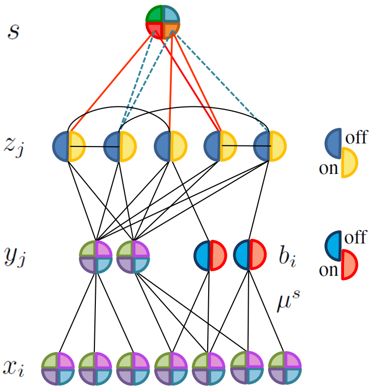
- Super-segment – Class:

$$\phi_{i,j}(y_i, z_j) = \begin{cases} -\infty & \text{if } y_i = j \wedge z_j = 0 \\ 0 & \text{otherwise.} \end{cases}$$

- Class – Scene:

$$\phi^{SC}(s, z_j) = \begin{cases} f_{s,z_j} & \text{if } z_j = 1 \wedge f_{s,z_j} > 0 \\ -\tau & \text{if } z_j = 1 \wedge f_{s,z_j} = 0 \\ 0 & \text{otherwise.} \end{cases}$$

where f_{s,z_j} represents the probability of occurrence of class z_j for scene type s



Pairwise potentials

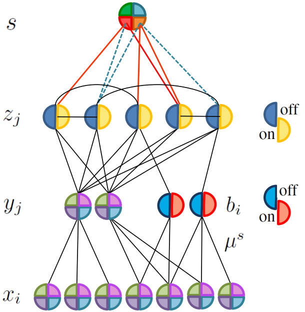
- Detection – Class:

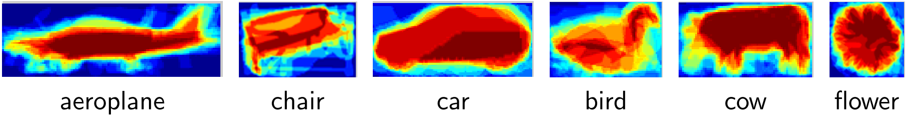
$$\phi^{BClass}_{i,j}(\beta_i, b_i, z_j) = \begin{cases} -\infty & \text{if } z_j = 0 \wedge c_i = j \wedge b_i = 1 \\ 0 & \text{otherwise.} \end{cases}$$

- Detection – Super-pixel (shape prior):

$$\phi^{sh}_l(x_j, b_i, \beta_i) = \begin{cases} \mu(x_j, \beta_i) & \text{if } x_j = c_i \wedge b_i = 1 \\ 0 & \text{otherwise.} \end{cases}$$

where $\mu(x_j, \beta_i) = \frac{1}{|A_j|} \sum_{p \in A_j} \mu(p, m_i)$, A_j is the set of pixels in the j-th segment, $|A_j|$ is the cardinality of this set, and $\mu(p, m_i)$ is the value of the mean mask for component m_i





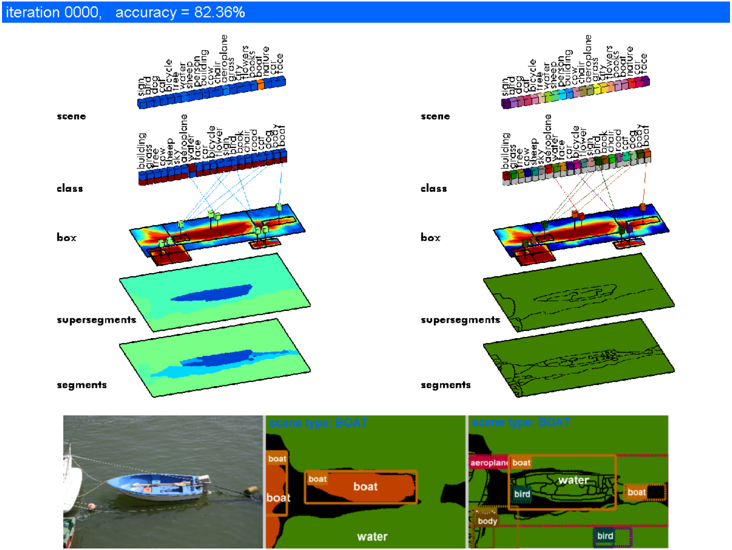
Loss function

Structure prediction problems require a specification for the loss. We define it as a weighted sum of task-specific losses, each of order at most 2.

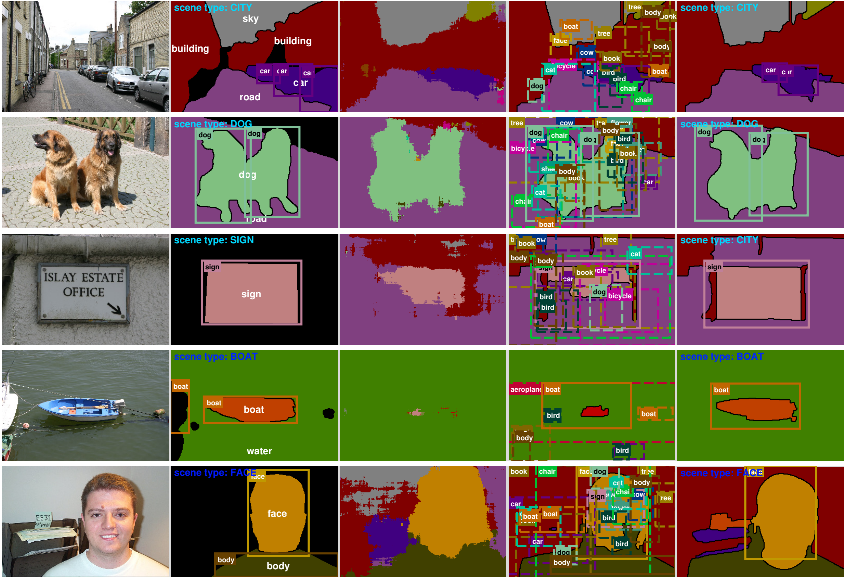
- Super-pixel and super-segment layers: loss is the total number of pixels that were wrongly predicted.
- Class: $0 - 1$ loss
- Scene: $0 - 1$ loss
- Detection:

$$\Delta_B(b_i, \hat{b}_i) = \begin{cases} 1 - \frac{intersection}{union} & \text{if } b_i = 1 \\ \frac{intersection}{union} & \text{otherwise} \end{cases}$$

Inference example



Joint Inference Results



Segmentation Results MSRC-21

[J. Yao, S. Fidler and R. Urtasun, CVPR12]

Table: MSRC-21 segmentation results

	building	grass	tree	cow	sheep	sky	aeroplane	water	face	car	bicycle	flower	sign	bird	book	chair	road	cat	dog	body	boat	**average**	**global**
	origMSRC dataset																						
Shotton et al	49	88	79	**97**	**97**	78	82	54	87	74	72	74	36	24	93	51	78	75	35	66	18	67	72
Jiang and Tu	53	97	83	70	71	98	75	64	74	64	88	67	46	32	92	61	89	59	66	64	13	68	78
Pixel-CRF	73	92	85	75	78	92	75	76	86	79	87	96	**95**	31	81	34	84	53	61	60	15	72	81
Hierarch. CRF	**80**	96	86	74	87	**99**	74	**87**	86	**87**	82	97	**95**	30	86	31	**95**	51	**69**	66	9	75	86
HCRF+Coocc.	74	**98**	90	75	86	**99**	81	84	**90**	83	91	**98**	75	49	95	63	91	71	49	**72**	18	77.8	**86.5**
Harmony pot.	60	78	77	91	68	88	87	76	73	77	93	97	73	**57**	95	**81**	76	81	46	56	**46**	75	77
Segm.+Class	72	**98**	**91**	77	82	93	86	86	82	82	93	97	71	50	96	59	88	78	51	67	0	76.2	85.1
Det 15 class	69	**98**	90	78	86	93	88	83	**90**	83	94	97	73	50	96	71	89	79	54	64	8	77.8	85.3
full model	71	**98**	90	79	86	93	**88**	86	90	84	**94**	**98**	76	53	**97**	71	89	**83**	55	68	17	**79.3**	86.2

Detection and Scene Classification Results

[J. Yao, S. Fidler and R. Urtasun, CVPR12]

Table: MSRC-21 object detection results

	cow	sheep	aeroplane	face	car	bicycle	flower	sign	bird	book	chair	cat	dog	body	boat	**average**
	Recall at equal FPPI															
FPPI rate	0.03	0.02	0.00	0.01	0.05	0.03	0.04	0.02	0.02	0.01	0.00	0.02	0.04	0.04	0.02	0.02
LSVM	84.6	73.9	84.6	**59.4**	50.0	63.6	16.9	40.0	**16.2**	23.7	50.0	20.0	20.0	**43.2**	18.8	44.3
cont. LSVM	76.9	17.4	23.1	50.0	50.0	**68.2**	15.3	40.0	8.1	18.4	50.0	30.0	**33.3**	38.6	**21.9**	36.1
Detection	**88.5**	78.3	**100.0**	43.8	**52.4**	63.6	**20.3**	**53.3**	**16.2**	42.1	**62.5**	**50.0**	26.7	38.6	6.3	49.5
full model	**88.5**	**82.6**	**100.0**	46.9	**52.4**	63.6	**20.3**	**53.3**	**16.2**	**44.7**	**62.5**	40.0	26.7	38.6	12.5	**49.9**
	Average Precision															
LSVM	**78.6**	76.5	96.2	56.4	**54.1**	**61.7**	19.9	45.0	**18.5**	30.0	59.2	31.4	28.0	**45.5**	22.1	48.2
cont.LSVM	75.8	37.0	85.1	**58.2**	52.1	60.8	19.1	38.5	12.3	28.6	60.5	32.1	**32.1**	41.7	**26.2**	44.0
Detection	78.1	72.7	**100.0**	45.5	53.1	60.9	**22.9**	**48.9**	18.2	42.9	**63.6**	**46.0**	27.3	34.3	9.1	48.2
full model	78.1	**81.8**	**100.0**	45.5	53.1	60.9	**22.9**	**48.9**	18.2	**44.4**	**63.6**	45.6	27.3	34.3	16.4	**49.4**

Table: MSRC-21 scene classification

	classifier	full m.
accuracy	79.5	**80.6**

More Results ...

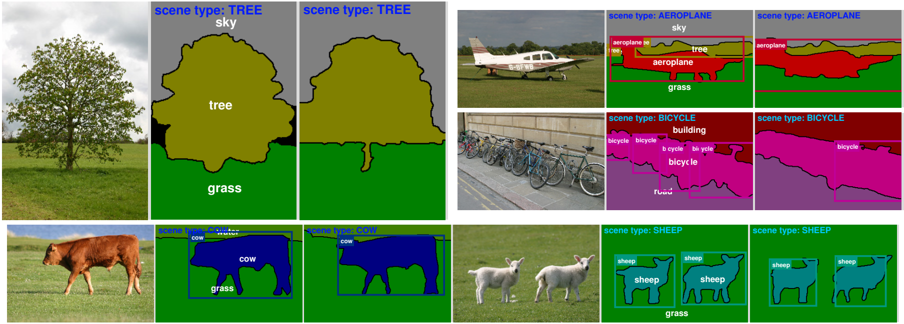


Figure: Segmentation examples: (image, groundtruth, our holistic scene model)

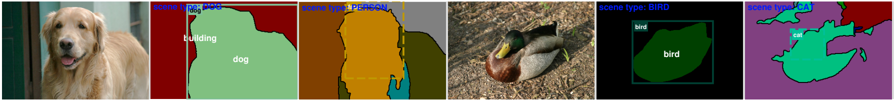


Figure: Examples of failure modes.

Let's talk about attributes

Zero-shot learning

- Can I leaned what a mule is without seen a single instance if I know what horses and donkeys are?
- Traditional paradigm is not very appropiate

[Source: D. Parikh]

Zero-shot learning

- Can I leaned what a mule is without seen a single instance if I know what horses and donkeys are?
- Traditional paradigm is not very appropiate

[Source: D. Parikh]

Zero-shot learning

- Can I leaned what a mule is without seen a single instance if I know what horses and donkeys are?
- Traditional paradigm is not very appropiate

[Source: D. Parikh]

Zero-shot learning

- Can I leaned what a mule is without seen a single instance if I know what horses and donkeys are?
- Traditional paradigm is not very appropiate

[Source: D. Parikh]

Zero-shot learning

- Can I leaned what a mule is without seen a single instance if I know what horses and donkeys are?
- Traditional paradigm is not very appropiate

[Source: D. Parikh]

Zero-shot learning

- Can I leaned what a mule is without seen a single instance if I know what horses and donkeys are?
- Traditional paradigm is not very appropiate

[Source: D. Parikh]

Zero-shot learning

- Can I leaned what a mule is without seen a single instance if I know what horses and donkeys are?
- Traditional paradigm is not very appropiate

[Source: D. Parikh]

Attributes

- Long history of attributes in vision, starting in 2007.
- They are typically simple classifiers
- The score of those classifiers is an alternative representation
- They are binary

Is furry

Has four-legs

Legs shorter than horses'

Tail longer than donkeys'

Has tail

[Oliva 2001] [Ferrari 2007] [Lampert 2009] [Farhadi 2009] [Kumar 2009] [Wang 2009] [Wang 2010] [Berg 2010] [Branson 2010] [Parikh 2010] [ICCV 2011] … 9

[Source: D. Parikh]

Attributes

- Long history of attributes in vision, starting in 2007.
- They are typically simple classifiers
- The score of those classifiers is an alternative representation
- They are binary

Is furry

Has four-legs

Legs shorter
than horses'

Tail longer
than donkeys'

Has tail

[Source: D. Parikh]

Attributes

- Some of them are relative

Is furry

Has four-legs

Legs shorter than horses'

Tail longer than donkeys'

Has tail

Image Search

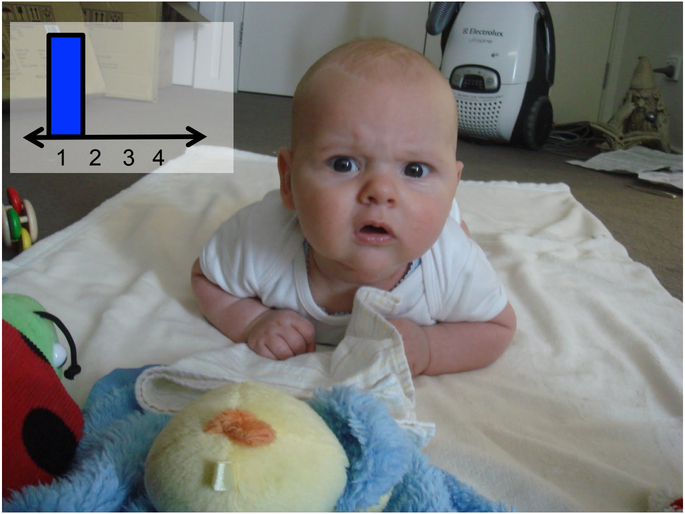
- I want to ask about an image of Chicago
- This might bee too crowded for my taste

Image Search

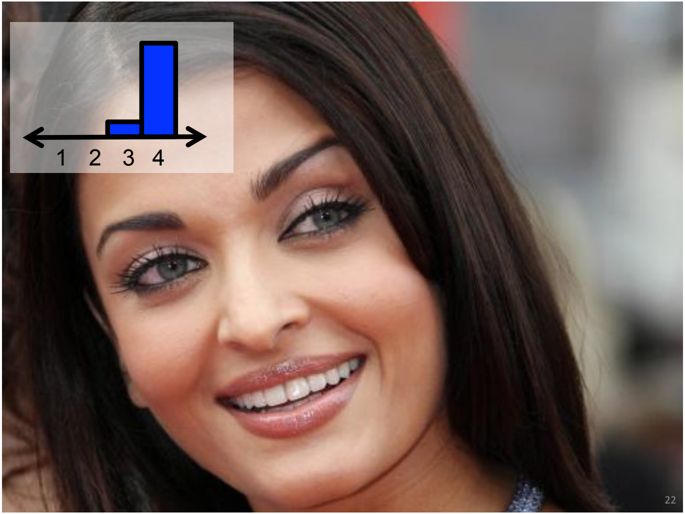
- I want to ask about an image of Chicago
- This might bee too crowded for my taste

How do we think about attributes?

How do we think about attributes?

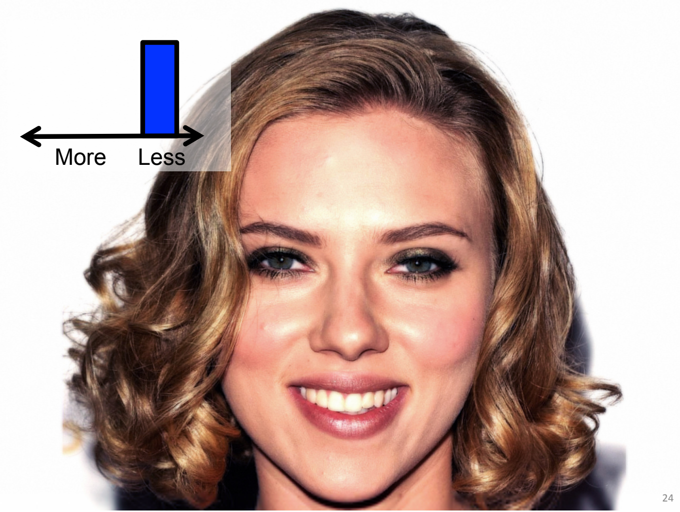


How do we think about attributes?



How do we think about attributes?

But it's easy to say...



Relative Attributes [Parikh et al. 11]

Relative attributes

- Allow relating images and categories to each other
- Learn ranking function for each attribute

Novel applications

- Zero-shot learning from attribute comparisons
- Automatically generating relative image descriptions

For each attribute a_m, open

Supervision is

$O_m : \{ ($  $\succ$  $), \ldots \}$,

$S_m : \{ \{$  $\sim$  $\}, \ldots \}$

Learning Relative Attributes

Learn a scoring function $r_m(\boldsymbol{x_i}) = \boldsymbol{w_m^T x_i}$

Learned parameters

that best satisfies constraints:

$$\forall (i,j) \in O_m : \boldsymbol{w_m^T x_i} > \boldsymbol{w_m^T x_j}$$

$$\forall (i,j) \in S_m : \boldsymbol{w_m^T x_i} = \boldsymbol{w_m^T x_j}$$

Max-margin learning to rank formulation

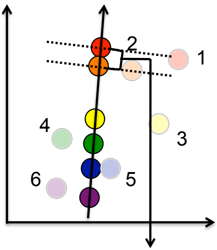
$$\min \quad \left(\frac{1}{2}||\boldsymbol{w_m^T}||_2^2 + C\left(\sum \xi_{ij}^2 + \sum \gamma_{ij}^2\right)\right)$$

$$\text{s.t} \quad \boldsymbol{w_m^T}(\boldsymbol{x_i} - \boldsymbol{x_j}) \geq 1 - \xi_{ij}, \forall (i,j) \in O_m$$

$$|\boldsymbol{w_m^T}(\boldsymbol{x_i} - \boldsymbol{x_j})| \leq \gamma_{ij}, \forall (i,j) \in S_m$$

$$\xi_{ij} \geq 0; \gamma_{ij} \geq 0$$

Based on [Joachims 2002]



Rank Margin

Image → Relative Attribute Score

Zero Shot Learning

Training: Images from **S seen** categories and

Descriptions of **U unseen** categories

Age: **Hugh** $\succ$ **Clive** $\succ$ **Scarlett** **Jared** $\succ$ **Miley**

Smiling: **Miley** $\succ$ **Jared**

Need not use all attributes, or all seen categories

Testing: Categorize image into one of **S**+**U** categories

30

Automatic Relative Description

Density 

Novel image 

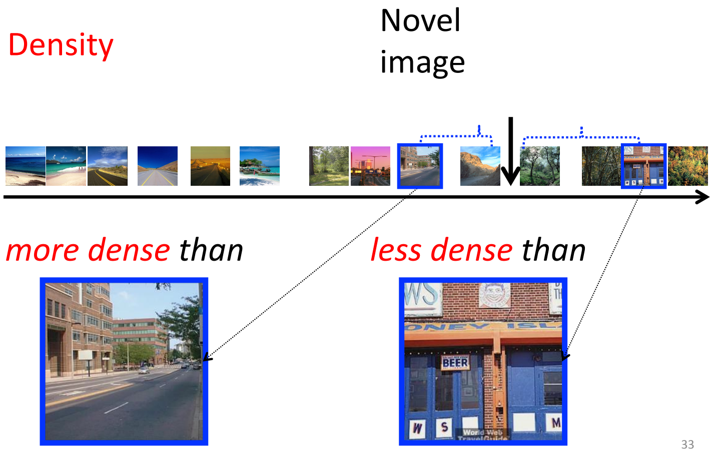
Conventional binary description: *not dense*

Dense: 

Not dense: 

32

Automatic Relative Description



Automatic Relative Description

more dense than ***Highways***, *less dense* than ***Forests***

Results

Binary (existing):

Not natural

Not open

Has perspective

Relative (ours):

More natural than insidecity
Less natural than highway

More open than street
Less open than coast

Has more perspective than highway
Has less perspective than insidecity

Results

Binary (existing):

Not natural

Not open

Has perspective

Relative (ours):

More natural than tallbuilding
Less natural than forest

More open than tallbuilding
Less open than coast

Has more perspective than tallbuilding

Results

Binary (existing):

Not Young

BushyEyebrows

RoundFace

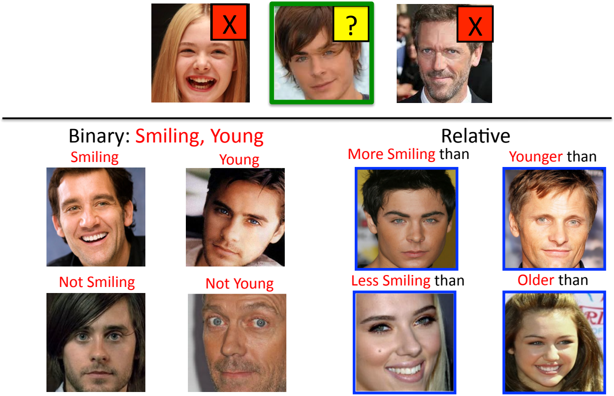
Relative (ours):

More Young than CliveOwen
Less Young than ScarlettJohansson

More BushyEyebrows than ZacEfron
Less BushyEyebrows than AlexRodriguez

More RoundFace than CliveOwen
Less RoundFace than ZacEfron

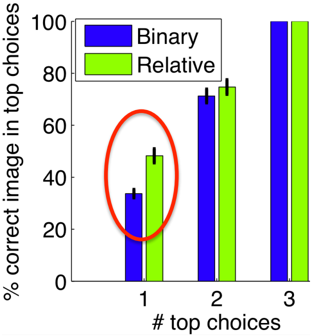
Human Studies: Which Image is described?



Automatic Relative Image Description

18 subjects

Test cases:
10 OSR, 20 PubFig



There is much more... for that you need to do a PhD on vision ;)