Max-margin Learning for Lower Linear Envelope Potentials in Binary Markov Random Fields

Stephen Gould
stephen.gould@anu.edu.au

Australian National University

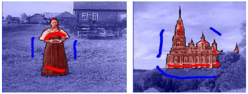
ICML, 29 June 2011

Motivation: Image Labeling

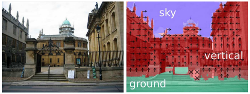
Image labeling: Label every pixel in an image with a class label from some pre-defined set, i.e., $y_p \in \mathcal{L}$.

Motivation: Image Labeling

Image labeling: Label every pixel in an image with a class label from some pre-defined set, i.e., $y_p \in \mathcal{L}$.

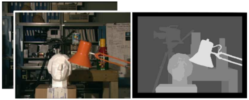


Interactive segmentation (Boykov and Jolly, 2001; Boykov and Funka-Lea, 2006)



Surface context (Hoiem et al., 2005)

Semantic labeling (He et al., 2004; Shotton et al., 2006; Gould et al., 2009)

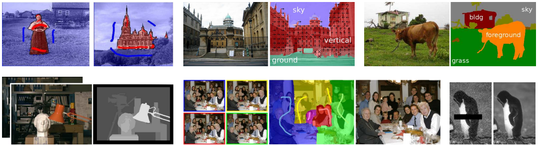


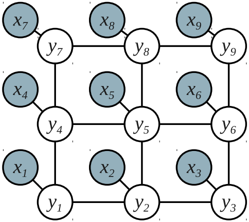
Stereo matching (Scharstein and Szeliski, 2002)

Photo montage (Agarwala et al., 2004)

Denoising

Motivation: Image Labeling



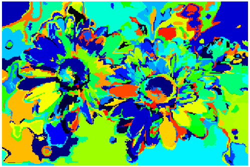


These problems are typically solved using a pairwise conditional Markov random field.

However, pairwise terms are often not expressive enough.

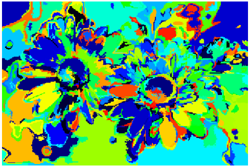
Consistency Potentials

Suppose an oracle told us which pixels belong together, e.g., for the figure-ground segmentation problem we might have


Consistency Potentials

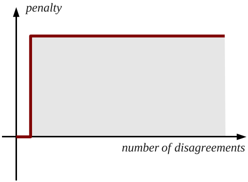
Suppose an oracle told us which pixels belong together, e.g., for the figure-ground segmentation problem we might have


Then we would only need to label the so-called *superpixels* rather than individual pixels.

Consistency Potentials

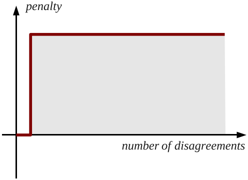
Unfortunately we don't have a perfect oracle. So what can we do?



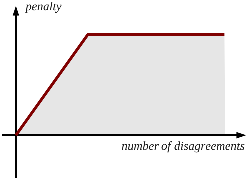
[Kohli et al., 2007]

Consistency Potentials

Unfortunately we don't have a perfect oracle. So what can we do?



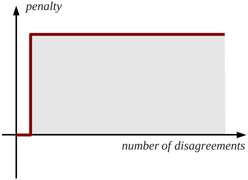
[Kohli et al., 2007]



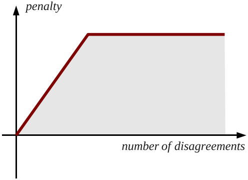
[Kohli et al., 2008]

Consistency Potentials

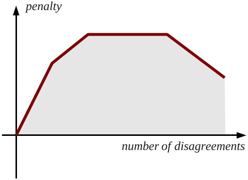
Unfortunately we don't have a perfect oracle. So what can we do?



[Kohli et al., 2007]



[Kohli et al., 2008]



[Kohli and Kumar, 2010]

Higher-order Markov Random Fields

The *energy function* for a higher-order MRF over discrete random variables $\mathbf{y} = \{y_1, \ldots, y_n\}$ can be written as:

$$E(\mathbf{y}; \mathbf{x}, \boldsymbol{\theta}) = \overbrace{\sum_c \psi_c(\mathbf{y}_c)}^{\text{clique potentials}}$$

Higher-order Markov Random Fields

The *energy function* for a higher-order MRF over discrete random variables $\mathbf{y} = \{y_1, \ldots, y_n\}$ can be written as:

$$
\begin{aligned}
E(\mathbf{y}; \mathbf{x}, \boldsymbol{\theta}) &= \overbrace{\sum_c \psi_c(\mathbf{y}_c)}^{\text{clique potentials}} \\
&= \underbrace{\sum_i \psi_i^U(y_i)}_{\text{unary}} + \underbrace{\sum_{ij} \psi_{ij}^P(y_i, y_j)}_{\text{pairwise}} + \underbrace{\sum_c \psi_c^H(\mathbf{y}_c)}_{\text{higher-order}}
\end{aligned}
$$

where the *potential* functions ψ_i^U, ψ_{ij}^P and ψ_c^H encode preferences for unary, pairwise and k-ary variable assignments, respectively.

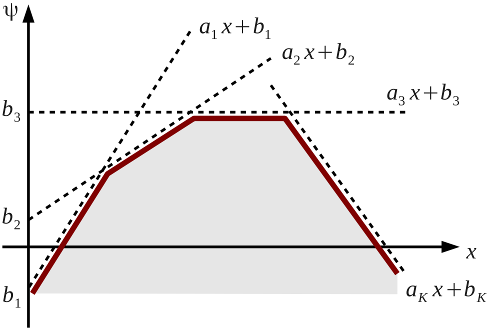
The goal of inference is to find $\mathbf{y}^\star = \operatorname{argmin}_{\mathbf{y}} E(\mathbf{y}; \mathbf{x}, \boldsymbol{\theta})$.

Binary Lower Linear Envelope MRFs

$$\psi_c^H(\mathbf{y}_c) \triangleq \min_k \left\{ a_k \sum_{i \in \mathcal{C}} y_i + b_k \right\}$$

Binary Lower Linear Envelope MRFs

$$\psi_c^H(\mathbf{y}_c) \triangleq \min_k \left\{ a_k \sum_{i \in \mathcal{C}} y_i + b_k \right\}$$



Energy Minimization ([Kohli and Kumar, CVPR 2010])

$$\psi_c^H(\mathbf{y}_c) \triangleq \min_k \left\{ a_k \sum_{i \in \mathcal{C}} y_i + b_k \right\} = \min_k \{ f_k(\mathbf{y}_c) \}$$

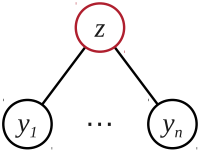
Energy Minimization ([Kohli and Kumar, CVPR 2010])

$$\psi_c^H(\mathbf{y}_c) \triangleq \min_k \left\{ a_k \sum_{i \in \mathcal{C}} y_i + b_k \right\} = \min_k \left\{ f_k(\mathbf{y}_c) \right\}$$

Introduce multi-valued auxiliary random variable $z \in \{1, \ldots, K\}$ and write

$$\widetilde{\psi}_c^H(\mathbf{y}_c, z) = \overbrace{\sum_k [\![z = k]\!] f_k(\mathbf{y}_c)}^{\text{unary and pairwise}}.$$

Now minimize jointly over $\mathbf{y}$ and z.



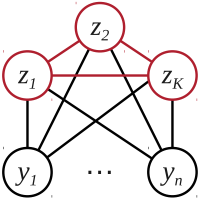
Energy Minimization (Attempt 2)

$$\psi_c^H(\mathbf{y}_c) \triangleq \min_k \left\{ a_k \sum_{i \in \mathcal{C}} y_i + b_k \right\} = \min_k \{f_k(\mathbf{y}_c)\}$$

Introduce auxiliary binary random variables $\mathbf{z} = (z_1, \ldots, z_K)$ with mutual exclusion constraint and write

$$\tilde{\psi}_c^H(\mathbf{y}_c, \mathbf{z}) = \overbrace{\sum_k z_k f_k(\mathbf{y}_c)}^{\text{unary and pairwise}} \quad \text{s.t.} \quad \overbrace{\sum_k z_k}^{\text{global}} = 1.$$

Now minimize jointly over $\mathbf{y}$ and $\mathbf{z}$.



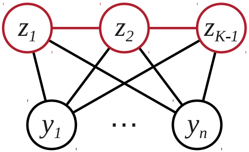
Energy Minimization (Attempt 3)

$$\psi_c^H(\mathbf{y}_c) \triangleq \min_k \left\{ a_k \sum_{i \in \mathcal{C}} y_i + b_k \right\} = \min_k \{ f_k(\mathbf{y}_c) \}$$

Assume sorted on a_k. Introduce auxiliary binary random variables $\mathbf{z} = (z_1, \ldots, z_{K-1})$ with *inclusion* constraints and write

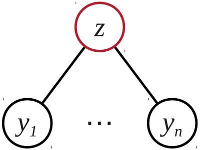
$$\widetilde{\psi}_c^H(\mathbf{y}_c, \mathbf{z}) = \overbrace{f_1(\mathbf{y}_c)}^{\text{unary}} + \overbrace{\sum_k z_k \left(f_{k+1}(\mathbf{y}_c) - f_k(\mathbf{y}_c) \right)}^{\text{unary and pairwise}} \text{ s.t. } \overbrace{z_k \geq z_{k+1}}^{\text{pairwise}}.$$

Now minimize jointly over $\mathbf{y}$ and $\mathbf{z}$.

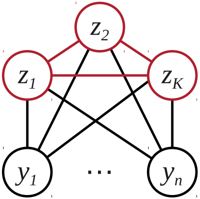


Relationship to Binary Pairwise MRFs

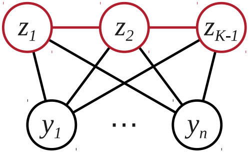
Each transformation results in a different latent variable Markov random field:



Attempt 1 (multi-valued; pairwise)



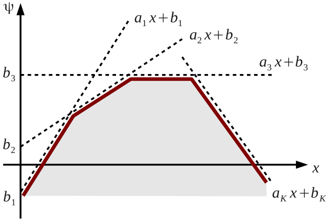
Attempt 2 (binary; non-pairwise)



Attempt 3 (binary; pairwise)

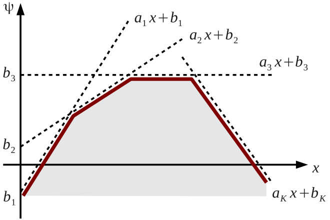
Exact Inference

- **Claim 1:** The binary pairwise MRF induced by "Energy Minimization Attempt 3" is submodular (see paper for proof)



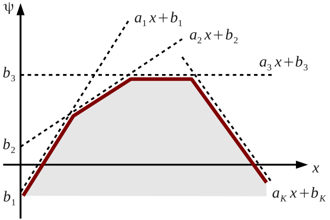
Exact Inference

- **Claim 1:** The binary pairwise MRF induced by "Energy Minimization Attempt 3" is submodular (see paper for proof)
- **Claim 2:** Submodular binary MRFs can be minimized in time polynomial in the number of variables ([Hammer, 1965])



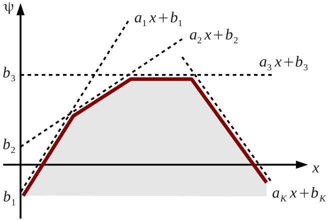
Exact Inference

- **Claim 1:** The binary pairwise MRF induced by "Energy Minimization Attempt 3" is submodular (see paper for proof)
- **Claim 2:** Submodular binary MRFs can be minimized in time polynomial in the number of variables ([Hammer, 1965])
  - Empirically, very fast algorithms exist for quadratic submodular problems ([Boykov and Kolmogorov, 2004])



Exact Inference

- **Claim 1:** The binary pairwise MRF induced by "Energy Minimization Attempt 3" is submodular (see paper for proof)
- **Claim 2:** Submodular binary MRFs can be minimized in time polynomial in the number of variables ([Hammer, 1965])
  - Empirically, very fast algorithms exist for quadratic submodular problems ([Boykov and Kolmogorov, 2004])
- We can perform *exact* inference in lower linear envelope binary Markov random fields



Max-margin Learning for Structured Prediction

Max-margin Learning for Structured Prediction

- **Energy function.** Parameterized by $\boldsymbol{\theta} \in \mathbb{R}^d$,

$$E(\mathbf{y}; \mathbf{x}, \boldsymbol{\theta}) = \underbrace{\sum_c \psi_c(\mathbf{y}_c; \mathbf{x}, \boldsymbol{\theta}_c)}_{\text{easy inference}} = \overbrace{\boldsymbol{\theta}^T \phi(\mathbf{y}, \mathbf{x})}^{\text{easy learning}}$$

Max-margin Learning for Structured Prediction

- **Energy function.** Parameterized by $\boldsymbol{\theta} \in \mathbb{R}^d$,

$$E(\mathbf{y}; \mathbf{x}, \boldsymbol{\theta}) = \underbrace{\sum_c \psi_c(\mathbf{y}_c; \mathbf{x}, \boldsymbol{\theta}_c)}_{\text{easy inference}} = \overbrace{\boldsymbol{\theta}^T \phi(\mathbf{y}, \mathbf{x})}^{\text{easy learning}}$$

- **Structured loss function.** e.g., $\Delta(\hat{\mathbf{y}}, \mathbf{y}) = \frac{1}{n} \sum_{i=1}^{n} [\![\hat{y}_i \neq y_i]\!]$

Max-margin Learning for Structured Prediction

- **Energy function.** Parameterized by $\boldsymbol{\theta} \in \mathbb{R}^d$,

$$E(\mathbf{y}; \mathbf{x}, \boldsymbol{\theta}) = \underbrace{\sum_c \psi_c(\mathbf{y}_c; \mathbf{x}, \boldsymbol{\theta}_c)}_{\text{easy inference}} = \overbrace{\boldsymbol{\theta}^T \phi(\mathbf{y}, \mathbf{x})}^{\text{easy learning}}$$

- **Structured loss function.** e.g., $\Delta(\hat{\mathbf{y}}, \mathbf{y}) = \frac{1}{n} \sum_{i=1}^n [\![\hat{y}_i \neq y_i]\!]$
- **Learning algorithm.** Given a training set $\{(\mathbf{x}_t, \mathbf{y}_t)\}_{t=1}^T$, solve the *margin-rescaling* optimization problem ([Taskar et al., 2005; Tsochantaridis et al., 2004]).

Max-margin Learning for Structured Prediction

QP for max-margin learning

$$\begin{array}{lll} \text{minimize} & \frac{1}{2}\|\boldsymbol{\theta}\|_2^2 + \frac{C}{T}\sum_t \xi_t & \\ \text{subject to} & \underbrace{\boldsymbol{\theta}^T \phi_t(\mathbf{y}) - \boldsymbol{\theta}^T \phi_t(\mathbf{y}_t)}_{\text{energy difference}} \geq \underbrace{\Delta(\mathbf{y}, \mathbf{y}_t) - \xi_t}_{\text{rescaled margin}}, & \forall t,\ \overbrace{\mathbf{y} \in \mathcal{Y}_t}^{\text{very large}} \\ & \xi_t \geq 0, & \forall t \end{array}$$

Max-margin Learning for Structured Prediction

QP for max-margin learning

$$\begin{array}{lll} \text{minimize} & \frac{1}{2}\|\boldsymbol{\theta}\|_2^2 + \frac{C}{T}\sum_t \xi_t & \\ \text{subject to} & \underbrace{\boldsymbol{\theta}^T\phi_t(\mathbf{y}) - \boldsymbol{\theta}^T\phi_t(\mathbf{y}_t)}_{\text{energy difference}} \geq \underbrace{\Delta(\mathbf{y}, \mathbf{y}_t) - \xi_t}_{\text{rescaled margin}}, & \forall t,\ \overbrace{\mathbf{y} \in \mathcal{Y}_t}^{\text{very large}} \\ & \xi_t \geq 0, & \forall t \end{array}$$

Re-writing constraints

$$\begin{array}{lll} \text{minimize} & \frac{1}{2}\|\boldsymbol{\theta}\|_2^2 + \frac{C}{T}\sum_t \xi_t & \\ \text{subject to} & \xi_t \geq \underbrace{\max_{\mathbf{y}\in\mathcal{Y}_t}\left\{\Delta(\mathbf{y}, \mathbf{y}_t) - \boldsymbol{\theta}^T\phi_t(\mathbf{y})\right\}}_{\text{loss-augmented inference (for given } \boldsymbol{\theta})} + \boldsymbol{\theta}^T\phi_t(\mathbf{y}_t), & \forall t \\ & \xi_t \geq 0, & \forall t \end{array}$$

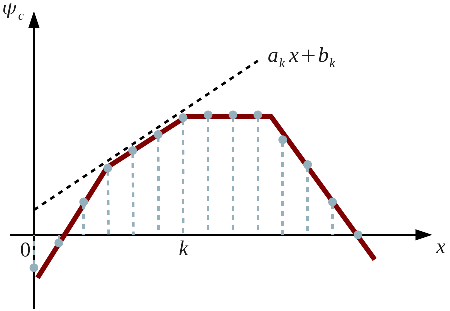
Lower Linear Envelope Representation

It remains to represent the lower linear envelope in a form that is amenable to learning.

$$\begin{aligned}\psi_c^H(\mathbf{y}_c) &\triangleq \min_k \left\{ a_k \sum_{i \in \mathcal{C}} y_i + b_k \right\} \\ &= \boldsymbol{\theta}^T \phi(\mathbf{y}_c)\end{aligned}$$

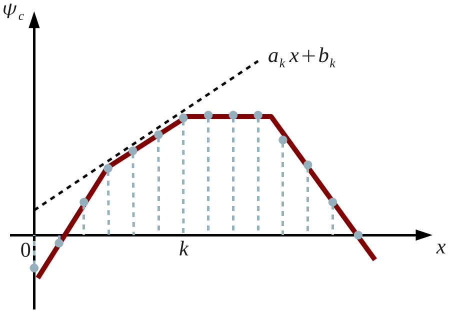
Lower Linear Envelope Representation for Learning

- Sample-based representation with concavity constraints: $2\theta_k - \theta_{k-1} - \theta_{k+1} \geq 0$



Lower Linear Envelope Representation for Learning

- Sample-based representation with concavity constraints: $2\theta_k - \theta_{k-1} - \theta_{k+1} \geq 0$



- Features $\phi(\mathbf{y}_c)$ are 1-of-n indicator vectors

0	0	0	0	0	0	**1**	0	0	0	0	0	0	0

- Can extend to be invariant of clique size

Max-margin Learning for Lower Linear Envelope MRFs

QP for lower linear envelope MRF learning

$$\begin{array}{lll} \text{minimize}_{\boldsymbol{\theta},\boldsymbol{\xi}} & \frac{1}{2}\|\boldsymbol{\theta}\|_2^2 + \frac{C}{T}\sum_t \xi_t & \\ \text{subject to} & \boldsymbol{\theta}^T \phi_t(\mathbf{y}) - \boldsymbol{\theta}^T \phi_t(\mathbf{y}_t) \geq \Delta(\mathbf{y}, \mathbf{y}_t) - \xi_t, & \forall t, \mathbf{y} \in \mathcal{Y}_t \\ & \xi_t \geq 0, & \forall t \\ & \mathbf{D}^2\boldsymbol{\theta} \geq 0 & \end{array}$$

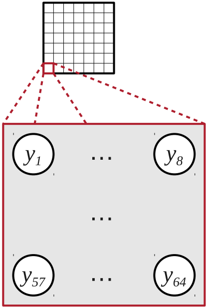
Max-margin Learning for Lower Linear Envelope MRFs

QP for lower linear envelope MRF learning

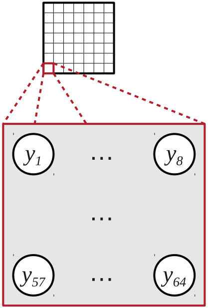
$$
\begin{array}{lll}
\text{minimize}_{\boldsymbol{\theta},\boldsymbol{\xi}} & \frac{1}{2}\|\boldsymbol{\theta}\|_2^2 + \frac{C}{T}\sum_t \xi_t & \\
\text{subject to} & \boldsymbol{\theta}^T \phi_t(\mathbf{y}) - \boldsymbol{\theta}^T \phi_t(\mathbf{y}_t) \geq \Delta(\mathbf{y}, \mathbf{y}_t) - \xi_t, & \forall t, \mathbf{y} \in \mathcal{Y}_t \\
 & \xi_t \geq 0, & \forall t \\
 & \mathbf{D}^2 \boldsymbol{\theta} \geq 0 &
\end{array}
$$

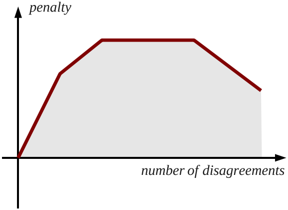
- Learning algorithm repeatedly
  - solves above QP using sampled representation $\boldsymbol{\theta}$
  - finds violated constraints using lower linear envelope representation $\{(a_k, b_k)\}$
- Variants of the feature representation and corresponding learning objective can also be used.

Synthetic Experiments



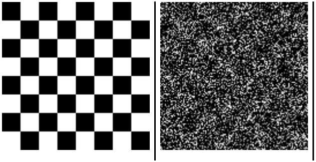
Synthetic Experiments



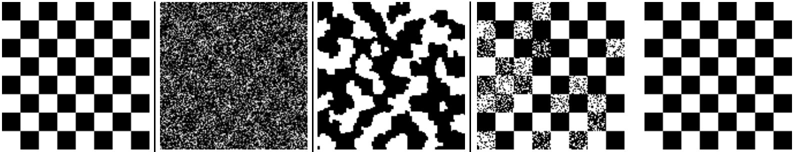


- In these experiments the ground-truth location of the squares are given.

Synthetic Results

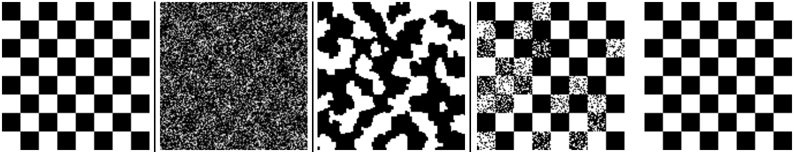


Synthetic Results

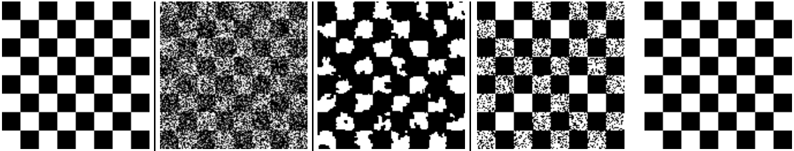


groundtruth | data | pairwise crf | 3rd iteration final iteration

Synthetic Results



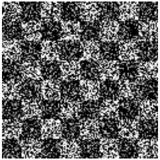
groundtruth | data | pairwise crf | 3rd iteration | final iteration



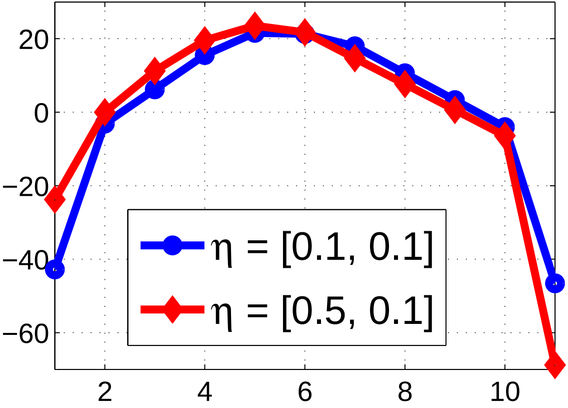
groundtruth | data | pairwise crf | 3rd iteration | final iteration

Learned Parameters for Synthetic Experiments

$\eta = (0.1, 0.1)$



$\eta = (0.5, 0.1)$



- η is the signal-to-noise ratio.

Interactive Image Segmentation

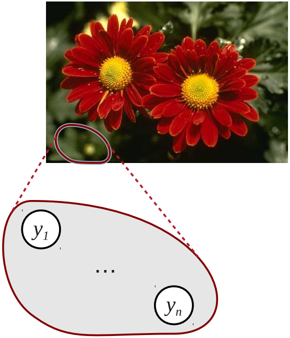
- "GrabCut" [Rother et al., SIGGRAPH 2004]

Interactive Image Segmentation

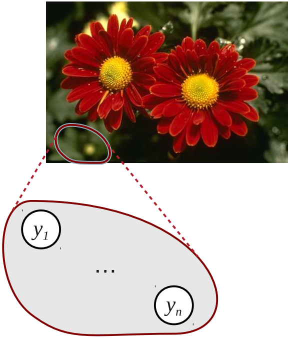
- "GrabCut" [Rother et al., SIGGRAPH 2004]

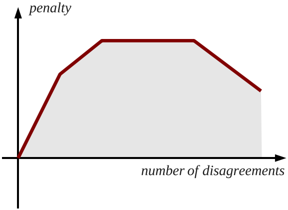
- Our experimental setup
  - leave-one-out cross-validation on 50 images
  - baseline: 8-neighborhood pairwise CRF
  - higher-order: lower linear envelope potential on non-overlapping superpixels

"GrabCut" Experiments



"GrabCut" Experiments





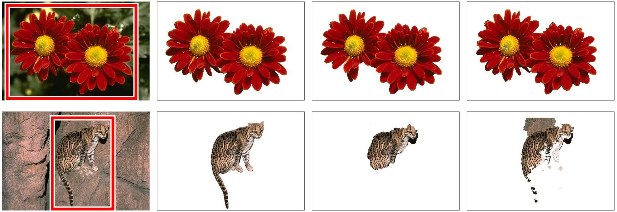
- Superpixels determined via a bottom-up unsupervised approach.

"GrabCut" Results

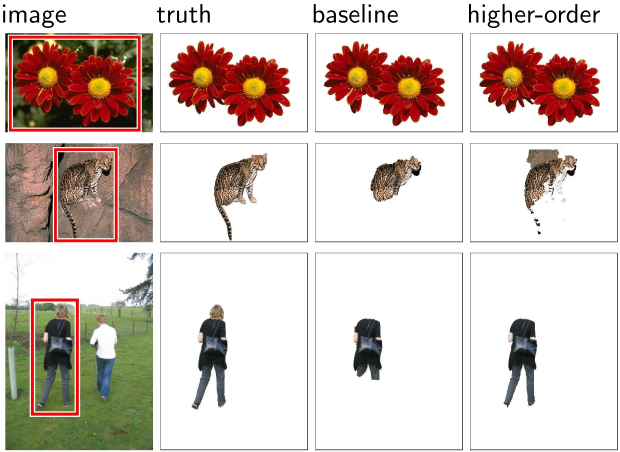
image | truth | baseline | higher-order

"GrabCut" Results

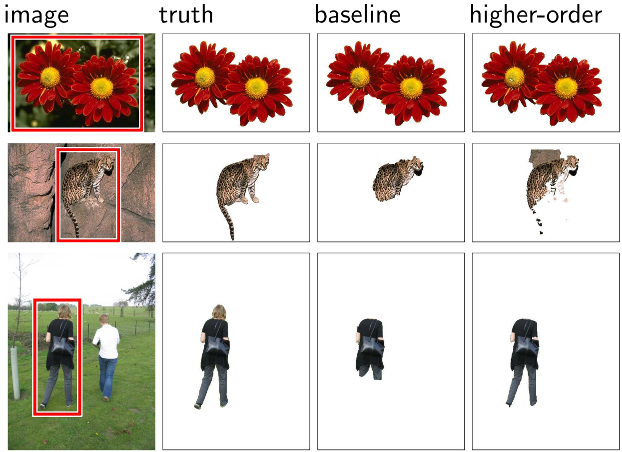
image truth baseline higher-order



“GrabCut” Results



"GrabCut" Results



- Quantitatively we see a 15% reduction in error rate.
- Simply enforcing hard consistency within superpixels results in 1% increase in error rate.

Summary

- **motivation**
  - higher-order models are important for image understanding

Summary

- **motivation**
  - higher-order models are important for image understanding
- **this work—binary lower linear envelope potentials**
  - telescoping-sum construction for exact MAP inference in time polynomial in the number of variables and number of linear envelope functions
  - representation for learning parameters of lower linear envelope potentials using max-margin framework
  - demonstrated in the context of figure-ground segmentation

Summary

- **motivation**
  - higher-order models are important for image understanding
- **this work—binary lower linear envelope potentials**
  - telescoping-sum construction for exact MAP inference in time polynomial in the number of variables and number of linear envelope functions
  - representation for learning parameters of lower linear envelope potentials using max-margin framework
  - demonstrated in the context of figure-ground segmentation
- **future work**
  - apply to multi-class setting
  - explore relationship with latent-variable SVMs

Summary

- **motivation**
  - higher-order models are important for image understanding
- **this work—binary lower linear envelope potentials**
  - telescoping-sum construction for exact MAP inference in time polynomial in the number of variables and number of linear envelope functions
  - representation for learning parameters of lower linear envelope potentials using max-margin framework
  - demonstrated in the context of figure-ground segmentation
- **future work**
  - apply to multi-class setting
  - explore relationship with latent-variable SVMs
- **questions?**