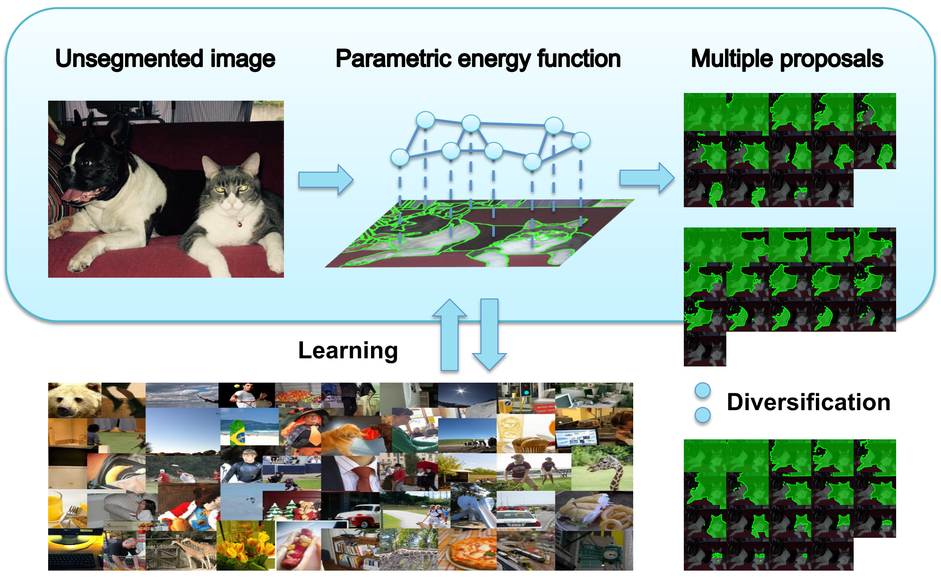
Learning to Combine Mid-level Cues for Object Proposal Generation

Tom Lee, Sanja Fidler, Sven Dickinson

UNIVERSITY OF TORONTO



Motivation

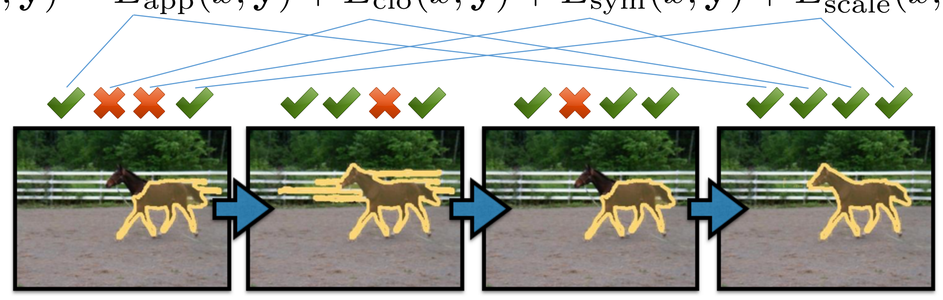
- Object proposals reduce an exhaustive set of hypotheses to a few plausible candidate segments.
- Object proposals are often predictions from parametric energy functions (CPMC [2] etc.)
- Parametric energy functions can encode relevant bottom-up grouping cues [4].
- But no previous approach exists for learning to predict multiple outputs with parametric energy functions.

Contributions

- A novel Parametric Min-Loss (PML) structured learning framework for parametric energy functions.
- PML learns to predict multiple outputs using a novel loss function.
- PML bridges the gap between learning and inference for parametric energy functions.
- PML is applicable to any domain that uses parametric energy functions.

Parametric energy function

$$E^{\lambda}(x, \mathbf{y}) = E_{\text{app}}(x, \mathbf{y}) + E_{\text{clo}}(x, \mathbf{y}) + E_{\text{sym}}(x, \mathbf{y}) + E^{\lambda}_{\text{scale}}(x, \mathbf{y})$$



- The appearance cue discourages division of similar colors and textures:

$$E_{\text{app}}(x, \mathbf{y}) = \mathbf{w}_{\text{app}}^{\mathsf{T}} \sum_{p,q} \phi_{\text{app}}(x, \mathbf{y}_{p,q})$$

$$\phi_{\text{app}}(x, \mathbf{y}_{p,q}) = \mathbb{1}_{[y_p \neq y_q]}(\text{sim}_{p,q}(\mathbf{h}^{\text{col}}), \text{sim}_{p,q}(\mathbf{h}^{\text{text}}))$$

- The closure cue discourages gaps along boundaries:

$$E_{\text{clo}}(x, \mathbf{y}) = w_{\text{clo}}^{\mathsf{T}} \left(\sum_{p} \phi_{\text{clo}}(x, y_p) - 2 \sum_{p,q} \phi_{\text{clo}}(x, \mathbf{y}_{p,q}) \right)$$

$$\phi_{\text{clo}}(x, y_p) = \sum_{b \in \partial(p)} \mathbb{1}_{[y_p=1]} g(x, b)$$

$$\phi_{\text{clo}}(x, \mathbf{y}_{p,q}) = \sum_{b \in \bar{\partial}(p,q)} \mathbb{1}_{[y_p=y_q=1]} g(x, b)$$

- The symmetry cue discourages division of symmetric parts:

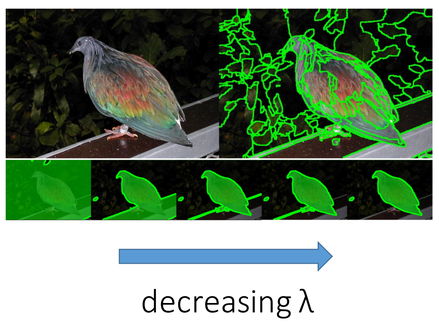
$$E_{\text{sym}}(x, \mathbf{y}) = w_{\text{sym}}^{\mathsf{T}} \sum_{p,q} \phi_{\text{sym}}(x, \mathbf{y}_{p,q})$$

$$\phi_{\text{sym}}(x, \mathbf{y}_{p,q}) = \mathbb{1}_{[y_p \neq y_q]} \max_{s \in S(p,q)} \text{score}(s)$$

- The energy is normalized by area by a factor λ:

$$E^{\lambda}_{\text{scale}}(x, \mathbf{y}) = \lambda \sum_{p} \phi_{\text{area}}(x, y_p)$$

Multiple-output prediction



decreasing λ

- Nonnegative weights and nonnegative λ coefficients guarantee a small set of solutions from parametric maxflow.
- One prediction for a specific λ:

$$\hat{\mathbf{y}} = \arg\min_{\mathbf{y}} E^{\lambda}(x, \mathbf{y}, \mathbf{w})$$

$$\hat{\mathbf{y}}(x, \mathbf{w}) = \arg\max_{\mathbf{y}} \mathbf{w}^{\mathsf{T}} \phi^{\lambda}(x, \mathbf{y})$$

- A set of predictions over a range of λ:

$$\hat{Y}(x, \mathbf{w}) = \{\hat{\mathbf{y}}^{\lambda}(x, \mathbf{w}) : \lambda \in [-1, 0]\}$$

Parametric Min-Loss learning

- Evaluate multiple predicted segments against one correct ground truth segment.
- Loss function ideally expresses a "min":

$$\mathcal{L}(\hat{Y}, \mathbf{y}) = \min_{\hat{\mathbf{y}} \in \hat{Y}} \ell(\hat{\mathbf{y}}, \mathbf{y})$$

- Inner loss function measures the error of a single predicted segment:

$$\ell(\hat{\mathbf{y}}, \mathbf{y}(g)) = \frac{1}{|g|} \sum_{p} |p| \begin{cases} v_p & \hat{y}_p = 0 \\ 1 - v_p & \hat{y}_p = 1, \end{cases}$$

- Upper bound for inner loss function (hinge loss):

$$h(\mathbf{w}, \lambda) = \max_{\hat{\mathbf{y}}} \ell(\hat{\mathbf{y}}, \mathbf{y}) + \mathbf{w}^{\mathsf{T}} \phi^{\lambda}(x, \hat{\mathbf{y}}) - \mathbf{w}^{\mathsf{T}} \phi^{\lambda}(x, \mathbf{y})$$

- Upper bound for loss function (min-hinge loss [3]):

$$H(\mathbf{w}) = \min_{\lambda \in [-1,0]} h(\mathbf{w}, \lambda)$$

- Regularized training objective:

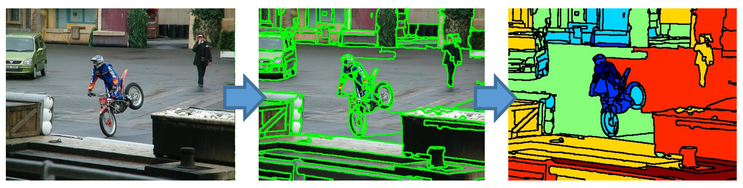
$$\min_{\mathbf{w}} \frac{1}{2}||\mathbf{w}||^2 + \frac{C}{N} \sum_{n=1}^{N} \min_{\lambda_n \in [-1,0]} h_n(\mathbf{w}, \lambda_n)$$

Block-coordinate descent

- w-block (S-SVM): $\arg\min_{\mathbf{w}} \frac{1}{2}||\mathbf{w}||^2 + \frac{C}{N} \sum_{n=1}^{N} h_n(\mathbf{w}, \lambda_n)$
- λ-block (loss-augmented parametric energy minimization):

$$\arg\min_{\lambda \in [-1,0]} h(\mathbf{w}, \lambda) \qquad \forall \lambda \in [-1, 0], \min_{\hat{\mathbf{y}}} -\ell(\hat{\mathbf{y}}, \mathbf{y}) - \mathbf{w}^{\mathsf{T}} \phi^{\lambda}(x, \hat{\mathbf{y}})$$

Location- and color-based diversification



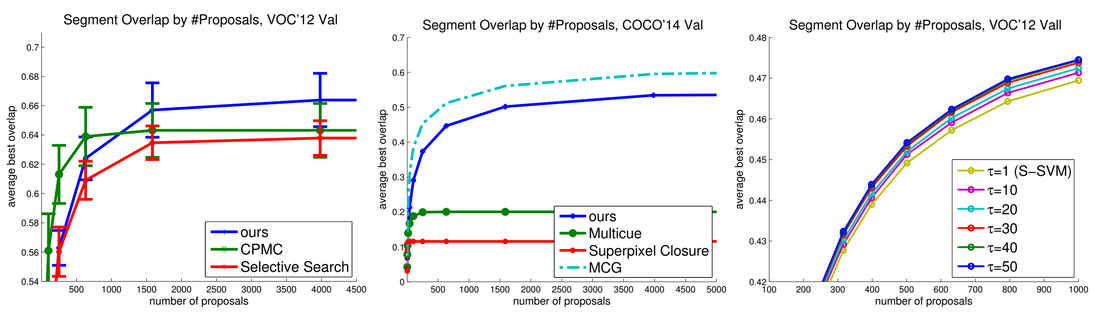
- Bias energy to different locations
- Maximum superpixel distance

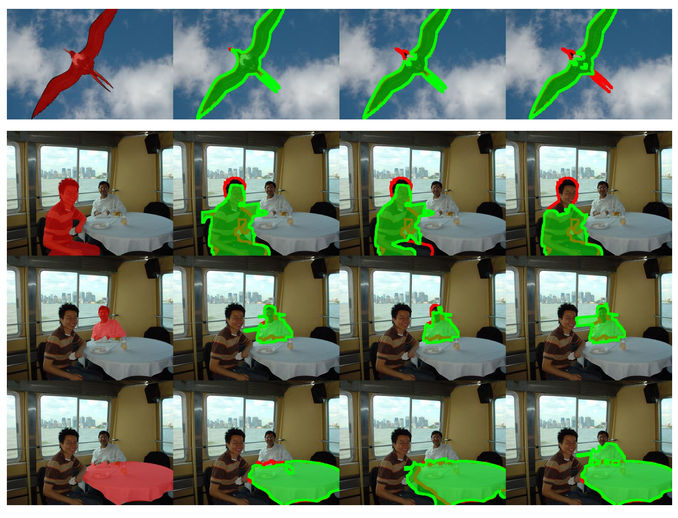
- Bias energy to different foreground-background color pairs
- Gaussian mixture model of superpixel colors

Postprocessing

- Discard non-maximum proposals among proposals with high overlap.
- Train SVM on deep features to assign an objectness score to each proposal.

Results





- We achieve results comparable with CPMC [2] and MCG [1]
- We outperform methods that lack learning, e.g. Selective Search [5]

[1] Arbelaez et al., CVPR 2014.
[2] Carreira & Sminchisescu, PAMI 2012.
[3] Guzman-Rivera et al., NIPS 2012.
[4] Lee et al., ACCV 2014.
[5] Uijlings et al., IJCV 2013.