Putting MAP back on the Map

Patrick Pletscher, Sebastian Nowozin, Pushmeet Kohli, Carsten Rother

ETH Eidgenössische Technische Hochschule Zürich Swiss Federal Institute of Technology Zurich

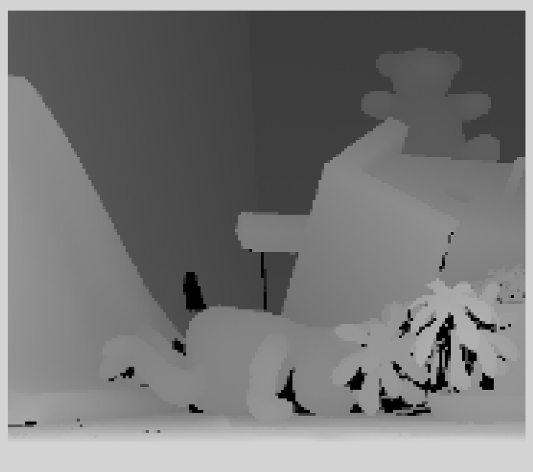
Microsoft Research

Energy Minimization in Computer Vision

Labeling $\mathbf{y}$ given an input image $\mathbf{x}$. Energy minimization of the form

$$\mathbf{y}^* = \operatorname*{argmin}_{\mathbf{y}} \sum_{i \in \mathcal{V}} \psi_i(\mathbf{x}, y_i) + \sum_{(i,j) \in \mathcal{E}} \psi_{ij}(\mathbf{x}, y_i, y_j).$$

semantic segmentation



stereo vision

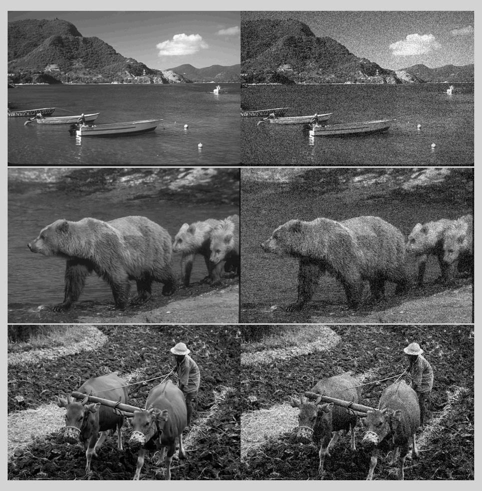


image denoising

How to choose ψ_i and ψ_{ij}? Learn them from data!

Our Work

- Empirical comparison of different learning and prediction paradigms.
- How well do the different approaches reproduce image statistics?
- Full learning of the potentials for image denoising.
- Study influence of misspecified models.

Running Example: A Simple Model for Image Denoising

Energy of a denoised image $\mathbf{y}$ for given image $\mathbf{x}$ and parameters $\mathbf{w}$:

$$E(\mathbf{y}, \mathbf{x}, \mathbf{w}) = -\sum_{i \in \mathcal{V}} w^u_{|y_i - x_i|} - \sum_{(i,j) \in \mathcal{E}} w^p_{|y_i - y_j|}.$$

Can be written in a linearized form $E(\mathbf{y}, \mathbf{x}, \mathbf{w}) = -\langle \mathbf{w}, \mathbf{s}(\mathbf{x}, \mathbf{y}) \rangle$ with:

- $\mathbf{s}(\mathbf{x}, \mathbf{y}) = [\mathbf{s}^u(\mathbf{x}, \mathbf{y})^\top, \mathbf{s}^p(\mathbf{y})^\top]^\top$ and
- $s^u_k(\mathbf{x}, \mathbf{y}) = \sum_{i \in \mathcal{V}} \delta_k(|x_i - y_i|)$, $s^p_k(\mathbf{y}) = \sum_{(i,j) \in \mathcal{E}} \delta_k(|y_i - y_j|)$.

Learning and Prediction: Overview

- Conditional Gibbs energy of a labeling/image pair:

$$P(\mathbf{y}|\mathbf{x}, \mathbf{w}) = \frac{1}{Z(\mathbf{x}, \mathbf{w})} \exp(-E(\mathbf{y}, \mathbf{x}, \mathbf{w})).$$

- *Loss* $\Delta(\mathbf{y}, \mathbf{y}^*)$, error when predicting $\mathbf{y}$ instead of $\mathbf{y}^*$.
  - Pixelwise zero-one error: $\Delta(\mathbf{y}, \mathbf{y}^*) = \sum_{i \in \mathcal{V}} [1 - \delta_{y_i^*}(y_i)]$.
  - Pixelwise mean squared error: $\Delta(\mathbf{y}, \mathbf{y}^*) = \sum_{i \in \mathcal{V}} (y_i - y_i^*)^2$.
  - Full image zero-one error: $\Delta(\mathbf{y}, \mathbf{y}^*) = [1 - \delta_{\mathbf{y}^*}(\mathbf{y})]$.
- Pre-dominant approaches in the literature:

$$\begin{tikzcd} \{(\mathbf{x}^n, \mathbf{y}^n)\}_{n=1}^N \arrow[dr] & & \Delta \arrow[d] & \mathbf{x} \arrow[d] \\ E(\mathbf{y}, \mathbf{x}, \mathbf{w}) \arrow[r] & P(\mathbf{y}|\mathbf{x}, \mathbf{w}^*) \arrow[r] & f_{\mathbf{w}^*}(\mathbf{x}) \arrow[r] & \mathbf{y}^* \end{tikzcd}$$

Incorporate loss only at predict time

$$\begin{tikzcd} & \{(\mathbf{x}^n, \mathbf{y}^n)\}_{n=1}^N \arrow[dr] & \Delta \arrow[d] & \mathbf{x} \arrow[d] \\ E(\mathbf{y}, \mathbf{x}, \mathbf{w}) \arrow[rr] & & f_{\mathbf{w}^*}(\mathbf{x}) \arrow[r] & \mathbf{y}^* \end{tikzcd}$$

Directly train a prediction function

Optimal Prediction for a Given Posterior

- Given the *true posterior* $P(\mathbf{y}|\mathbf{x})$. How should we predict?
- Depends on the *loss* $\Delta(\mathbf{y}, \mathbf{y}^*)$: error when predicting $\mathbf{y}$ instead of $\mathbf{y}^*$.

$$\begin{tikzcd}[arrows=dash] & P(\mathbf{y}|\mathbf{x}) \arrow[dl] \arrow[d] \arrow[dr] & \\ \text{zero-one} \arrow[d] & \text{pixelwise zero-one} \arrow[d] & \text{pixelwise squared} \arrow[d] \\ \text{MAP} & \text{max marginal} & \text{MMSE} \end{tikzcd}$$

- Best predictor: minimize the risk:

$$\mathbf{y}(\mathbf{x}) = \operatorname*{argmin}_{\mathbf{y}} \sum_{\mathbf{y}'} \Delta(\mathbf{y}, \mathbf{y}') P(\mathbf{y}'|\mathbf{x}).$$

- Maximum-A-Posteriori (MAP) prediction:

$$\mathbf{y}^* = \operatorname*{argmax}_{\mathbf{y}} E(\mathbf{y}, \mathbf{x}, \mathbf{w}).$$

  Efficient approximate algorithms exist (graph-cut or belief propagation).
- Minimum Mean Squared Error (MMSE) prediction:

$$y_i^* = \mathbb{E}_{P(y_i|\mathbf{x})}[y_i] \quad \forall i \in \mathcal{V}.$$

  Marginals for example computed using Gibbs sampling.
- Problem: In reality we do not have the true posterior!

Learning

- Maximum Likelihood (MLE):

$$\mathbf{w}^{mle} = \operatorname*{argmin}_{\mathbf{w}} -\frac{1}{N} \sum_{n=1}^{N} \log P(\mathbf{y}^n|\mathbf{x}^n, \mathbf{w}) + \frac{\lambda}{2} \|\mathbf{w}\|^2.$$

- Maximum Pseudo-likelihood (MPLE), tractable MLE approximation:

$$\mathbf{w}^{mple} = \operatorname*{argmin}_{\mathbf{w}} -\frac{1}{N} \sum_{n=1}^{N} \sum_{i \in \mathcal{V}} \log P(y_i^n|\mathbf{y}^n_{\mathcal{N}(i)}, \mathbf{x}^n, \mathbf{w}) + \frac{\lambda}{2} \|\mathbf{w}\|^2.$$

- Maximum Margin (MM):

$$\mathbf{w}^{mm} = \operatorname*{argmin}_{\mathbf{w}} \frac{1}{N} \sum_{n=1}^{N} \max_{\mathbf{y}} [\langle \mathbf{w}, \mathbf{s}(\mathbf{x}^n, \mathbf{y}) - \mathbf{s}(\mathbf{x}^n, \mathbf{y}^n) \rangle + \Delta(\mathbf{y}, \mathbf{y}^n)] + \frac{\lambda}{2} \|\mathbf{w}\|^2.$$

Insights on Image Statistics Matching

- Without regularization MLE can be understood as statistics matching:

$$\frac{1}{N} \sum_{n=1}^{N} \mathbb{E}_{P(\mathbf{y}|\mathbf{x}^n, \mathbf{w}^{mle})}[\mathbf{s}(\mathbf{x}^n, \mathbf{y})] = \frac{1}{N} \sum_{n=1}^{N} \mathbf{s}(\mathbf{x}^n, \mathbf{y}^n).$$

- However: *Image statistics only matched in expectation, not for a single labeling.*

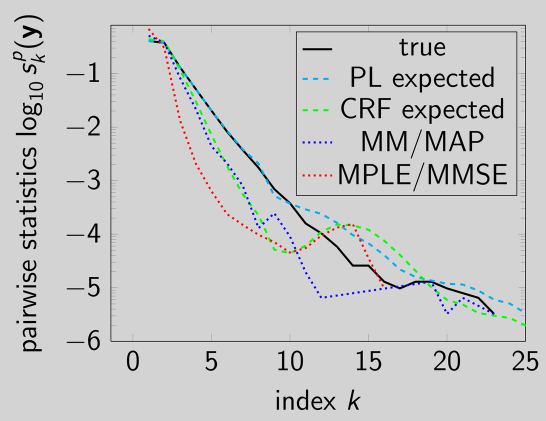
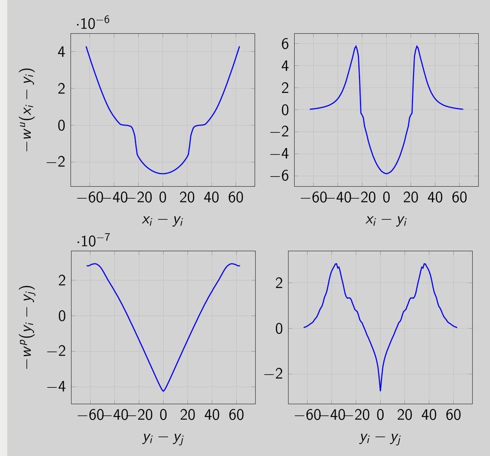
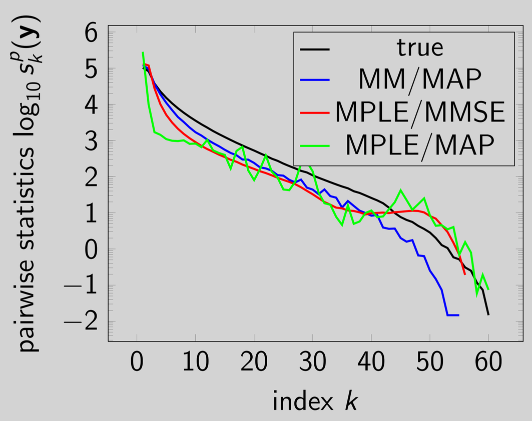


Image Denoising

noisy image,20.29dB — original image — MM/MAP,26.03dB — MPLE/MMSE,25.44dB — MPLE/MAP,23.11dB



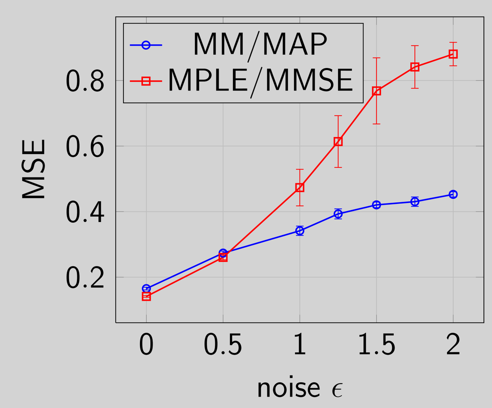


	standard model		with BM3D	
method	MSE	PSNR	MSE	PSNR
MM/MAP	**8.65**	**27.05**	**6.86**	**28.23**
MPLE/MAP	17.42	24.30	13.31	25.5
MPLE/MMSE	10.04	26.65	8.47	27.54
BM3D only	-	-	6.95	28.19

- Berkeley dataset with $\sigma = 25$ noise level. Reduced to 64 labels.
- MM/MAP performs at least as good as MPLE/MMSE.
- Linear pairwise costs are optimal for MAP.

Synthetic Data: Misspecified Models

- Synthetic experiment: Data generating model different than assumed model.
- Amount of misspecification controlled by ϵ.
- MM/MAP more robust to misspecifications than MPLE/MMSE.



Conclusions

- Only use MAP if model trained using max-margin!
- Properly trained MAP performs on par with MMSE.
- MMSE not better suited for reproducing image statistics.

References

- I. Tsochantaridis et al. (2004). "Support vector machine learning for interdependent and structured output spaces". In: *ICML*, p. 104
- Taskar, Guestrin, and Koller (2003). "Max-Margin Markov Networks". In: *NIPS*
- John Lafferty, Andrew Mccallum, and Fernando Pereira (2001). "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data". In: *ICML*, pp. 282–289
- Uwe Schmidt, Qi Gao, and Stefan Roth (2010). "A Generative Perspective on MRFs in Low-Level Vision". In: *CVPR*
- D. Martin et al. (2001). "A Database of Human Segmented Natural Images". In: *ICCV*
- Vladimir Kolmogorov (2006). "Convergent tree-reweighted message passing for energy minimization." In: *PAMI*