Discriminative Dictionary Learning with Spatial Priors

Nazar Khan[1,2]   Marshall Tappen[2]

[1]Lahore University of Management Sciences, Pakistan

[2]University of Central Florida, USA

ICIP 2013, Melbourne, Australia
September 18, 2013

Introduction

- Traditional sparse coding has assumed independent image patches.
- But real-world image patches are not independent – a Markovian dependency (*i.e.* spatial prior) is often assumed.
- We show how spatial priors can be incorporated for learning dictionaries.
- We retain discriminability in the spatial prior.

Sparse Coding

- Finding the sparse vector of coefficients $\mathbf{s}^*$ in an over-complete basis.

$$\mathbf{y} \xrightarrow{\mathbf{D}} \mathbf{s}^*$$

  where $|\mathbf{s}^*| > |\mathbf{y}|$ and $\mathbf{s}^*$ is sparse.
- Basis for new space is the so called dictionary $\mathbf{D}$.
- We use ℓ_1-sparse coding

$$\mathbf{s}^* = \arg\min_{\mathbf{s}} \underbrace{||\mathbf{y} - \mathbf{D}\mathbf{s}||_F^2}_{\text{reconstruction error}} + \underbrace{\lambda||\mathbf{s}||_1}_{\text{sparsity constraint}}$$

Dictionary Learning (DL)

- Finding the over-complete basis $\mathbf{D}^*$ that optimally reconstructs a set $\mathbf{Y}$ of signals *in a sparse coding manner*.

$$\underbrace{\{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_N\}}_{\mathbf{Y}} \longrightarrow \left\{ \begin{array}{c} \mathbf{D}^* \\ \mathbf{s}_1^*, \mathbf{s}_2^*, \ldots, \mathbf{s}_N^* \end{array} \right\}$$

- Find dictionary as well as sparse codes that optimally reconstruct the set $\mathbf{Y}$.
- Formally,

$$\mathbf{D}^*, \mathbf{A}^* = \arg\min_{\mathbf{D},\mathbf{S}} \frac{1}{2}||\mathbf{Y} - \mathbf{DS}||_F^2 + \lambda \sum_{j=1}^{N} ||\mathbf{s}_j||_1$$

Classification via Dictionaries

- Training (N classes)

$$\begin{array}{cccc} \text{Class } 1 & \text{Class } 2 & \dots & \text{Class } N \\ \uparrow & \uparrow & \dots & \uparrow \\ \mathbf{D}_1 & \mathbf{D}_2 & \dots & \mathbf{D}_N \end{array}$$

- Testing (signal $\mathbf{y}$)

$$\arg \min_{i \in \{1 \dots N\}} \mathcal{R}_i$$

where $\mathcal{R}_i = \frac{1}{2}||\mathbf{y} - \mathbf{D}_i \mathbf{s}_i^*||_F^2$.

- Learn dictionaries for each class and classify test signal into class with least reconstruction error.
- **Final goal is classification but dictionaries are learned in a reconstructive manner.**

Reconstructive vs. Discriminative Dictionary Learning

Reconstructive

	Class 1	Class 2	Class3
$\mathbf{D}_1$	✓	?	?
$\mathbf{D}_2$	?	✓	?
$\mathbf{D}_3$	?	?	✓

$\mathbf{D}_i$ good for class i but nothing stops it from being good for some other class too.

Discriminative

	Class 1	Class 2	Class3
$\mathbf{D}_1$	✓	×	×
$\mathbf{D}_2$	×	✓	×
$\mathbf{D}_3$	×	×	✓

$\mathbf{D}_i$ good for class i and bad for other classes.

Data Term: Discriminative Deviation

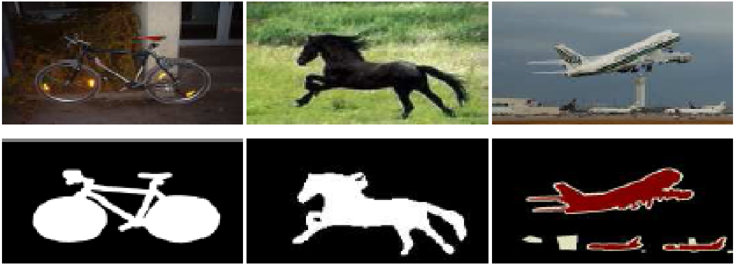
- Let $\mathcal{R}_i = \frac{1}{2}||\mathbf{y} - \mathbf{D}_i\mathbf{s}_i||^2$ be the reconstruction error of signal $\mathbf{y}$ under dictionary $\mathbf{D}_i$.
- For the vector of reconstruction errors $\mathcal{R} = [\mathcal{R}_1, \mathcal{R}_2, \ldots, \mathcal{R}_N]$, define *discriminative deviation*

$$\mathcal{D}_t = \mathcal{R}_t - \bar{\mathcal{R}}$$

  where t is the true class.
- Minimizing $\mathcal{D}_t$ encourages reconstruction error for the true class to be lower than those for all other classes.
- Alternatively, dictionaries for other classes should do worse than true class.
- Joint, discriminative learning over all classes.

Smoothness Term

- Real-world images are characterized by a spatial smoothness prior.
  - Even stronger prior for image labellings.



- The DL problem should respect this prior.

Learning with Spatial Priors

- If adjacent labels are same, sparse codes under this label should be more similar than under all other labels of the neighbor.
  - Use discriminative deviation function again.
- If adjacent labels are different, sparse codes under both labels should not be similar. Leads to *boundary preservation*.

$$\psi(s_i, s_j) \propto \begin{cases} -\mathcal{D}(s_i^T s_j), & \text{if labels are same} \\ s_i^T s_j, & \text{if labels are different} \end{cases}$$

- Spatial prior $\implies$ CRF Energy Formulation

Learning with Spatial Priors

- During learning, prior is
  - useful for dictionary learning because it includes the sparse codes,
  - discriminative because it is label-dependent,
- During inference, prior is
  - boundary preserving

CRF Energy Formulation

- Consider image $\mathbf{I}$ as a *structured* grid ($\mathbf{I} \longrightarrow G(\mathcal{V}, \mathcal{E})$) with labelling $\mathbf{y}$ corresponding to C classes.
- When $\mathbf{y}$ represents ground-truth, we want

$$\min_{\{\mathbf{D}\}_1^C} \sum_{\mathbf{I} \in \text{training images}} \underbrace{\left(\sum_{i \in \mathcal{V}} \mathcal{D}_{\mathbf{y}_i} + \mathcal{R}_{\mathbf{y}_i} + \sum_{(i,j) \in \mathcal{E}} \psi_{ij} \right)}_{E(\mathbf{y}, \mathbf{I}, \{\mathbf{D}\}_1^C)}$$

- $\mathcal{D}_{\mathbf{y}_i}$ encourages discrimination.
- $\mathcal{R}_{\mathbf{y}_i}$ encourages reconstruction.
- ψ_{ij} encourages spatial coherence with boundary preservation.
- All three objectives can be weighted by $\boldsymbol{\kappa} = \{\boldsymbol{\kappa}_{\mathcal{D}}, \boldsymbol{\kappa}_{\mathcal{R}}, \boldsymbol{\kappa}_{\psi}\}$.
- Parameters to be learned are the dictionaries $\{\mathbf{D}\}_1^C$ and the CRF parameters $\boldsymbol{\kappa}$.

Learning on a CRF

- $P(\mathbf{y}|\mathbf{I}) \propto e^{-E(\mathbf{y},\mathbf{I})}$.
- Intractable partition function.
  - Maximize *pseudolikelihood* to learn $\{\{\mathbf{D}\}^*, \boldsymbol{\kappa}^*\}$.
- Potential problem with over-smoothness [VSSM06].
  - Handled via learning of optimal $\boldsymbol{\kappa}$.
- Requires gradient of the non-differentiable ℓ_1 sparse coding procedure
  - Use implicit differentiation.

Smoothness ⇒ stability

- Sparse codes with very large entries ⇒ ill-conditioned dictionary [DXW11].
- Conversely, by requiring adjacent sparse codes to be (typically) similar, the dictionaries are encouraged to be well-conditioned.
- This is useful since *discriminative* DL is inherently unstable.
    - Reconstruction-discrimination tradeoff.

Experiments and Results

Pixelwise classification into foreground/background for Graz02 bike dataset.

Data Term+Prior		Data Term		Shape Mask
Ours	[YY12]	[KT12]	[RSS10]	[MS12]
72.1	62.4	69.5	68	61.8

Table: Comparison of Equal Error Rate (EER %) of precision-recall curves for pixel-level classfication of Graz02 bike test set. Our results exceed the state-of-the-art in top-down dictionary learning based approaches and match the bottom-up super-pixel based segmentation accuracy from [FVS09].

Pixelwise classification on Graz02 bike dataset

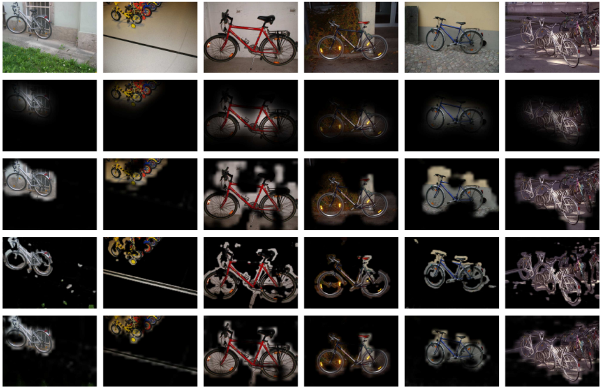
Original

Data Term + Post Filtering [KT12]

CRF + Potts model [YY12]

Ours

Ours (coarser grid)



Learning the spatial prior is beneficial

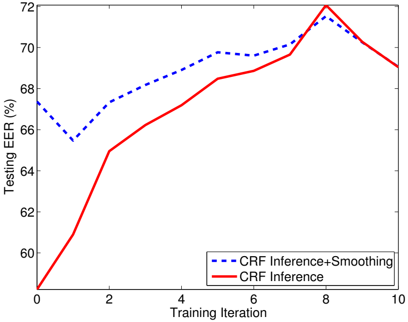


Figure: Benefit of training iterations on the equal error rate (EER) of the precision-recall curve of the test data for Graz02 bike category. Our learning procedure (in red) without additional smoothing was able to learn CRF parameters that out-perform manual smoothing after 8 iterations.

$\mathbf{D}^*$ and $\boldsymbol{\kappa}^*$ in isolation

	No Spatial Term	$\boldsymbol{\kappa}_0$	$\boldsymbol{\kappa}^*$
$\mathbf{D_0}$	55.1	58.2	66.7
$\mathbf{D}^*$	62.3	63.2	72.1

Table: **Column-wise**: For inference, learned $\boldsymbol{\kappa}$ is better than fixed $\boldsymbol{\kappa}$ which is better than unary beliefs. **Row-wise**: DDL with spatial priors is better than fixed k-means dictionaries, *even when inferring without a spatial prior (*62.3% *vs.* 55.1%*)*.

Other datasets

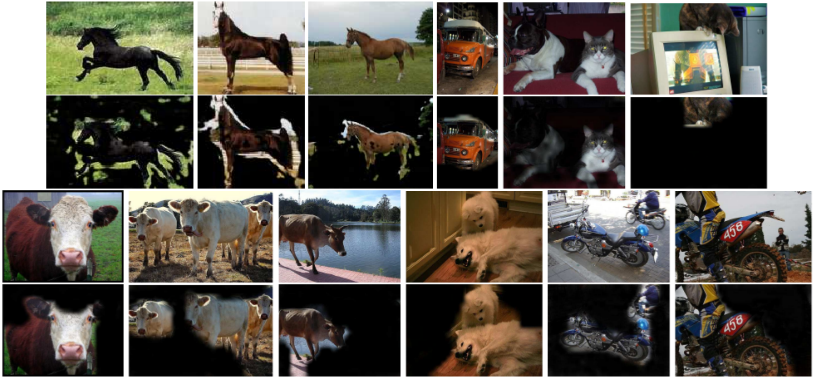


Figure: Some sample results on the Weizmann Horse dataset and VOC 2007 dataset.

VOC 2007 dataset

Class	KSVD[AEB05]	Ours
aeroplane	35.2	**43.7**
bicycle	28.3	**41.2**
bird	35.3	**42.3**
boat	26.3	**35.5**
bottle	16.1	**30.2**
bus	43.7	**69.0**
car	29.1	**43.2**
cat	39.9	**63.3**
chair	9.1	**10.6**
cow	46.0	**70.0**

Table: EER values for figure-ground segmentation on the VOC 2007 dataset.

VOC 2007 dataset

Class	KSVD[AEB05]	Ours
dining table	38.8	**52.7**
dog	33.3	**51.5**
horse	36.6	**42.0**
motorbike	47.2	**62.9**
person	28.3	**43.0**
potted plant	23.0	**31.4**
sheep	47.5	**54.3**
sofa	21.8	**28.0**
train	54.3	**74.0**
tv/monitor	16.3	**29.1**

Table: EER values for figure-ground segmentation on the VOC 2007 dataset.

Conclusion

- A spatial smoothness prior is beneficial for learning discriminative dictionaries for the pixel classification task.
- Issues raised:
  - Structures can exist at multiple scales. Are pairwise, single scale spatial constraints too restrictive?
  - In the language of the seminal sparse coding works by Field *et al.* [OF96, OF97]
    - do simple-cell receptive field properties still emerge when sparsity *and spatial constraints* are used for learning?

References I

- Michal Aharon, Michael Elad, and Alfred Bruckstein, *K-svd: Design of dictionaries for sparse representation*, SPARS, 2005.
- Wei Dai, Tao Xu, and Wenwu Wang, *Simultaneous codeword optimization (simco) for dictionary update and learning*, CoRR **abs/1109.5302** (2011).
- B. Fulkerson, A. Vedaldi, and S. Soatto, *Class segmentation and object localization with superpixel neighborhoods*, Proc. ICCV, 2009.
- Nazar Khan and Marshall Tappen, *Stable discriminative dictionary learning via discriminative deviation*, ICPR, 2012.
- Marcin Marszalek and Cordelia Schmid, *Accurate object recognition with shape masks*, International Journal of Computer Vision (2012), 191–209.

References II

- Bruno A Olshausen and David J Field, *Emergence of simple-cell receptive field properties by learning a sparse code for natural images*, Nature **381** (1996), no. 6583, 607–609.
- ______, *Sparse coding with an overcomplete basis set: A strategy employed by v1?*, Vision research **37** (1997), no. 23, 3311–3325.
- Ignacio Ramirez, Pablo Sprechmann, and Guillermo Sapiro, *Classification and clustering via dictionary learning with structured incoherence and shared features.*, CVPR, 2010.
- S. V. N. Vishwanathan, Nicol N. Schraudolph, Mark W. Schmidt, and Kevin P. Murphy, *Accelerated training of conditional random fields with stochastic gradient methods*, In ICML, 2006, pp. 969–976.
- Jamie Yang and Ming-Hsuang Yang, *Top-down visual saliency via joint CRF and dictionary learning.*, CVPR, 2012.

Questions?