The University of Western Australia
Achieve International Excellence

Australian National University

Joint Discriminative Bayesian Dictionary and Classifier Learning

Naveed Akhtar, Ajmal Mian and Fatih Porikli
{naveed.Akhtar, fatih.porikli}@anu.edu.au; {naveed.akhtar, ajmal.mian}@uwa.edu.au

IEEE 2017 Conference on Computer Vision and Pattern Recognition
CVPR July 21-26 Honolulu 2017

Introduction:

- Discriminative dictionaries are effective for classification. However, classification performance depends on:
  - Dictionary size (pre-defined),
  - Relationship between the dictionary atoms and the class labels (pre-defined), and
  - Suitable classifier for the sparse codes (often learned separately).
- A Bayesian approach is proposed to learn a discriminative dictionary jointly with a classifier.
- The learning process
  - Automatically determines the required dictionary size,
  - Adaptively learns the relationships between the dictionary atoms and the class labels, and
  - Strongly couples the classifier with the dictionary.

Core intuition:

- In Beta-Bernoulli Process (BP) [1], dictionary atoms relate to the training data under a set of Bernoulli distributions.
- Different sets of Bernoulli distribution are used for separate classes to induce discrimination in the dictionary.
- The same Bernoulli distributions are forced to be used in joint learning of the classifier, to encode class-wise dictionary atom popularity in the classifier.
- The sparse codes of test samples use the popular atoms of the correct class more frequently, thereby easily classified by the classifier.
- The non-parametric nature of the BP is exploited in determining the desired dictionary and classifier sizes.

References:

[1] J. Paisley and L. Carin. Nonparametric factor analysis with beta process priors. ICML, 2009.

Acknowledgement:

This research was supported by ARC Discovery grant DP160101458 and it received funding from ARENA as part of ARENA's Research and Development Programme.

Proposed model:

$\forall i \in \mathcal{I}_c$ and $\forall k \in \mathcal{K} = \{1, ..., K\}$:

$$\mathbf{y}_i^c = \boldsymbol{\Phi}(\mathbf{z}_i^c \odot \mathbf{s}_i^c) + {}^{\mathrm{y}}\boldsymbol{\epsilon}_i \qquad \mathbf{h}_i^c = \boldsymbol{\Psi}(\mathbf{z}_i^c \odot \mathbf{t}_i^c) + {}^{\mathrm{h}}\boldsymbol{\epsilon}_i$$

$$z_{ik}^c \sim \mathrm{Bernoulli}(z_{ik}^c | \pi_{k_o}^c)$$

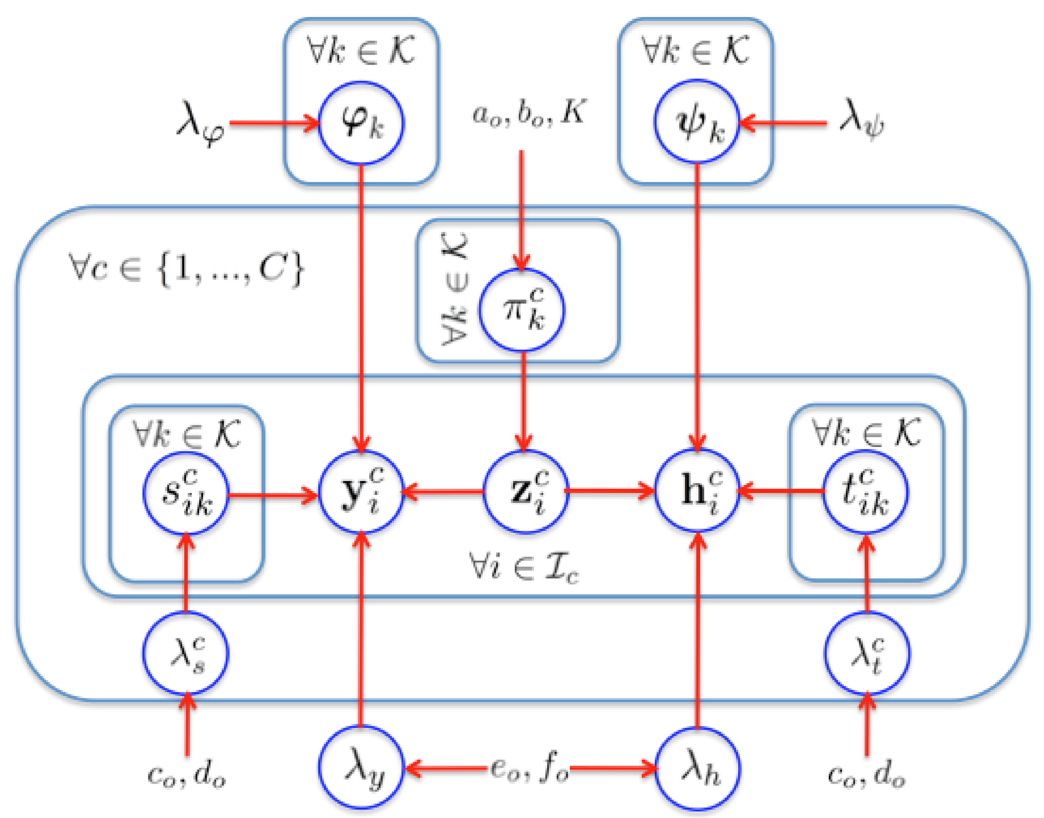
$$\pi_k^c \sim \mathrm{Beta}(\pi_k^c | a_o/K, b_o(K-1)/K)$$

$$s_{ik}^c \sim \mathcal{N}(s_{ik}^c | 0, 1/\lambda_{s_o}^c) \qquad t_{ik}^c \sim \mathcal{N}(t_{ik}^c | 0, 1/\lambda_{t_o}^c)$$

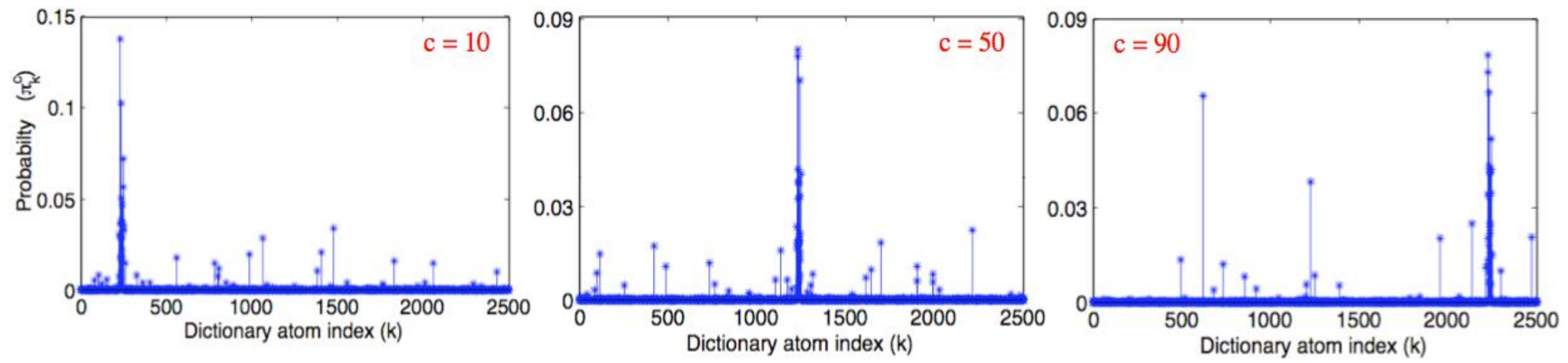
$$\boldsymbol{\varphi}_k \sim \mathcal{N}(\boldsymbol{\varphi}_k | \mathbf{0}, \boldsymbol{\Lambda}_{\varphi_o}^{-1}) \qquad \boldsymbol{\psi}_k \sim \mathcal{N}(\boldsymbol{\psi}_k | \mathbf{0}, \boldsymbol{\Lambda}_{\psi_o}^{-1})$$

$${}^{\mathrm{y}}\boldsymbol{\epsilon}_i \sim \mathcal{N}({}^{\mathrm{y}}\boldsymbol{\epsilon}_i | \mathbf{0}, 1/\lambda_{y_o}\mathbf{I}_L) \qquad {}^{\mathrm{h}}\boldsymbol{\epsilon}_i \sim \mathcal{N}({}^{\mathrm{h}}\boldsymbol{\epsilon}_i | \mathbf{0}, 1/\lambda_{h_o}\mathbf{I}_C)$$

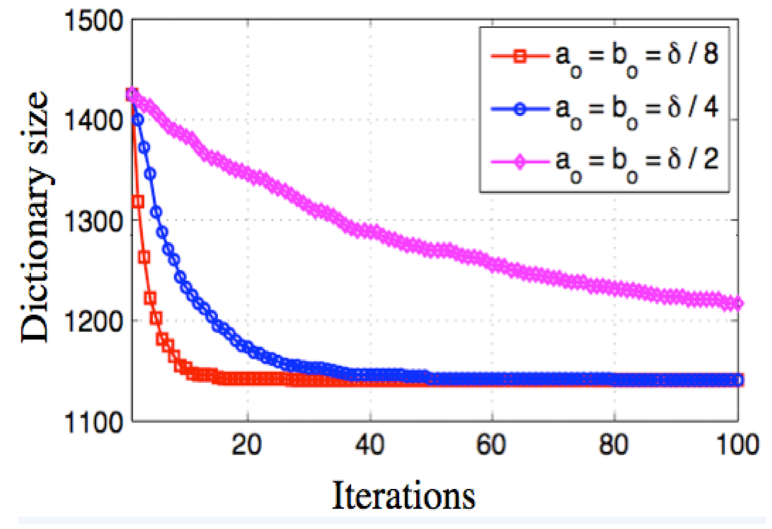
Separate priors over dictionary atoms and classifier parameters, but coupling under the same Beta and Bernoulli distributions.



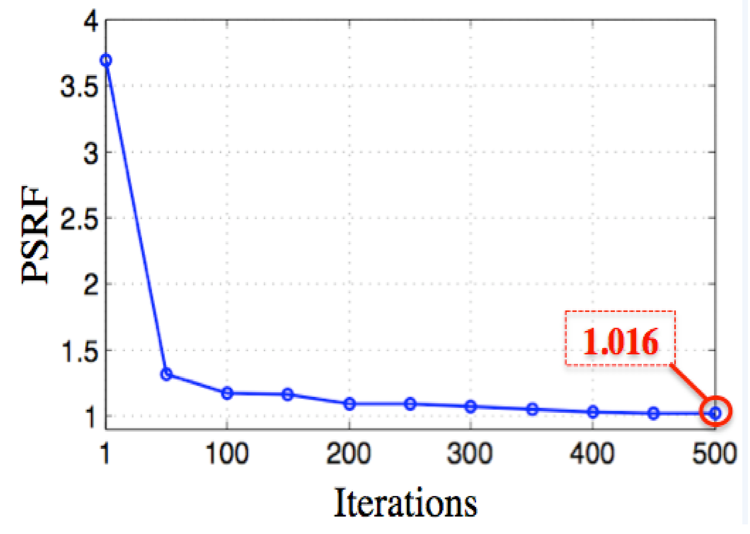
Graphical representation of the proposed joint representation model.



Larger values of Bernoulli parameters cluster well for different classes, indicating more frequent use of certain popular dictionary atoms for each class. Atom sharing is also evident from the plots.



Dictionary size reduces with iterations. Different parameter values result in different reduction rates, but convergence to similar dictionaries.



With appropriate initialization (see paper for the procedure) dictionary converges within 500 Gibbs sampling iterations.

Results:

Face Recognition on Extended YaleB

- Training:15 per subject.
- Random face features
- Average learned atoms 567

Method	Accuracy %	Average Time (ms)
DL-COPAR	86.47 ± 0.69	31.11
LC-KSVD1	87.76 ± 0.60	0.61
LC-KSVD	89.73 ± 0.59	0.60
D-KSVD	89.77 ± 0.57	0.61
SRC	89.71 ± 0.45	50.19
FDDL	90.01 ± 0.69	42.82
DBDL	91.09 ± 0.59	1.07
JBDC (Proposed)	**92.14 ± 0.52**	1.02

Action Recognition on Fifteen Scene Category

- Five-fold cross validation.
- Spatial Pyramid Features

Method	Accuracy %	Time
FDDL	94.08 ± 0.43	57.99
D-KSVD	95.12 ± 0.18	0.58
LC-KSVD1	95.37 ± 0.28	0.59
SRC	95.41 ± 0.13	78.33
DL-COPAR	96.02 ± 0.28	55.67
LC-KSVD	96.38 ± 0.29	0.59
DBDL	96.98 ± 0.28	0.71
JBDC (Proposed)	**97.73 ± 0.21**	0.70

Object Recognition on Caltech-101

- Deep features as input
- Classification time per sample is 6.2ms.
- 3001 atoms learned for training with 30 samples per class.

Training samples	5	10	15	20	25	30
SRC	76.23	79.99	81.27	83.48	84.00	84.51
DL-COPAR	76.11	80.40	83.44	84.01	84.85	85.03
FDDL	78.31	81.37	83.37	84.76	85.66	85.98
LC-KSVD1	79.03	82.86	84.13	84.65	86.10	86.94
D-KSVD	79.69	83.11	84.99	86.01	86.80	87.72
LC-KSVD	79.74	83.13	85.20	85.98	86.77	87.81
DBDL	80.11	84.03	85.99	86.71	87.97	88.81
JBDC (Proposed)	**81.64**	**85.70**	**86.96**	**87.88**	**88.72**	**89.59**

Demo/Code available at:
http://staffhome.ecm.uwa.edu.au/~00053650/code.html