Learning Inhomogeneous FRAME Models for Object Patterns

Jianwen Xie[1], Wenze Hu[2], Song-Chun Zhu[1], Ying Nian Wu[1]

[1] University of California, Los Angeles (UCLA), USA [2] Google Inc, USA

IEEE 2014 Conference on Computer Vision and Pattern Recognition

CVPR 2014

1 Abstract

We investigate an inhomogeneous version of the FRAME (Filters, Random field, And Maximum Entropy) model and apply it to modeling object patterns. The inhomogeneous FRAME is a non-stationary Markov random field model that reproduces the observed marginal distributions or statistics of filter responses at all the different locations, scales and orientations. Our experiments show that the inhomogeneous FRAME model is capable of generating a wide variety of object patterns in natural images. We then propose a sparsified version of the inhomogeneous FRAME model where the model reproduces observed statistical properties of filter responses at a small number of selected locations, scales and orientations. We propose to select these locations, scales and orientations by a shared sparse coding scheme, and we explore the connection between the sparse FRAME model and the linear additive sparse coding model. Our experiments show that it is possible to learn sparse FRAME models in unsupervised fashion and the learned models are useful for object classification.

5 Conclusion

The sparse inhomogeneous FRAME model has the following properties.

1. It can ***reconstruct*** the training images.
2. It can ***synthesize*** new images.
3. It separates ***appearance variations*** and ***shape deformations***.
4. It gives ***interpretable*** sketches.
5. Dictionaries or codebooks of models can be learned in ***unsupervised*** manner.
6. It combines rich traditions of *harmonic analysis* and *Markov random field* models.

6 Reproducibility

http://www.stat.ucla.edu/~jxie/iFRAME.html

2 Inhomogeneous FRAME Model

Model

It is a generative model that seeks to represent object patterns

with the form of

$$p(\mathbf{I};\lambda) = \frac{1}{Z(\lambda)} \exp\left(\sum_{x,s,\alpha} \lambda_{x,s,\alpha} \left|\langle \mathbf{I}, B_{x,s,\alpha}\rangle\right|\right) q(\mathbf{I})$$

where $B_{x,s,\alpha}$ is a basis function centered at pixel x, and tuned to scale s and orientation α, $Z(\lambda)$ is normalizing constant, $q(\mathbf{I})$ is a known reference density Gaussian white noise model.

MLE Learning:

(1) ***Parameters*** $\boldsymbol{\lambda}$: update equation by stochastic gradient algorithm:

$$\lambda_{x,s,\alpha}^{(t+1)} = \lambda_{x,s,\alpha}^{(t)} + \gamma_t \left(\frac{1}{M} \sum_{m=1}^{M} \left|\langle \mathbf{I}_m, B_{x,s,\alpha}\rangle\right| - \mathrm{E}_{p(\mathbf{I};\lambda^{(t)})}\left[\left|\langle \mathbf{I}, B_{x,s,\alpha}\rangle\right|\right] \right)$$

$$\mathrm{E}_{p(\mathbf{I};\lambda)}\left[\left|\langle \mathbf{I}, B_{x,s,\alpha}\rangle\right|\right] \approx \frac{1}{\tilde{M}} \sum_{m=1}^{\tilde{M}} \left|\langle \tilde{\mathbf{I}}_m, B_{x,s,\alpha}\rangle\right|$$

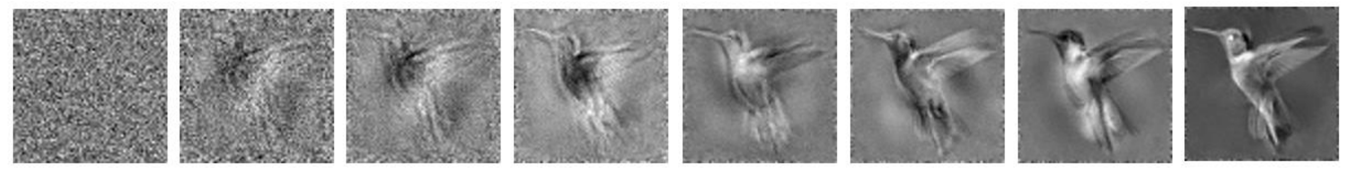
where $\{\mathbf{I}_m, m = 1,\dots,M\}$ are training images, $\{\tilde{\mathbf{I}}_m, m = 1,\dots,\tilde{M}\}$ are synthesized images sampled from $p(\mathbf{I};\lambda^{(t)})$ by Hamiltonian Monte Carlo (HMC), γ_t is the step size.

(2) ***Normalizing constant Z***: start from $\lambda^{(0)} = 0$, $\log Z(\lambda^{(0)}) = 0$. Compute $\log Z(\lambda^{(t)})$ along the learning process by iteratively updating its value as follows:

$$\log Z(\lambda^{(t+1)}) = \log Z(\lambda^{(t)}) + \log \frac{Z(\lambda^{(t+1)})}{Z(\lambda^{(t)})}$$

$$\frac{Z(\lambda^{(t+1)})}{Z(\lambda^{(t)})} \approx \frac{1}{\tilde{M}} \sum_{m=1}^{\tilde{M}} \left[\exp\left(\sum_{x,s,\alpha} \left(\lambda_{x,s,\alpha}^{(t+1)} - \lambda_{x,s,\alpha}^{(t)}\right) \times \left|\langle \tilde{\mathbf{I}}_m, B_{x,s,\alpha}\rangle\right| \right) \right]$$

Learning process (t = 1,7,10,20,50,100,300,and 500)



Images synthesis

3 Sparse FRAME Model

Model

Sparsified inhomogeneous FRAME model

$$p(\mathbf{I};\lambda) = \frac{1}{Z(\lambda)} \exp\left(\sum_{i=1}^{n} \lambda_i \left|\langle \mathbf{I}, B_{x_i,s_i,\alpha_i}\rangle\right|\right) q(\mathbf{I})$$

Deformable shared sparse coding

$$\mathbf{I}_m = \sum_{i=1}^{n} c_{m,i} B_{x_i+\Delta x_{m,i},\, s_i,\, \alpha_i+\Delta\alpha_{m,i}} + \epsilon_m$$

where $(\Delta x_{m,i}, \Delta\alpha_{m,i})$ are the perturbations (varying within limited ranges) of the location and orientation of the i-th basis function.

Learning

Step 1: Selecting $\mathbf{B} = (B_{x_i,s_i,\alpha_i},\ i = 1,\dots,n)$ by ***deformable shared matching pursuit*** algorithm to minimize

$$\sum_{m=1}^{M} \left\| \mathbf{I}_m - \sum_{i=1}^{n} c_{m,i} B_{x_i+\Delta x_{m,i},\, s_i,\, \alpha_i+\Delta\alpha_{m,i}} \right\|^2$$

[0] Initialize $\epsilon_m \leftarrow \mathbf{I}_m$, $i \leftarrow 0$.

[1] Let $i \leftarrow i+1$.

[2] Select $(x_i, s_i, \alpha_i) = \arg\max_{x,s,\alpha} \sum_{m=1}^{M} \max_{\Delta x,\Delta\alpha} \left|\langle \epsilon_m, B_{x+\Delta x,\, s,\, \alpha+\Delta\alpha}\rangle\right|^2$

[3] Let $(\Delta x_{m,i}, \Delta\alpha_{m,i}) = \arg\max_{\Delta x,\Delta\alpha} \left|\langle \epsilon_m, B_{x_i+\Delta x,\, s_i,\, \alpha_i+\Delta\alpha}\rangle\right|^2$.

[4] Let $c_{m,i} \leftarrow \langle \epsilon_m, B_{x_i+\Delta x_{m,i},\, s_i,\, \alpha_i+\Delta\alpha_{m,i}}\rangle$

[5] Update $\epsilon_m \leftarrow \epsilon_m - c_{m,i} B_{x_i+\Delta x_{m,i},\, s_i,\, \alpha_i+\Delta\alpha_{m,i}}$

[6] Stop if $i = n$, else go back to step [1].

Step 2: Estimating $\lambda = (\lambda_i, i = 1,\dots,n)$ given selected $\mathbf{B}$.

$$\lambda_i^{(t+1)} = \lambda_i^{(t)} + \gamma_t \left(\frac{1}{M} \sum_{m=1}^{M} \max_{\Delta x,\Delta\alpha} \left|\langle \mathbf{I}_m, B_{x_i+\Delta x,\, s_i,\, \alpha_i+\Delta\alpha}\rangle\right| - \frac{1}{\tilde{M}} \sum_{m=1}^{\tilde{M}} \left|\langle \tilde{\mathbf{I}}_m, B_i\rangle\right| \right)$$

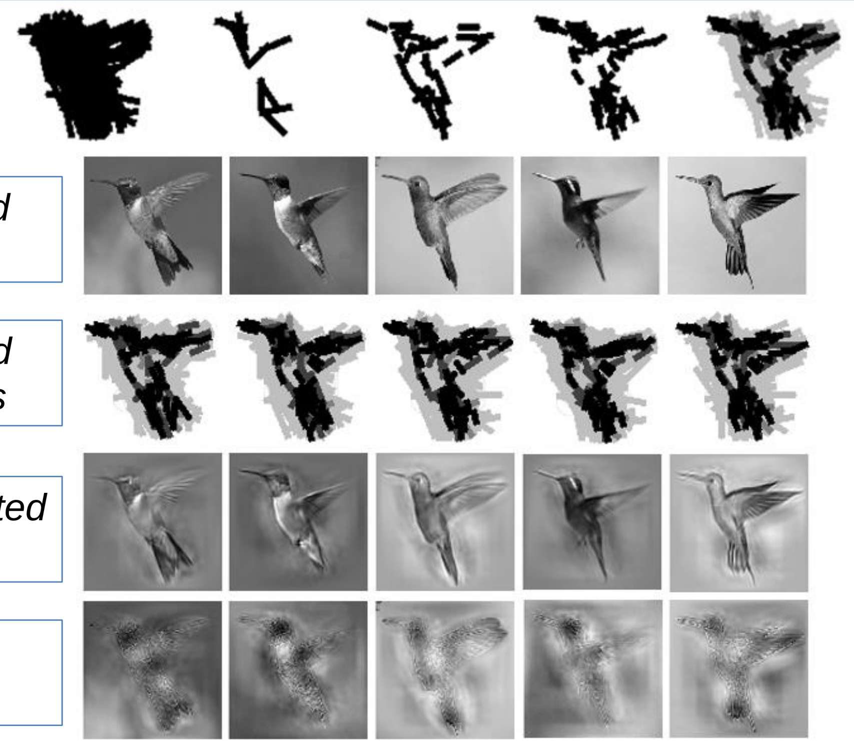
symbolic sketches of selected basis at 4 different scales and the superposition of the 4 scales

observed images

deformed sketches

reconstructed images

residual images

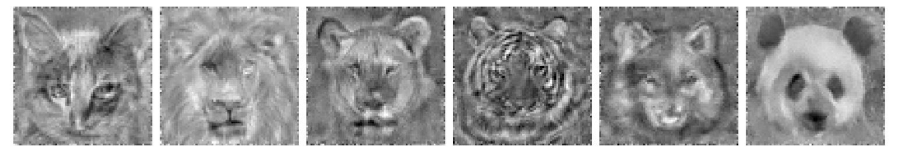


synthesized images

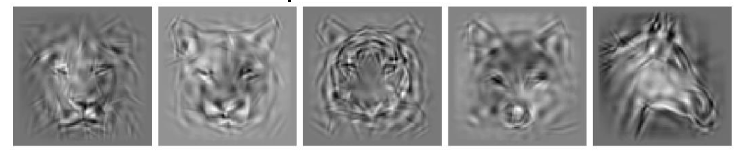
4 Experiments

Image synthesis

Dense FRAME



Sparse FRAME



Detection by deformable template matching

Template matching score

$$\sum_{i=1}^{n} \lambda_i \max_{\Delta x,\Delta\alpha} \left|\langle \mathbf{I}_m, B_{x_i+\Delta x, s_i, \alpha_i+\Delta\alpha}\rangle\right| - \log Z(\lambda)$$

Model-based EM-type Clustering

Let x be the true category of the image, y be the inferred category.

Conditional purity $\sum_y p(y) \max_x p(x|y)$

Conditional entropy $\sum_y p(y) \sum_x p(x|y) \log(\frac{1}{p(x|y)})$

Summary of comparison with k-mean on 7 clustering tasks

	purity	entropy
K-mean	0.820	0.347
FRAME	**0.912**	**0.158**

Exp1

Exp3

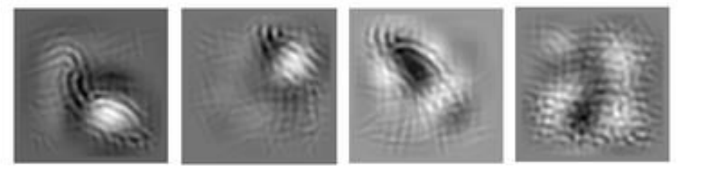
Codebook Learning

The learning algorithm iterates two steps:

Step 1: *Image encoding*: given the current codebook, encode the training images by spatially translated, rotated, scaled versions of the models in the codebook.

Step 2: *Codebook re-learning*: re-learn each model in the codebook from the image patches currently covered by this template

codebook



observed images

reconstructed images

Image classification on domain dataset

The learned codebook can serve as "words" in the "bag-of-word" method for object classification. We test it by image classification on domain adaptation tasks.

Method	C→A	C→D	A→C	A→W	W→C	W→A	D→A	D→W
Metric [1]	33.7±0.8	35.0±1.1	27.3±0.7	36.0±1.0	21.7±0.5	32.3±0.8	30.3±0.8	55.6±0.7
SGF [2]	40.2±0.7	36.6±0.8	37.7±0.5	37.9±0.7	29.2±0.7	38.2±0.6	39.2±0.7	69.5±0.9
GFK [3]	46.1±0.6	55.0±0.9	39.6±0.4	56.9±1.0	32.8±0.7	46.2±0.7	46.2±0.6	80.2±0.4
FDDL [4]	39.3±2.9	55.0±2.8	24.3±2.2	50.4±3.5	22.9±2.6	41.1±2.6	36.7±2.5	65.9±4.9
ours	62.2±1.6	52.2±4.0	46.7±2.5	53.2±4.9	39.1±3.0	53.2±4.4	55.3±2.9	72.4±3.1

C: caltech
A: amazon
W:webcam
D:DSLR

[1] K. Saenko, B. Kulis, M. Fritz, and T. Darrell. Adapting visual category models to new domains. ECCV, 213-226, 2010.
[2] R. Gopalan, R. Li, and R. Chellappa. Domain adaptation for object recognition: an unsupervised approach. ICCV, 999-1006, 2011.
[3] B. Gong, Y. Shi, F. Sha, and K. Grauman. Geodesic flow kernel for unsupervised domain adaptation. CVPR, 2066–2073, 2012.
[4] M. Yang, L. Zhang, X. Feng, and D. Zhang. Fisher discrimination dictionary learning for sparse representation. ICCV, 543-550, 2011.