Generative Hierarchical Learning of Sparse FRAME Models

Jianwen Xie[1], Yifei Xu[2], Erik Nijkamp[1], Ying Nian Wu[1], Song-Chun Zhu[1]

[1] University of California, Los Angeles (UCLA), USA [2] Shanghai Jiao Tong University, Shanghai, China

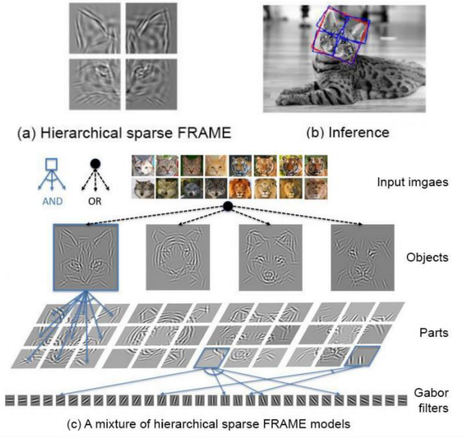
IEEE 2017 Conference on Computer Vision and Pattern Recognition

CVPR July 21-26 2017 HONOLULU

Abstract

This paper proposes a method for generative learning of hierarchical random field models. The resulting model, which we call the hierarchical sparse FRAME (Filters, Random field, And Maximum Entropy) model, is a generalization of the original sparse FRAME model by decomposing it into multiple parts that are allowed to shift their locations, scales and rotations, so that the resulting model becomes a hierarchical deformable template. The model can be trained by an EM-type algorithm that alternates the following two steps: (1) Inference: Given the current model, we match it to each training image by inferring the unknown locations, scales, and rotations of the object and its parts by recursive sum-max maps, and (2) Re-learning: Given the inferred geometric configurations of the objects and their parts, we re-learn the model parameters by maximum likelihood estimation via stochastic gradient algorithm. Experiments show that the proposed method is capable of learning meaningful and interpretable templates that can be used for object detection, classification and clustering.

Illustration



Reproducibility

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

Hierarchical Sparse FRAME Model

Representation

Let $B_{x,s,\alpha}$ denote a basis function (e.g., Gabor wavelets) centered at pixel x, tuned to scale s, and orientation α. Given a dictionary of basis functions $\{B_{x,s,\alpha}, \forall x, s, \alpha\}$. The model is a probability distribution defined on image $\mathbf{I}$,

$$p(\mathbf{I};\mathbf{H},\lambda) = \frac{1}{Z(\lambda)}\exp[f(\mathbf{I};\mathbf{H},\lambda)]q(\mathbf{I}),$$

where the scoring function $f(\mathbf{I};\mathbf{H},\lambda)$ is

$$f(\mathbf{I};\mathbf{H},\lambda) = \sum_{j=1}^{K}\sum_{i=1}^{n_j}\lambda_i^{(j)}|<\mathbf{I}, B_{x_i^{(j)},s_i^{(j)},\alpha_i^{(j)}}>|,$$

where $\mathbf{H} = \{(B_{x_i^{(j)},s_i^{(j)},\alpha_i^{(j)}}, i = 1,\dots,n_j), j = 1,\dots,K\}$ is a template of K groups of selected basis functions. Each group represents a part template. n_j is the number of basis functions in group j. $\lambda = \left\{\left(\lambda_i^{(j)}\right), i = 1,\dots,n_j\right), j = 1,\dots,K\}$ collects the parameters. $q(\mathbf{I})$ is a known Gaussian white noise reference distribution.

In current implementation, we simply divide the image domain into $K = d \times d$ non-overlapping parts, so that the basis functions within each part form a group. The parts and the basis functions are allowed to perturb their shift to account for shape deformation.

Unsupervised Learning

Objective function

The learning of the hierarchical sparse FRAME model is to learn K part templates $\left\{\mathbf{B}^{(j)}, j = 1,\dots,K\right\}$ from non-aligned training images $\{\mathbf{I}_m, m = 1,\dots,M\}$, while inferring the object locations χ_m, the part perturbations $(\Delta X_{m,j}, \Delta S_{m,j}, \Delta A_{m,j})$, and the perturbations of basis functions, by maximizing the objective function defined as the sum of the log-likelihood given $\mathbf{H}$ over all the training images, $\sum_{m=1}^{M} L(\mathbf{I}_m|\mathbf{H}_{\chi_m})$, subject to the constraint that there are no overlapping parts in each $\mathbf{I}_m$.

EM-type learning

E-step: Inference. Given the current $\mathbf{H}$, we match it to each image $\mathbf{I}_m$ by inferring the location of the object and the perturbations in locations, scales and rotations of K parts, as well as the perturbations of all basis functions in each part by recursive SUM-MAX maps:

Up-1: compute the Gabor wavelet matching score

$$\mathrm{SUM1}(x,s,\alpha) = |<\mathbf{I}, B_{x,s,\alpha}>|, \forall\, x, s, \alpha$$

Up-2: compute MAX1 by local maximization to account for shifts of Gabor wavelets

$$\mathrm{MAX1}(x,s,\alpha) = \max_{\Delta x,\Delta\alpha}\mathrm{SUM1}(x+\Delta x, s, \alpha+\Delta\alpha), \forall\, x, s, \alpha$$

Up-3: compute matching scores of K part templates

$$\mathrm{SUM2}^{(j)}(X) = \sum_{i=1}^{n_j}\lambda_i^{(j)}\mathrm{MAX1}(X + x_i^{(j)}, s_i^{(j)}, \alpha_i^{(j)}) - \log Z(\lambda^{(j)}), \forall\, X, j$$

Up-4: compute MAX2 by local maximization to account for shifts of part templates

$$\mathrm{MAX2}^{(j)}(X) = \max_{\Delta X}\mathrm{SUM2}^{(j)}(X+\Delta X),\ \forall X, j$$

Up-5: compute the matching score of the object template for all locations

$$\mathrm{SUM3}(\chi) = \sum_{j=1}^{K}\mathrm{MAX2}^{(j)}(\chi + X_j), \forall \chi$$

Up-6: compute the optimum matching score of H

$$\mathrm{MAX4} = \max_{\chi}\mathrm{SUM3}(\chi)$$

Down-1: compute the location of the object on the image

$$\tilde{\chi} = \arg\max_{\chi}\mathrm{SUM3}(\chi)$$

Down-2: compute the perturbations of all parts on the image

$$\Delta X_j = \arg\max_{\Delta X}\mathrm{SUM2}^{(j)}(\tilde{\chi} + X_j + \Delta X),\ \forall j$$

Down-3: compute the perturbations of Gabor wavelets in all parts on the image

$$\left(\Delta x_i^{(j)}, \Delta\alpha_i^{(j)}\right) = \arg\max_{\Delta x,\Delta\alpha}\mathrm{SUM1}\left(\tilde{\chi} + X_j + \Delta X_j + x_i^{(j)} + \Delta x, s_i^{(j)}, \alpha_i^{(j)} + \Delta\alpha\right), \forall\, i, j$$

M-step: Re-learning. Given the inferred deformation, we first align the objects and parts by morphing the corresponding images patches, and re-learn the model by the two-stage algorithm:

Stage-1: a shared sparse coding scheme is used to select $\mathbf{B} = \{B_{x_i,s_i,\alpha_i}, i = 1,\dots,n\}$ by minimizing the overall least squares reconstruction error

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

Stage-2: After selecting $\mathbf{B} = \{B_{x_i,s_i,\alpha_i}, i = 1,\dots,n\}$, we (1) estimate the corresponding weight parameters

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

where $\{\tilde{\mathbf{I}}_m, m = 1,\dots,\tilde{M}\}$ are synthesized images sampled from $p(\mathbf{I};\lambda^{(t)})$ by Hamiltonian Monte Carlo (HMC), γ_t is the step size; and (2) estimate the normalizing constant Z by starting from $\lambda^{(0)} = 0$, $\log Z\left(\lambda^{(0)}\right) = 0$ and computing $\log Z\left(\lambda^{(t)}\right)$ along the learning process by

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

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

Experiments

Evaluating mixture models by clustering tasks

A mixture of hierarchical sparse FRAME models can be trained in an unsupervised manner by an EM-type algorithm that iterates: (1) classifying images into different clusters based on the current model (2) re-learning the model of each cluster from images. Conditional purity and conditional entropy are used to measure the clustering performance. Table below summarizes the comparisons with other methods by showing the average accuracies based on 5 repetitions for 12 clustering tasks.

Measure	ours	Sparse FRAME	Generative Boosting [1]	Active basis [2]	Two-step EM [3]	K-means + HoG	AOT [4]
purity	**0.962**	0.928	0.923	0.815	0.798	0.788	0.849
entropy	**0.083**	0.174	0.159	0.365	0.419	0.408	0.291

Object, part, and key point localization

We evaluate the accuracy of the inference of our model on detection tasks. The performance of detection is measured by evaluating the accuracy of localizing key points, parts, and objects. We plot imprecision-recall curves (e.g., 2 examples shown in Fig. 1) and use area under curve (AUC) to measure the performance of the localization of key points. Table below displays the average AUCs over 8 categories. Fig. 2 display some detection results.

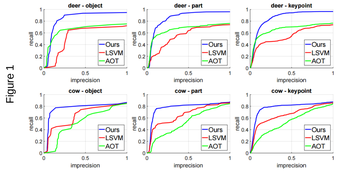


Figure 1

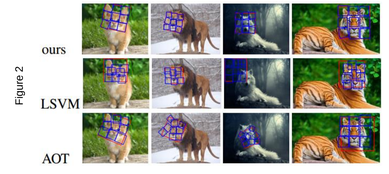


Figure 2

tasks	Object			part			Key point		
	ours	AOT [4]	LSVM [5]	ours	AOT	LSVM	ours	AOT	LSVM
Avg.	0.859	0.785	0.727	0.868	0.793	0.732	0.867	0.795	0.730

Evaluating unsupervisedly learned models via classification

We use the LHI-Animal-Faces dataset. For each category, we learn a mixture model of 5 or 11 hierarchical sparse FRAME models with 2×2 moving parts in an unsupervised manner. We then combine the object templates from all the learned mixture models into a codebook of $20 \times 5 = 100$ or $20 \times 11 = 220$ codewords. The maps of template matching scores from all the codewords in the codebook are computed for each image and then they are fed into spatial pyramid matching (SPM) to obtain feature vectors. SVM classifiers with L2 loss are trained on these vectors, and are evaluated on the testing data in terms of classification accuracies.

# clusters	AOT [4]	Ours w/o parts	Ours
5	65.80%	70.62%	**74.33%**
11	62.54%	72.56%	**75.83%**

[1] Xie, Jianwen, et al. "Inducing wavelets into random fields via generative boosting." ACHA, 2016
[2] Wu, Ying Nian, et al. "Learning active basis model for object detection and recognition." IJCV, 2010.
[3] Barbu, Adrian, et al. "Learning mixtures of bernoulli templates by two-round EM with performance guarantee." Electronic Journal of Statistics, 2014.
[4] Si, Zhangzhang, et al. "Learning and-or templates for object recognition and detection." PAMI, 2013.
[5] Felzenszwalb, Pedro F., et al. "Object detection with discriminatively trained part-based models." PAMI, 2010.

Conclusion

This paper proposes a generative learning framework applied to hierarchical representations of object patterns. Our model is defined as a hierarchical extension of the original sparse FRAME model. The model is capable of capturing geometric deformations and can be learned in an unsupervised manner. It can be visualized by MCMC sampling. Compared to previous generative hierarchical leaning methods, our method performs better in terms of accuracies of localization of object, parts, and key points in detection, object classification, and clustering.