Hierarchical Joint Max-Margin Learning of Mid- and Top-Level Representations for Visual Recognition

Hans Lobel
Machine Learning Group
Pontificia Universidad Católica de Chile

René Vidal
Center for Imaging Science
The Johns Hopkins University

Álvaro Soto
Machine Learning Group
Pontificia Universidad Católica de Chile

Our Contribution

- Common recognition methods use a 2 level hierarchy:
  - A visual dictionary learned in an unsupervised fashion.
  - A top-level classifier, generally using a one-vs-all scheme.
- We present a method to learn this hierarchy, where both levels are jointly optimized to find a task-specific representation.
- Experimental evaluation shows an important performance increase, leading to state-of-the-art performance using smaller dictionaries.

Image Encoding

- We model each image with:
  - A set of L rectangular regions defined over an image I.
  - A set of local features, $\{v_j^I\}$, extracted from an image I.
  - A visual dictionary $\Theta = [\theta_1\ \theta_2\ \theta_3\ \dots\ \theta_K]$, formed by linear classifiers.
- We encode a rectangular region l defined over I as follows:

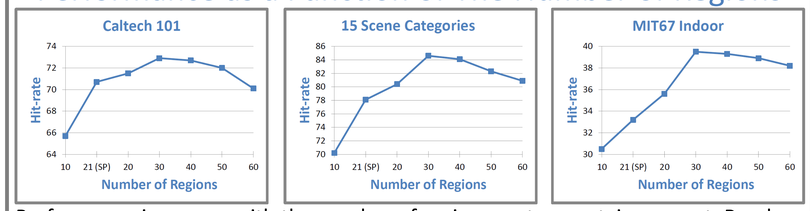
$$x_{l,\Theta,z}(I) = [\langle \theta_1, v^I_{z_{(l,1)}} \rangle, \langle \theta_2, v^I_{z_{(l,2)}} \rangle, \ \dots, \langle \theta_K, v^I_{z_{(l,K)}} \rangle]$$

where variables $z^I_{(l,k)} = \underset{j \in N_l^I}{\operatorname{argmax}} \langle \theta_k, v_j^I \rangle$ induce max-pooling.

Results For Scene/Object Categorization

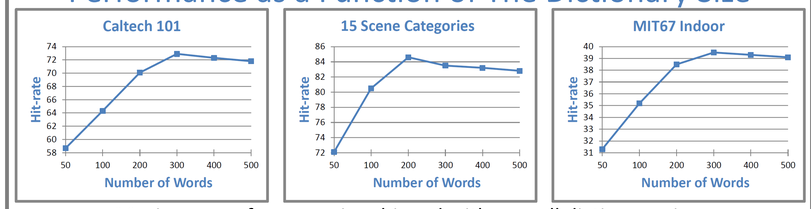
- Datasets: Caltech 101, 15 Scene Categories, MIT67 Indoor.
- Images downsampled to no more than 300 pixels.
- HOG+LBP descriptors extracted densely.
- L rectangular regions selected randomly.

Performance as a Function of The Number of Regions



Performance increases with the number of regions up to a certain amount. Random regions give better performance than Spatial Pyramid at similar number of regions.

Performance as a Function of The Dictionary Size



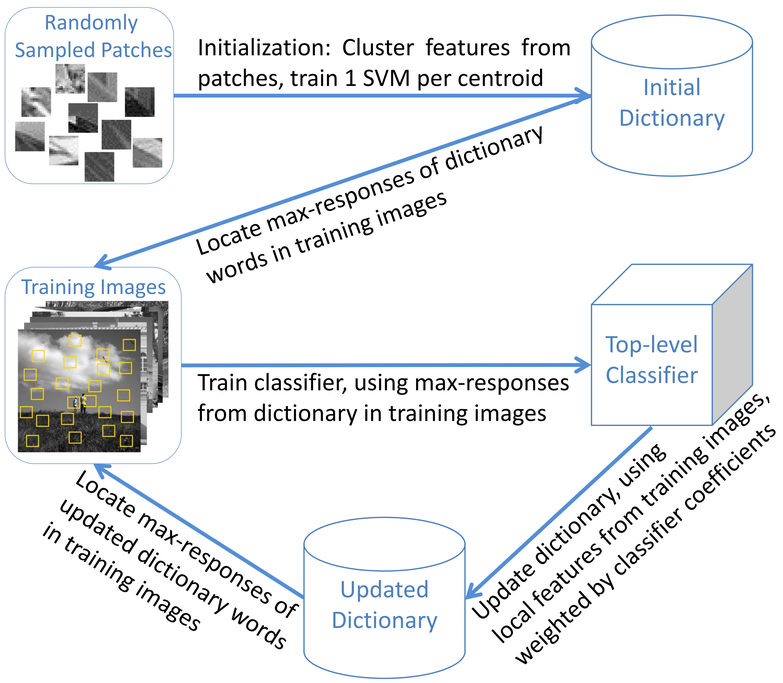
Maximum performance is achieved with a small dictionary size.

Comparison to Other Methods

Method	# Words	Dataset: Caltech101	Dataset: 15 Scenes	Dataset: MIT67
Baseline	400	63.9 ± 0.6	78.1 ± 0.3	33.2
SPM	400	64.6 ± 0.8	81.4 ± 0.5	-
LLC	2048	**73.4**	80.5 ± 0.6	-
LCSR	1024	73.2 ± 0.8	82.7 ± 0.5	-
ScSPM	1024	73.2 ± 0.5	80.3	-
Max-margin	5250	-	82.7 ± 0.5	-
Object Bank	200 per class	-	80.9	37.6
Reconfigurable Models	200	-	78.6 ± 0.7	37.9
Discriminative Patches	210 per class	-	-	38.1
Proposed	200 or 300	72.9 ± 0.6	**84.6 ± 0.4**	**39.5**

Compared to other related methods, we obtain state-of-the-art performance on two out of three datasets, using a much smaller dictionary size.

A Quick Overview



Learning Problem

- Given a set of training examples $\{I^i, y^i\}_{i=1}^N$, we find object classifier parameters W and dictionary parameters Θ by solving a regularized max-margin learning problem:

$$\min_{W,\Theta,\{\xi_i\}} \ \frac{1}{2}\|W\|_F^2 + \frac{C_1}{2K}\|\Theta\|_F^2 + \frac{C_2}{N}\sum_{i=1}^N \xi_i$$

$$\text{s.t.} \quad \sum_l^L \sum_k^K w_{y^i,l,k} \langle \theta_k, v^{I^i}_{z^{I^i}_{(l,k)}} \rangle - \sum_l^L \sum_k^K w_{y,l,k} \langle \theta_k, v^{I_i}_{z_{(l,k)}} \rangle \geq \Delta(y_i, y) - \xi_i,$$
$$\forall i \in \{1,\dots,N\} \wedge \forall y \in \{1,\dots,M\} \wedge \forall z.$$
$$w_{y,l,k} \geq 0, \forall y \in \{1,\dots,M\} \wedge \forall l \in \{1,\dots,L\} \wedge \forall k \in \{1,\dots,K\}.$$

- We use an alternating minimization approach:
  - When Θ is fixed, we solve a standard multi-class SVM for W.
  - When W is fixed, we solve a Latent Structural SVM for Θ.