Joint Dictionary and Classifier Learning for Categorization of Images using a Max-margin Framework

Hans Lobel*, René Vidal†, Domingo Mery*, Alvaro Soto*

* grima Machine Intelligence Group
Pontificia Universidad Católica, Chile

† Center for Imaging Science
Johns Hopkins University, Baltimore, Maryland

Bag of Visual Words is one of the most popular recognition techniques

- Mid-level dictionary and top-level classifier.
- The dictionary is commonly built with a generative process (k-means, sparse coding).
- Generally uses KNN or a set of one-vs-all classifiers for multiclass categorization.
- Learning processes are separate, i.e., no coupling between dictionary and classifiers.

Jointly learning dictionary and classifiers is a key aspect

- Can increase dictionary discriminativity, thus facilitating classifiers work.
- Few works explore this in a recognition context:
  - ✓ Categorization -> Lian et al. (2010)
  - ✓ Segmentation -> Jain et al. (2012)
  - ✓ Saliency -> Yang & Yang (2012)
- Results show a clear performance improvement.

True multiclass classification can also improve the dictionary

- Takes advantage of meaningful correlations between categories.
- Can induce word sharing behavior, which is a very desirable property (Ott & Everingham, 2011).
- Almost no work explores this using BoVW.

A joint max-margin framework for recognition

- Discriminative dictionary composed of multiple linear SVMs.
- Multiclass SVM for categorization, using scores of dictionary words activations.
- Max-pooling over a spatial pyramid.
- Jointly learnt using a regularized max-margin energy minimization problem.

Outline

- Image model: Encoding and categorization
- Learning problem
- Experiments and results
- Conclusions

Visual descriptors are encoded according to the dictionary

- Assume a dictionary $\Theta = [\theta_1 \; \theta_2 \; \theta_3 \; \ldots \; \theta_K]$, composed of linear SVMs, and a set of visual descriptors extracted from squared image sectors.

- A visual descriptor v is encoded as follows:

$$c_{\Theta}(v) = [v^T \theta_1, \; \ldots, v^T \theta_K] = v^T \Theta$$

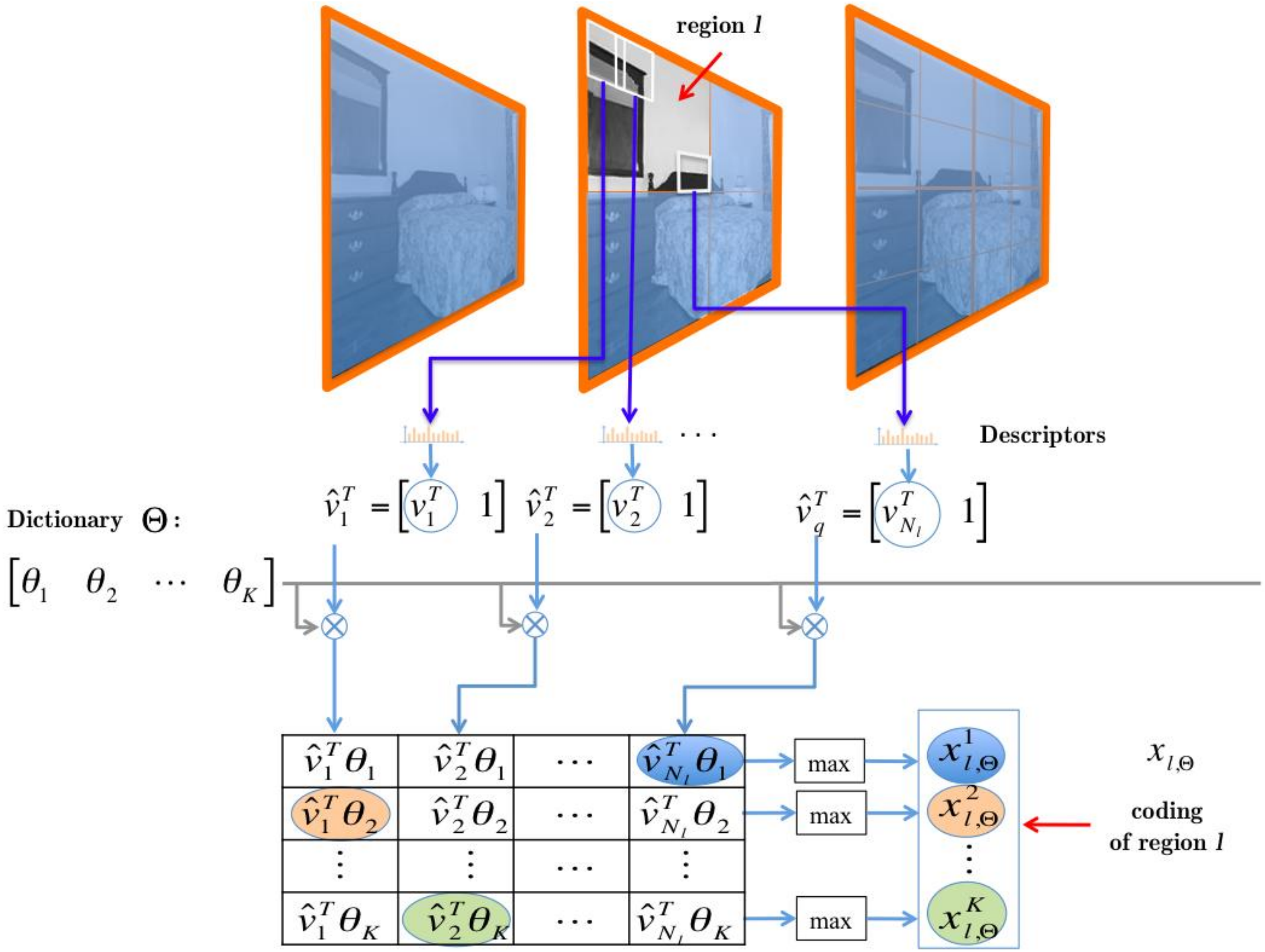
Images are encoded based on a spatial pyramid and max-pooling

- Given a spatial pyramid formed by L regions, a region l is encoded the following way:

$$x_{l,\Theta} = [\max_{j=1}^{N_l} v_{l,j}^T \theta_1,\ \max_{j=1}^{N_l} v_{l,j}^T \theta_2,\ \ldots,\ \max_{j=1}^{N_l} v_{l,j}^T \theta_K]^T$$

- Finally, the encoding of an image, $x_\Theta(I)$, is obtained by concatenating the encoding of each region.

Images are encoded based on a spatial pyramid and max-pooling



Energy is given by a linear combination of max functions

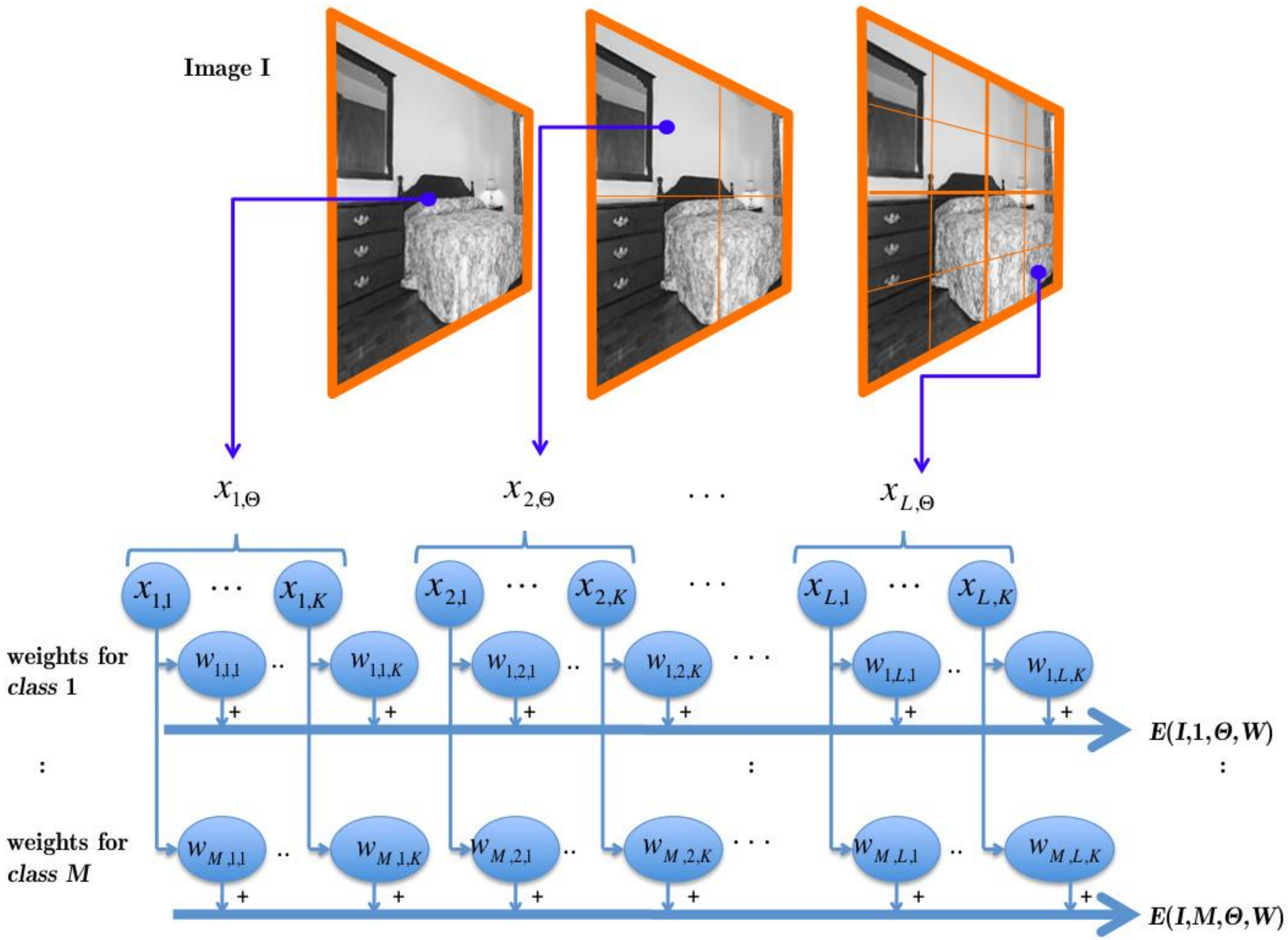
- Given a set of linear classifiers $W = [w_1 \; w_2 \; \cdots \; w_M]$ we define the energy of an image with L regions as follows:

$$E(I, y, \Theta, W) = \quad w_y^T x_\Theta(I) = \quad \sum_l^L \sum_k^K w_{y,l,k} \cdot \max_{j=1}^{N_l}(v_{l,j}^T \theta_k)$$

- An image is finally categorized the following way:

$$y^* = \underset{y}{\operatorname{argmax}} \, E(I, y, \Theta, W)$$

Energy is given by a linear combination of max functions



A regularized max-margin energy minimization learning problem

- Given a set of training examples $\{I_i, y_i\}_{i=1}^{N}$, we find W and Θ by solving the following max-margin problem:

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

$$\text{s.t.} \quad E(I_i, y_i, \Theta, W) - E(I_i, y, \Theta, W) \geq \Delta(y_i, y) - \xi_i,$$

$$\forall i \in \{1, \ldots, N\} \wedge \forall y \in \{1, \ldots, M\}.$$

Non-convexity does not allow using the same approach

- Our problem is similar to a Structural SVM, but differs in two fundamental points:
  i. Constraints not linear on Θ
  ii. Optimization is not jointly convex on W and Θ

- We solve the problem using an alternating minimization approach

Classifier learning is straight forward

- By fixing Θ, the problem reduces to a standard multiclass SVM. Can be efficiently solved using a cutting plane algorithm.

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

$$\text{s.t.} \quad E(I_i, y_i, \Theta, W) - E(I_i, y, \Theta, W) \geq \Delta(y_i, y) - \xi_i, \\ \forall i \in \{1, \dots, N\} \wedge \forall y \in \{1, \dots, M\}.$$

Dictionary learning requires a different approach

- By fixing W, we are required to solve the following problem:

$$\min_{\Theta} \frac{C_1}{2K} \|\Theta\|_F^2 + \frac{C_2}{N} \sum_{i}^{N} E(I_i, \hat{y}_i, \Theta, W) + \Delta(y_i, \hat{y}_i) - E(I_i, y_i, \Theta, W)$$

where

$$\hat{y}_i = \operatorname*{argmax}_{y} E(I_i, y, \Theta, W) + \Delta(y_i, y)$$

Dictionary learning requires a different approach

- To solve the last problem, we use an interior point optimization method, which requires the problem to be differentiable.
- To achieve that, we approximate the max function with a convex soft-max version, given by the log-sum-exponential function:

$$\max_{i=1}^{N}(z_i) \approx \frac{1}{r}\log\left(\sum_{i}^{N}\exp(r z_i)\right)$$

Some implementation details before the results

- Evaluation performed on three datasets: 15 scene categories, MIT67 and Caltech101.
- HOG+LBP descriptors are extracted on dense grid of regions of 16x16 pixels, with a spacing of 8 pixels in each direction.
- The initial dictionary is obtained by clustering a subset of descriptors and the training a linear SVM for each centroid.

Performance benefits from more words, up to certain point

	Number of Words		
Dataset	**50**	**100**	**200**
Caltech101	63.1 ±0.8	72 ± 0.5	73.1 ± 0.5
15 Scenes	72.2 ± 0.5	83.7 ± 0.2	84.8 ± 0.2
MIT67	31.2	38.3	39.9

State-of-the-art result with far less words than other methods

		Datasets		
Method	**# Words**	**Caltech101**	**15 Scenes**	**MIT67**
Baseline	200	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	-	82.17± 0.5	-
Object Bank	200	-	80.9	37.6
Reconfigurable Models	200	-	78.6 ± 0.7	37.9
Discriminative Patches	210*	-	-	38.1
Proposed	200	73.1 ± 0.5	**84.8 ± 0.2**	**39.9**

Jointly learning dictionary and classifiers actually works

- Performance is notably increased when compared to the baseline method.
- The proposed scheme produces a strong sharing of visual words among the target classes.
- This sharing allows us to use smaller dictionaries and achieve state-of-the-art performance.

Joint Dictionary and Classifier Learning for Categorization of Images using a Max-margin Framework

Hans Lobel*, René Vidal†, Domingo Mery*, Alvaro Soto*

* Machine Intelligence Group
Pontificia Universidad Católica, Chile

† Center for Imaging Science
Johns Hopkins University, USA