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

Hans Lobel[†] René Vidal[‡] Alvaro Soto[†]
[†]Department of Computer Science, Pontificia Universidad Católica de Chile
[‡]Center for Imaging Science, Johns Hopkins University

Abstract

Currently, Bag-of-Visual-Words (BoVW) and part-based methods are the most popular approaches for visual recognition. In both cases, a mid-level representation is built on top of low-level image descriptors and top-level classifiers use this mid-level representation to achieve visual recognition. While in current part-based approaches, mid- and top-level representations are usually jointly trained, this is not the usual case for BoVW schemes. A main reason for this is the complex data association problem related to the usual large dictionary size needed by BoVW approaches. As a further observation, typical solutions based on BoVW and part-based representations are usually limited to extensions of binary classification schemes, a strategy that ignores relevant correlations among classes. In this work we propose a novel hierarchical approach to visual recognition based on a BoVW scheme that jointly learns suitable mid- and top-level representations. Furthermore, using a max-margin learning framework, the proposed approach directly handles the multiclass case at both levels of abstraction. We test our proposed method using several popular benchmark datasets. As our main result, we demonstrate that, by coupling learning of mid- and top-level representations, the proposed approach fosters sharing of discriminative visual words among target classes, being able to achieve state-of-the-art recognition performance using far less visual words than previous approaches.

1. Introduction

The success of recognition methods based on visual descriptors and off-the-shelf machine learning techniques [27, 6] is one of the main reasons for the new enthusiasm in computer vision technologies. These methods have shown robustness against visual complexities, such as changes in illumination, scale, affine distortions, and mild intraclass and pose variations. Unfortunately, the visual world is highly complex and problems such as object deformations, partial occlusions, and severe intra-class and pose variations, require more elaborate solutions [8].

In terms of the main goals of visual recognition, such as object, scene, or action recognition, currently the two most popular approaches are: Bag-of-Visual-Words (BoVW) [24] and part-based methods [7]. In both cases, a mid-level representation is built on top of low-level image descriptors, such as SIFT or HoG features.

BoVW models build a mid-level representation that corresponds to the output of a pooling scheme acting on a visual dictionary that encodes appearance information from local image patches. Early BoVW approaches were based on vector quantization, generally using K-Means to cluster low-level keypoint descriptors [24, 5]. Afterwards, several variations have been proposed using alternative quantization methods, discriminative dictionaries, or different pooling strategies [12, 14, 30, 20]. Additionally, spatial information has also been incorporated by concatenating BoVW representations from different local image areas and different scales [15]. Recently, sparse coding schemes have emerged as a powerful alternative to vector quantization providing dictionaries that achieve lower reconstruction errors and attractive computational properties. In particular, [31] shows that a combination of sparse coding, spatial pyramidal decomposition, max-polling, and high dimensional linear SVM classifiers provide a powerful scheme to perform object and scene recognition. Discriminative sparse representations have also been proposed [19, 4], mostly building particular dictionaries for each class.

In the case of part-based models, the mid-level representation corresponds to basic semantic visual structures that can usually be mapped to relevant object components. Common strategies to obtain these parts are manual selection [7], greedy latent models [8], or the output of a large set of part-based classifiers trained using a costly labeling process [3]. Spatial information is also incorporated into the models by learning common spatial configurations among parts [8]. After the seminal work in [8], latent models have

Acknowledgment: This work was partially funded by FONDECYT grant 1120720 and NSF grant 11-1218709.

been used to jointly learn parts and object classifiers under a common optimization scheme that maximizes object classification performance. Recently, [29] proposes an extension to [8] that directly considers the multiclass classification case, but in the context of an action recognition application.

In both cases, BoVW and part-based models, the final recognition is generally based on a classifier that is trained on top of the mid-level representation usually augmented with spatial information. As a common fact, most of these models do not consider the multiclass classification problem directly, [29] being a notable exception. As an alternative, most methods achieve multiclass classification by using variations of the binary classification case, for example, using a one-versus-all classification strategy. Unfortunately, these types of strategies do not consider relevant correlations among classes. More importantly, they usually rely on solutions that employ a particular mid-level representation for each target class, a strategy that does not scale properly with the number of classes.

In terms of hierarchical compositional models, our work is related to recent recognition approaches based on deep belief networks (DBN) [9, 13], where the training process also incorporates hierarchical estimation of latent variables, spatial pooling schemes, and intermediate representations based on linear filters. DBNs are usually applied over a raw image representation using several layers of generic structures. As a consequence, DBNs have many parameters and they are usually difficult to train. In contrast, we embed semantic knowledge to our model by explicitly exploiting compositional relations among low level visual features, visual words, and high label classifiers. This leads to simpler architectures and allows us to potentially incorporate labeled data at intermediate layers. Furthermore, our max-margin approach is based on a Hinge loss, and not a quadratic or a logistic function commonly used to train DBNs, leading to a different optimization setup.

We believe that it is still not clear what is the most suitable level of abstraction to implement intermediate level representations. While part-based approaches have demonstrated excellent performance by using a small set of highly discriminative parts augmented with relevant spatial constraints, they also present problems due to common visual complexities, such as high intraclass part appearance variations, as well as, data association problems related to missing parts. In these cases, the redundancy and greater flexibility offered by the statistical properties of BoWs, based on a suitable dictionary of visual words, represent an attractive alternative or complement to part-based approaches.

In this work we present a novel hierarchical approach to visual recognition that jointly learns a suitable mid-level representation together with top-level classifiers using a multiclass max-margin approach. We formulate our problem as an energy minimization problem, where structural hierarchical relations are modeled by sub-energy terms acting at different levels of abstraction. In terms of dictionary construction, we depart from the usual vector quantization [24] or sparse coding schemes [31] commonly used in BoVW models, and similarly to [8], we use linear SVMs classifiers to characterize each word of the dictionary underlying the BoVW representation. Furthermore, we complement the dictionary construction with a max-pooling strategy, because it shows superior performance than alternative techniques, as suggested in [31] and [4]. More importantly, by coupling learning of a mid-level dictionary and top-level classifiers, we are able to obtain a mid-level representation that fosters word sharing among target classes. As shown by our experiments, this allows us to achieve state-of-the-art recognition performance in several common benchmarks datasets, using an order of magnitude less words than previous approaches. We believe that this is a critical issue to the scalability of visual recognition algorithms.

From a machine learning perspective our hierarchical formulation allows us to combine labeled and unlabeled data under a common framework. In particular, high level semantic information at the level of object or scene class labels can be propagated to guide the otherwise unsupervised search for relevant visual words at the level of image patches. We believe that this is a powerful learning scheme that can be extended to further levels of abstraction providing a rich hypothesis space to build visual compositional schemes [2].

Consequently, this work makes two main contributions:

- Introduce a hierarchical method that jointly learns suitable BoVW mid-level representations and top-level classifiers using a multiclass max-margin framework.

- Demonstrate that by coupling learning of a mid-level dictionary and top level classifiers, we are able to achieve state-of-the-art results with a significant reduction in dictionary size respect to previous approaches.

2. Model Description

One of the motivations for our approach is the work of Yang *et al*. [31]. In this work, the authors achieve excellent results in image classification by combining a linear spatial pyramid matching (SPM) kernel based on sparse coding (ScSPM) with linear SVMs. Their key insights are to use sparse dictionary learning techniques to construct the dictionary (as opposed to k-means), and to use *max pooling* to construct the descriptors over a spatial pyramid (as opposed to average pooling, which leads to histograms).

An important disadvantage of [31] is that the dictionary of visual features is learned independently from the classification parameters for the different object categories. In this

work, we propose an approach in which the visual dictionary is learned jointly with the top level visual classifiers.

2.1. Image Representation

As in [31], we assume that visual descriptors [18, 1] are extracted from images, either centered at interest points or by using a dense sampling scheme, and that each of these descriptors has size T. Inspired by [10], we define a visual dictionary Θ of K words,

$$\Theta = [\theta_1\ \theta_2\ \theta_3\ \ldots\ \theta_K] \in \mathbb{R}^{(T+1)\times K}, \tag{1}$$

where each word θ_k is represented as a linear classifier with bias:

$$\theta_k = [\theta_{k,1}, \theta_{k,2}, \ldots, \theta_{k,T}, b_k]^T \in \mathbb{R}^{T+1}. \tag{2}$$

Unlike [31], instead of using *sparse coding* to encode each descriptor, we use an encoding scheme based on the classification score obtained by each dictionary word. More specifically, if v is a descriptor vector, its coding $c_\Theta(v)$ using dictionary Θ is given by:

$$c_\Theta(v) = [\langle v, \theta_1\rangle,\ \ldots,\ \langle v, \theta_K\rangle] = v^T\Theta. \tag{3}$$

Intuitively, if the visual words are sufficiently discriminative, the descriptor v should be similar only to a few words in the dictionary. Therefore, we expect the vector $c_\Theta(v)$ to have only few values that are greater than zero.

Given a dictionary Θ and a set of L rectangular regions defined over an image, we represent the image using *max spatial pooling*. Generally, these regions are defined using a spatial pyramidal decomposition. Instead, we use regions randomly defined over an image. For each region $l = 1, \ldots, L$, let v_j^l be a descriptor vector extracted from region l, where $j \in [1 \ldots N_l]$ indexes the N_l descriptors extracted from region l. Thus, given a dictionary Θ, we encode region l using *max spatial pooling* as:

$$x_{l,\Theta} = [\max_{j\in N_l}\langle v_j^l, \theta_1\rangle,\ \ldots,\ \max_{j\in N_l}\langle v_j^l, \theta_K\rangle]^T \in \mathbb{R}^K. \tag{4}$$

Notice that, unlike the sparse coding scheme, which assigns zero weights to dictionary words that do not contribute to image reconstruction, our scheme assigns negative weights to dictionary words with low similarity. This property can potentially lead to over-fitting. To avoid this, we assume that each region contains a null feature vector $\vec{0}$, whose classification score is zero for any of the dictionary words. In this way, regions where none of the extracted feature vectors obtains a positive score obtain a zero weight by the max-pooling operator.

Finally, the complete descriptor of image I given dictionary Θ, $x_\Theta(I)$, is obtained by concatenating the descriptors of its L regions, *i.e.*,

$$x_\Theta(I) = [x_{1,\Theta}, x_{2,\Theta}, \ldots, x_{L,\Theta}]^T \in \mathbb{R}^{KL}. \tag{5}$$

2.2. Image Classification

Given a descriptor for image I, $x_\Theta(I)$, we define an image classification score, or energy function, for an image I as:

$$E(I, y, \Theta, W) = \quad w_y^T x_\Theta(I). \tag{6}$$

Here, $w_y \in \mathbb{R}^{KL}$ represents the parameters of a classifier learnt for object class $y \in \{1, 2, \ldots, M\}$ and

$$W = [w_1\ w_2\ \cdots\ w_M] \in \mathbb{R}^{KL\times M} \tag{7}$$

represents all the object classifier parameters.

If w_y is divided into L sub-vectors of size K, each one assigned to a different region, we can rewrite the energy in the following form:

$$E(I, y, \Theta, W) = \quad \sum_l^L \sum_k^K w_{y,l,k} \cdot \max_{j\in N_l}\langle v_j^l, \theta_k\rangle, \tag{8}$$

where $w_{y,l,k}$ refers to the k-th element of the l-th sub-vector of w_y. This formulation makes explicit the fact that the total energy of an image is a linear combination of max functions. It can also be seen that the energy function shows a nonlinear dependence between the weights w_y and the dictionary words θ_k. Given the parameters of the classifiers for the different object categories, W, and the parameters of the classifiers for the different visual words, Θ, we classify an image I as follows

$$y^* = \operatorname*{argmax}_y E(I, y, \Theta, W). \tag{9}$$

3. Learning

The model described in the previous section depends on two sets of parameters: the object classifiers W and the visual words classifiers Θ. Rather than first learning the visual words and then learning the object classifiers, our goal is to learn both of them simultaneously, so that the visual words are discriminative for the visual classification task.

More specifically, given a set of training examples $\{I_i, y_i\}_{i=1}^N$, where I_i is the i-th image and y_i is its class, we propose to find Θ and W by solving the following regularized max-margin learning problem:

$$\min_{W,\Theta,\{\xi_i\}} \frac{1}{2}\|W\|_F^2 + \frac{C_1}{2K}\|\Theta\|_F^2 + \frac{C_2}{N}\sum_{i=1}^N \xi_i \tag{10}$$

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

The objective function encourages the construction of visual words that behave like linear SVMs, *i.e.*, classifiers

that jointly maximize the margin and minimize the loss. On the other hand, the constraints encourage the score for an image according to its ground truth label, $E(I_i, y_i, \Theta, W)$, to be higher than the score according to any other label, $E(I_i, y, \Theta, W)$, by a loss function $\Delta(y_i, y)$ given by

$$\Delta(y_1, y_2) = \begin{cases} 0 & \text{if } y_1 = y_2 \\ 1 & \text{otherwise} \end{cases}. \tag{11}$$

The slack variables $\xi_i \geq 0$ allow for a violation of these constraints.

Although similar, this problem cannot be solved as a particular case of Structural SVM (S-SVM) [26]. As the constraints are non-linear on the parameters (W, Θ). Even when fixing the weights W and solving only for Θ, the constraints are not linear on the parameters.

To tackle this issue, we can solve a relaxed version of the previous problem using latent variables to avoid the non-linearity. If we recall, the descriptor of a region l is given by Eq. (4). We modify this expression by removing the *max* operator and adding a set of latent variables $z = \{z_{(l,k)}\}$, for $l \in \{1, \ldots, L\} \wedge k \in \{1, \ldots, K\}$:

$$x_{l,\Theta,z} = [\langle \theta_1, v^l_{z_{(l,1)}} \rangle,\ \langle \theta_2, v^l_{z_{(l,2)}} \rangle,\ \ldots,\ \langle \theta_K, v^l_{z_{(l,K)}} \rangle]^T, \tag{12}$$

where $z_{(l,k)}$ is the index of the descriptor extracted from region l with maximum response, when θ_k is applied to it, i.e:

$$z_{(l,k)} = \underset{j \in N_l}{\text{argmax}} \langle \theta_k, v^l_j \rangle. \tag{13}$$

Using this, Eq. (8) modifies to:

$$E(I, y, \Theta, W) = \sum_l^L \sum_k^K w_{y,l,k} \cdot \langle \theta_k, v^l_{z_{(l,k)}} \rangle. \tag{14}$$

Based on this energy formulation, we can now state the problem in a form similar to Eq. (10):

$$\min_{W, \Theta, \{\xi_i\}} \frac{1}{2}\|W\|_F^2 + \frac{C_1}{2K}\|\Theta\|_F^2 + \frac{C_2}{N}\sum_{i=1}^N \xi_i \tag{15}$$

$$\text{s.t.} \quad \sum_l^L \sum_k^K w_{y_i,l,k} \cdot \langle \theta_k, v^l_{z_{i,(l,k)}} \rangle - \sum_l^L \sum_k^K w_{y,l,k} \cdot \langle \theta_k, v^l_{z_{(l,k)}} \rangle \geq \Delta(y_i, y) - \xi_i,$$

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

This new problem is similar to a Latent Structural SVM (LS-SVM)[32], but it is still non-linear on (W, Θ). Nonetheless, we can solve it using alternating minimization, by fixing W or Θ, transforming each of these problems into a proper LS-SVM.

According to the CCCP algorithm [33] used to solve the LS-SVM, if the problem can be factored as a sum of a convex and a concave term, it can be efficiently solved by iterating between the optimization of the concave and the convex parts leading to a local minimum or saddle point.

Returning to our case, Eq. (15) can be rewritten as two different unconstrained problems, fixing Θ and W, respectively:

$$\min_W \frac{1}{2}\|W\|_F^2 + \frac{C_2}{N}\sum_{i=1}^N \max_{y,z} \sum_{l=1}^L \sum_{k=1}^K w_{y,l,k} \cdot \langle \theta_k, v^l_{z_{(l,k)}} \rangle + \Delta(y_i, y) - \frac{C_2}{N}\sum_{i=1}^N \max_{z_i} \sum_{l=1}^L \sum_{k=1}^K w_{y_i,l,k} \cdot \langle \theta_k, v^l_{z_{i,(l,k)}} \rangle \tag{16}$$

and

$$\min_\Theta \frac{C1}{2K}\|\Theta\|_F^2 + \frac{C_2}{N}\sum_{i=1}^N \max_{y,z} \sum_{k=1}^K \langle \theta_k, \sum_{l=1}^L w_{y,l,k} \cdot v^l_{z_{(l,k)}} \rangle + \Delta(y_i, y) - \frac{C_2}{N}\sum_{i=1}^N \max_{z_i} \sum_{k=1}^K \langle \theta_k, \sum_{l=1}^L w_{y_i,l,k} \cdot v^l_{z_{i,(l,k)}} \rangle \tag{17}$$

The structure of the above formulations corresponds to the difference of two convex terms. This gives rise to a strategy for solving the complete problem by alternating an optimization based on the CCCP algorithm. The proposed approach can be summarized in the following steps that are repeated until the energy defined by (15) stops decreasing:

1. Given fixed values of Θ and W, compute for each example the optimum value for its latent variables z_i as:

$$z_i = \underset{z}{\text{argmax}} \sum_{l=1}^L \sum_{k=1}^K w_{y_i,l,k} \cdot \langle \theta_k, v^l_{z_{(l,k)}} \rangle. \tag{18}$$

Eq. (18) shows that the descriptor vector selected for a region l and dictionary word θ_k depends directly on the values of $w_{y_i,l,k}$ and the inner product $\langle \theta_k, v^l_{z_{(l,k)}} \rangle$. To avoid problems related to negative weights, we enforce non-negativity constraints on W:

$$w_{y,l,k} \geq 0, \forall y, l, k. \tag{19}$$

In this way, the value of the inner product will only be scaled by $w_{y,l,k}$, thus preserving the semantics of max-pooling. As a consequence z_i now only depends on Θ, thus making it unnecessary to update the latent variables after recomputing W.

2. Given fixed values of Θ and $\{z_i\}$, solve the following minimization that corresponds to a standard Structural SVM with non-negativity constraints on W,

$$\min_{W} \frac{1}{2}\|W\|_F^2 + \frac{C_2}{N}\sum_{i=1}^{N}\max_{y,z}\sum_{l=1}^{L}\sum_{k=1}^{K} w_{y,l,k}\cdot\langle\theta_k, v^l_{z_{(l,k)}}\rangle + \Delta(y_i,y) - \frac{C_2}{N}\sum_{i=1}^{N}\sum_{l=1}^{L}\sum_{k=1}^{K} w_{y_i,l,k}\cdot\langle\theta_k, v^l_{z_{i,(l,k)}}\rangle \quad (20)$$
$$\text{s.t.} \quad w_{y,l,k} \geq 0, \forall y, l, k.$$

3. Given fixed values of W and $\{z_i\}$, solve the following minimization that corresponds to a standard Structural SVM,

$$\min_{\Theta} \frac{C1}{2K}\|\Theta\|_F^2 + \frac{C_2}{N}\sum_{i=1}^{N}\max_{y,z}\sum_{k=1}^{K}\langle\theta_k, \sum_{l=1}^{L} w_{y,l,k}\cdot v^l_{z_{(l,k)}}\rangle + \Delta(y_i,y) - \frac{C_2}{N}\sum_{i=1}^{N}\sum_{k=1}^{K}\langle\theta_k, \sum_{l=1}^{L} w_{y_i,l,k}\cdot v^l_{z_{i,(l,k)}}\rangle. \quad (21)$$

Although the convergence to a local minimum or saddle point can not be theoreticaly guaranteed for our block coordinate descent method [25], experimentally we found that for a suitable selection of the regularization parameters, our procedure does converge. In practice, we repeat the three steps of the algorithm until the energy decrease is lower than a fixed threshold.

4. Experiments

We perform a series of evaluations on 3 different datasets: *Caltech 101*, *15 Scene Categories*, and *MIT67 Indoor*. We first focus on the analysis of the sensitivity of the results to the values taken by the regularization constants, C_1 and C_2, the size of the dictionary, K, and the number of pooling regions, L. Additionally, we compare our approach to alternative state-of-the art techniques.

4.1. Implementation Details

Feature extraction

Each image is first downsized to 300 pixels in each direction. Local descriptors are extracted from each image over a dense grid of regions of 16x16 pixels, with a spacing of 8 pixels in each direction. We use the HOG + LBP descriptor. To construct the descriptor for each image, we randomly select L rectangular regions of sizes between 25% to 100% of the image size and low overlap.

Starting dictionary

We sample 75 descriptors per training image and cluster them using the standard K-Means algorithm. A linear SVM is trained for each centroid using, as positive examples, the ones belonging to that centroid and, as negative examples, a random sample of descriptors belonging to other centroids.

Datasets details

- *Caltech101*: This dataset contains 102 object categories (101 objects plus background). We use 10 random splits of the data, using 30 images for training and the rest for testing.
- 15 *Scene Categories*: This dataset contains images of 15 natural scene categories. We use 10 random splits of the data, using 100 images for training and the rest for testing.
- *MIT67 Indoor*: This dataset contains 67 indoor scene categories having a large intra-class variation. We use the standard evaluation procedure, using 80 images per class for training and 20 for testing.

4.2. Results

4.2.1 Effects of the regularization constants

Regularization constants, C_1 and C_2, play a significant role in keeping generalization at high levels and avoiding overfitting to training data. Tuning their values corresponds to a key aspect in the performance of our algorithm.

For the constant C_2, our results show that values in the interval $(0.1, 1)$ lead to superior and stable performance. Values below 0.1 tend to dramatically decrease performance, and values greater than 1 lead to overfitting, quickly saturating the performance on the training set.

As for C_1, we consider the fact that each word in the starting dictionary is trained independently. In this way, the resulting dictionary has a large norm compared to the 0-1 Δ function we use. This means that a high value of C_1 compared to C_2, privileges the reduction of the term $||W||_F^2$, without taking into account the potential increase of the loss controlled by C_2.

In order to avoid the latter situation, we fix the value of C_1 based on the norm of the first estimation obtained for W. We found that a suitable rule to set the starting value of C_1 is given by:

$$\frac{C_1 \times ||\Theta||_F^2}{K} \approx ||W||_F^2 \quad (22)$$

Eq. (22) achieves a certain level of balance between generalization and learning, similar to the one obtained by the estimation of W. Indirectly, this rule also takes into account the value of the constant C_2, which has a direct impact on the value of the term $||W||_F^2$.

4.2.2 Effects of the number pooling regions and dictionary words

We use from 10 to 60 pooling regions and average the hit rate obtained in the three datasets. We fix the size of the dictionary to 300. Table 1 shows our results, where we also include the performance of the standard pyramidal decomposition using 3 levels (SP).

	Dataset		
# Regions	**Caltech 101**	**15 Scenes**	**MIT67**
10	65.7 ± 0.8	70.2 ± 0.6	30.5
SP (21)	70.7 ± 0.3	78.1 ± 0.4	33.2
20	71.5 ± 0.6	80.4 ± 0.3	35.6
30	72.9 ± 0.6	84.6 ± 0.4	39.5
40	72.7 ± 0.7	84.1 ± 0.5	39.3
50	72.0 ± 0.8	82.3 ± 0.4	38.9
60	70.1 ± 0.8	80.9 ± 0.4	38.2

Table 1. Recognition accuracy as a function of the number of pooling regions.

Table 1 shows a consistent improvement in performance as the number of regions increases; the best performance being reached for 30 pooling regions. It is interesting to note that the performance obtained by randomly selecting regions is consistently higher than the standard pyramidal decomposition, a result also shown in [11].

The size of the dictionary (number of words) is a central issue in this research. We perform tests changing the number of dictionary words between 50 and 500 and measure the average hit rate in the three datasets. We fix the number of pooling regions to 30, as our previous experiments suggested. Table 2 shows our results.

	Dataset		
# Words	**Caltech 101**	**15 Scenes**	**MIT67**
50	58.7 ± 0.8	72.1 ± 0.5	31.3
100	64.3 ± 0.7	80.5 ± 0.5	35.2
200	70.1 ± 0.5	84.6 ± 0.4	38.5
300	72.9 ± 0.6	83.5 ± 0.5	39.5
400	72.3 ± 0.5	83.2 ± 0.6	39.3
500	71.8 ± 0.5	82.8 ± 0.5	39.1

Table 2. Performance as a function of the dictionary size.

As it can be seen, there is a clear increase in performance when the dictionary size increases from 50 to 200 words. Near 200 words seems to be minimum size for this method to work correctly. After that, performance starts to slowly decrease on 15 Scenes. On Caltech 101 and MIT67 the situation is a bit different, as the performance continues to scale until it reaches its peak at around 300 words. This seems as a natural progression, since these two datasets have a notably larger amount of categories. These results confirm a key property of our method, as the increase in dictionary size is lower than linear with respect to the increase of target classes.

4.3. Classification performance

The next experiment compares our results against alternative methods based on BoVW representations. Table 3 shows the results. We also include a baseline method in the comparison. This method is the same approach described in this paper, but without the dictionary updating step (step 3 of the algorithm description). We observe that our proposed method achieves state-of-the-art performance in 15 Scene Categories and MIT67 Indoor, while obtaining competitive scores for Caltech101. An important aspect of our results is that they are achieved using only 200 or 300 dictionary words, while alternative methods usually use more than a thousand. Another interesting fact is the clear performance advantage that is achieved over the baseline method. Therefore, by updating the dictionary using information from the top level classifiers our method not only improves its raw performance compared to the baseline, but also achieves its peak performance using less visual words.

5. Conclusions and Future Work

In this work we present a novel method for visual recognition that uses a joint optimization scheme to learn a discriminative visual dictionary and the weights of a multiclass max-margin classifier. When compared with alternative BoVW approaches, our method achieves state-of-the-art performance on two of the three datasets used in our experiments (15 Scenes and MIT67).

Another key result of our method is the generation of a common dictionary of discriminative visual words among the target visual categories. This allows us to achieve state-of-the-art performance using an order of magnitude less visual words than previous approaches. This is a relevant issue that not only reduces the complexity of the underlying optimization problem, but also has an impact at testing time due to the reduced number of visual words. Future work will focus on using multiscale patches to enrich our hypothesis space to search for suitable visual words. We also plan to handle larger dictionaries to allow a suitable representation of datasets with a large number of visual categories.

References

[1] H. Bay, T. Tuytelaars, and L. V. Gool. Surf: Speeded up robust features. In *ECCV*, 2006. 3

[2] I. Biederman. Recognition-by-components: A theory of human image understanding. *Psychological Review*, 94:115–147, 1987. 2

[3] L. Bourdev, S. Maji, T. Brox, and J. Malik. Detecting people using mutually consistent poselet activations. In *European Conference on Computer Vision (ECCV)*, 2010. 1

Method	# Words	Dataset		
		Caltech101	15 Scenes	MIT67
Baseline	400	63.9 ± 0.6	78.1 ± 0.3	33.2
SPM [15]	400	64.6 ± 0.8	81.4 ± 0.5	-
LLC [28]	2048	**73.4**	80.5 ± 0.6	-
LCSR [22]	1024	73.2 ± 0.8	82.7 ± 0.5	-
ScSPM [31]	1024	73.2 ± 0.5	80.3	-
Max-margin [17]	5250	-	82.7 ± 0.5	-
Object Bank [16]	200 per class	-	80.9	37.6
Reconfigurable Models [21]	200	-	78.6 ± 0.7	37.9
Discriminative Patches [23]	210 per class	-	-	38.1
Proposed	200 or 300	72.9 ± 0.6	**84.6 ± 0.4**	**39.5**

Table 3. Our proposed method achieves state-of-the-art performance in 2 out of 3 datasets. In 15 Scenes, we only use 200 words, while in the other two sets we use 300 words.

[4] Y. Boureau, F. Bach, Y. LeCun, and J. Ponce. Learning mid-level features for recognition. In *CVPR*, 2010. 1, 2

[5] G. Csurka, C. R. Dance, L. Fan, J. Willamowski, and C. Bray. Visual categorization with bags of keypoints. In *In Workshop on Statistical Learning in Computer Vision, ECCV*, 2004. 1

[6] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005 (CVPR 2005).*, volume 1, pages 886–893, 2005. 1

[7] P. Felzenszwalb and D. Huttenlocher. Pictorial structures for object recognition. *International Journal of Computer Vision*, 61(1):55–79, 2005. 1

[8] P. Felzenszwalb, D. Mcallester, and D. Ramanan. A discriminatively trained, multiscale, deformable part model. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008. 1, 2

[9] G. E. Hinton and S. Osindero. A fast learning algorithm for deep belief nets. *Neural Computation*, 18:2006, 2006. 2

[10] A. Jain, L. Zappella, P. McClure, and R. Vidal. Visual dictionary learning for joint object categorization and segmentation. In *ECCV*, 2012. 3

[11] Y. Jia, C. Huang, and T. Darrell. Beyond spatial pyramids: Receptive field learning for pooled image features. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3370–3377, 2012. 6

[12] F. Jurie and B. Triggs. Creating efficient codebooks for visual recognition. In *ICCV*, 2005. 1

[13] A. Krizhevsky, I. Sutskever, and G. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012. 2

[14] S. Lazebnik and M. Raginsky. Supervised learning of quantizer codebooks by information loss minimization. *PAMI*, 31(7):1294–1309, 2009. 1

[15] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2169–2178, 2006. 1, 7

[16] E. P. X. Li-Jia Li, Hao Su and L. Fei-Fei. Object bank: A high-level image representation for scene classification & semantic feature sparsification. In *Neural Information Processing Systems (NIPS)*, Vancouver, Canada, December 2010. 7

[17] X.-C. Lian, Z. Li, B.-L. Lu, and L. Zhang. Max-margin dictionary learning for multiclass image categorization. In *European Conference on Computer Vision (ECCV)*, ECCV'10, pages 157–170, 2010. 7

[18] D. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004. 3

[19] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman. Supervised dictionary learning. In *Advances in Neural Information Processing Systems 21*, pages 1033–1040. 2008. 1

[20] F. Moosmann, B. Triggs, and F. Jurie. Fast discriminative visual codebooks using randomized clustering forests. In *Neural Information Processing Systems (NIPS)*, 2007. 1

[21] S. Parizi, J. Oberlin, and P. Felzenszwalb. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2775–2782, 2012. 7

[22] A. Shabou and H. Le-Borgne. Locality-constrained and spatially regularized coding for scene categorization. In *CVPR*, 2012. 7

[23] S. Singh, A. Gupta, and A. Efros. Unsupervised discovery of mid-level discriminative patches. In *ECCV*, 2012. 7

[24] J. Sivic and A. Zisserman. Video google: A text retrieval approach to object matching in videos. In *ICCV*, 2003. 1, 2

[25] P. Tseng. Convergence of a block coordinate descent method for nondifferentiable minimization. *Journal of Optimization Theory and Applications*, 109(3):475–494, June 2001. 5

[26] I. Tsochantaridis, T. Hofmann, T. Joachims, and Y. Altun. Support vector machine learning for interdependent and structured output spaces. In *ICML*, 2004. 4

[27] P. Viola and M. Jones. Robust real-time face detection. *International Journal of Computer Vision*, 57(2):137–154, 2004. 1

[28] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong. Locality-constrained linear coding for image classification. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010. 7

[29] Y. Wang and G. Mori. Hidden part models for human action recognition: Probabilistic versus max margin. *PAMI*, 33(7):1310–1323, 2011. 2

[30] J. Winn, A. Criminisi, and T. Minka. Object categorization by learned universal visual dictionary. In *ICCV*, pages 1800–1807, 2005. 1

[31] J. Yang, K. Yu, Y. Gong, and T. Huang. Linear spatial pyramid matching using sparse coding for image classification. In *CVPR*, 2009. 1, 2, 3, 7

[32] C. J. Yu and T. Joachims. Learning structural svms with latent variables. In *Proceedings of the 26th Annual International Conference on Machine Learning (ICML)*, 2009. 4

[33] A. L. Yuille and A. Rangarajan. The concave-convex procedure. *Neural Computation*, 15(4):915–936, Apr. 2003. 4