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

Hans Lobel[1], René Vidal[2], Domingo Mery[1], and Alvaro Soto[1]

[1] Department of Computer Science, Ponficia Universidad Católica de Chile
[2] Center for Imaging Science, Johns Hopkins University

Abstract. The Bag-of-Visual-Words (BoVW) model is a popular approach for visual recognition. Used successfully in many different tasks, simplicity and good performance are the main reasons for its popularity. The central aspect of this model, the visual dictionary, is used to build mid-level representations based on low level image descriptors. Classifiers are then trained using these mid-level representations to perform categorization. While most works based on BoVW models have been focused on learning a suitable dictionary or on proposing a suitable pooling strategy, little effort has been devoted to explore and improve the coupling between the dictionary and the top-level classifiers, in order to generate more discriminative models. This problem can be highly complex due to the large dictionary size usually needed by these methods. Also, most BoVW based systems usually perform multiclass categorization using a one-vs-all strategy, ignoring relevant correlations among classes. To tackle the previous issues, we propose a novel approach that jointly learns dictionary words and a proper top-level multiclass classifier. We use a max-margin learning framework to minimize a regularized energy formulation, allowing us to propagate labeled information to guide the commonly unsupervised dictionary learning process. As a result we produce a dictionary that is more compact and discriminative. We test our method on several popular datasets, where we demonstrate that our joint optimization strategy induces a word sharing behavior among the target classes, being able to achieve state-of-the-art performance using far less visual words than previous approaches.

1 Introduction

Bag-of-Visual-Words (BoVW) [24] is currently one of the most popular approaches for solving visual recognition problems, like scene or object detection and categorization. BoVW models encode information using a mid-level representation based on a dictionary of visual words, which encodes appearance information from local patches [15]. To perform categorization, these models are combined with top-level supervised classifiers that are trained using the mid-level representations. Spatial information is

Acknowledgment: This work was partially funded by FONDECYT grant 1120720.

also usually incorporated into these models by concatenating information from different spatial areas [11].

The construction of the visual dictionary is one the most important aspect of these models. Commonly, it is built using generative approaches that minimize errors in patch reconstruction, such as vector quantization [24] or sparse coding [32] techniques, In these cases dictionary construction is decoupled from the training of top-level classifiers [24,18]. As an alternative, discriminative dictionary construction strategies have also been proposed but mostly considering a weak link to the training of top-level classifiers [30,11].

Regarding the categorization process, BoVW models generally use a one-versus-all classification strategy. Unfortunately, this scheme does not considers relevant correlations among classes. Furthermore, they usually employ a visual dictionary for each target class, an aspect that turns out to be critical when the number of classes increases.

In this work we introduce a novel approach for visual recognition that presents two main contributions. Our first contribution is a method that jointly learns a suitable BoVW representations and top-level classifiers using a multiclass max-margin approach. The mathematical formulation behind this method is very general, however, we focus on a BoVW representation using a Spatial Pyramid Matching scheme. In contrast to previous works, we model dictionary words as linear SVMs [22] using a direct multiclass scheme. This allow us to pose our formulation as an energy minimization problem. Furthermore, we combine the responses of these discriminative dictionary words using a max-pooling strategy, as several recent works have shown the superiority of max-pooling over alternative polling strategies [31,2].

Our second contribution is a direct result of our learning scheme. By using a joint optimization strategy, we are able to induce word sharing among target classes. This allows us to achieve state-of-the-art categorization performance in several common benchmarks datasets, using an order of magnitude less words than previous approaches. We believe that word sharing is a critical issue to the scalability of visual recognition algorithms.

2 Related Work

BoVW model has remained throughout the last years as one of the the most common strategies for visual recognition, mainly for its simplicity and good results [24,18]. At the heart of this strategy lies the visual dictionary, which is used to quantize descriptor vector extracted from images. Dictionary learning is generally performed using an unsupervised method, like *K-Means*, to cluster the extracted descriptors [24,3,9,11,30,17,12].

In order to increase BoVW performance, sparse coding techniques have also been used as a method for learning a visual dictionary and to quantize feature descriptors. A remarkable example of this is [31], where sparse coding, max-pooling, and linear SVMs are used to achieve excellent categorization performance. Discriminative sparse representations have also been proposed [16], mostly building particular dictionaries for each class. In [2,14], the coupling of dictionary and classifier learning is explored. Altough similar in spirit to our work, here we explore a stronger form of coupling, consisting of a shared dictionary of linear SVMs and a multiclass classifier, instead of

the standard one-vs-all framework. Also, they use different methods to obtain the visual words and to build the mid-level representation used by the top-level classifier.

Deep belief networks (DBN) [7,10] applied to visual recognition also present some similarities to our work, mainly regarding spatial pooling schemes and intermediate representations based on linear filters. As a consequence of a multi-layered generic structure, DBNs have many parameters and they are usually difficult to train. This presents a main difference to our work, as we embed semantic knowledge to our model by explicitly considering compositional relations among low level visual features, mid-level visual words, and top-level classifiers, leading to a more meaningful and simpler architecture. Also, the Hinge loss function used in our work leads to a different mathematical formulation.

Max-margin schemes have also been successfully used for visual recognition. Currently, one of the most used schemes is the one presented in [6], where a latent SVM is used to learn a mixture of multiscale deformable part models for binary classification. Also, in the area of action recognition, [29] proposes an extension to the method of [6] that uses a multiclass classification scheme. Although our discriminative BoVW model is also based on a max-margin approach, our hierarchical formulation, mid-level pooling scheme, and training scheme are highly different that the ones used by part-based approaches. In particular, our formulation is able to scale to cases that involve hundreds of parts, while part-based approaches are designed to operate with a reduced set of parts.

3 Model Description

3.1 Image Representation

We assume that visual descriptors [4,1] 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. Also, inspired by the work of [8], 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:

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

To encode each descriptor, we use an encoding scheme based on the classification score obtained by each dictionary word, similar to the one presented in [13]. More specifically, if v is a descriptor vector, its coding based on the dictionary Θ, $c_\Theta(v)$, is defined as:

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

In order to use dictionary words with a bias term as defined above, every descriptor vector v has a constant 1 appended at the end. Intuitively, if the visual words are sufficiently discriminative, the descriptor v should be similar to only a few words from the dictionary. Therefore, we expect the vector $c_\Theta(v)$ to have only a few values greater than zero.

Given a dictionary Θ and a spatial pyramidal decomposition of the image into L regions, we represent the image using *max spatial pooling*. For each region l, $l \in [1, L]$, let $v_{l,j}$ be its j-th descriptor vector, where $j \in [1, N_l]$ and N_l is the number of descriptors extracted from region l. Given a dictionary Θ, we encode region l using *max spatial pooling* as:

$$x_{l,\Theta} = [\max_{j=1}^{N_l} v_{l,j}^T \theta_1,\ \max_{j=1}^{N_l} v_{l,j}^T \theta_2,\ \ldots,\ \max_{j=1}^{N_l} v_{l,j}^T \theta_K]^T \in \mathbb{R}^K. \tag{4}$$

As we are working in a discriminative setting instead of a generative one, like sparse coding generated dictionaries, our scheme assigns negative weights instead of a zero-weight to dictionary words with low similarity. As this can potentially lead to overfitting issues, we assume that each region l contains a null feature vector $\mathbf{0}$, whose classification score is equal to zero for any of the dictionary words. Using this trick, if in a given region none of the extracted feature vectors obtains a positive score, the max score over the region will be obtained by the null feature vector, thus putting a zero weight on the region descriptor.

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

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

Figure 1 shows a diagram of the creation of the coding of an image region.

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

$$E(I, y, \Theta, W) = \quad \sum_{l}^{L} \sum_{k}^{K} w_{y,l,k} \cdot \max_{j=1}^{N_l} (v_{l,j}^T \theta_k). \tag{8}$$

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 linear dependence on the weights w_y, but a nonlinear one on the dictionary words. Figure 2 shows a schematic view the construction of the energy function.

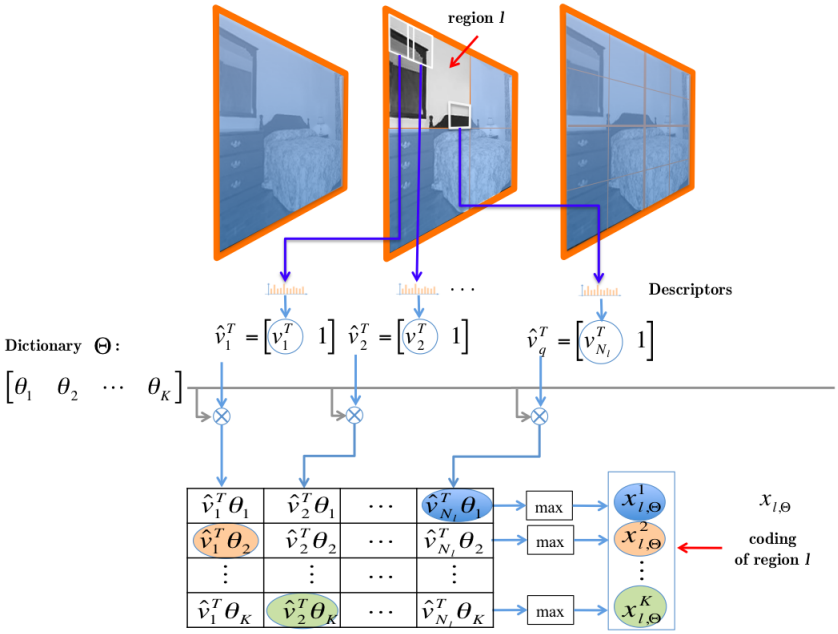

Fig. 1. Diagram of the coding of an image region l.

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

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

4 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:

$$\begin{aligned} \min_{W,\Theta,\{\xi_i\}} & \frac{1}{2}\|W\|_F^2 + \frac{C_1}{2K}\|\Theta\|_F^2 + \frac{C_2}{N}\sum_{i=1}^{N}\xi_i \\ \text{s.t.} \quad & E(I_i, y_i, \Theta, W) - E(I_i, y, \Theta, W) \geq \Delta(y_i, y) - \xi_i, \\ & \forall i \in \{1, \ldots, N\} \wedge \forall y \in \{1, \ldots, M\}. \end{aligned} \tag{10}$$

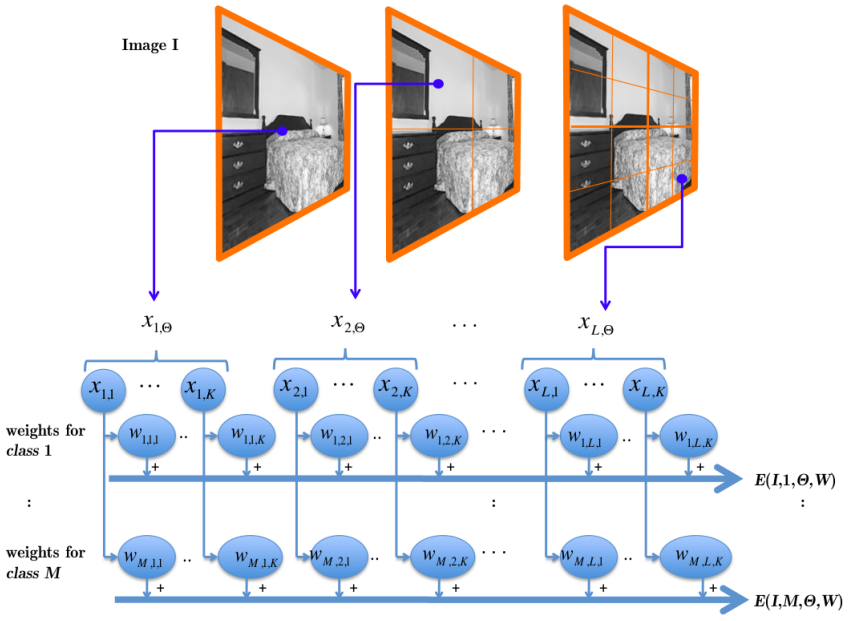

Fig. 2. Schematic view the construction of the energy function.

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.

At first sight, one could think that the formulation in (10) is a particular case of Structural SVM (S-SVM) [25]. However, this is not the case due to two fundamental differences. First, the constraints are not linear in Θ, while in S-SVMs the constraints are always linear in the parameters. Second, the optimization problem is not jointly convex on Θ and W.

To see the latter, notice that the optimal solution for the slack variables is given by

$$\xi_i^*(\Theta, W) = \max_y (E(I_i, y, \Theta, W) + \Delta(y_i, y)) - E(I_i, y_i, \Theta, W) \tag{12}$$

and recall that the point-wise maximum of convex functions is convex. Therefore, given $W \geq 0$, the energy in (8) is convex in Θ because it is the weighted sum of convex functions with nonnegative weights. By the same argument, the first term in (12) is

convex, however the second one is concave. As a consequence, the objective function for Θ given $W \geq 0$ is the sum of a convex and a concave function in Θ. Otherwise, the cost function is generally non-convex. On the other hand, notice that given Θ, the objective function is convex in W, because the first term of (12) is a maximum of convex functions, while the second term is linear in W.

Motivated by the above analysis, we propose to solve the problem in (10) using an alternating minimization approach where we alternate between the computation of W given a fixed Θ and the computation of Θ given a fixed W. Due to the non-convexity of the cost function, there is no theoretical analysis guaranteeing its convergence. However, our experiments show that for suitable selection of the parameters, our procedure does converge in practice.

Given $\Theta = \Theta^{(t)}$ at iteration t, the computation of W reduces to a standard multi-class SVM problem, which can be efficiently solved, *e.g.*, with a cutting-plane algorithm [25]. Such methods typically produce a solution for both the classifier parameters $W^{(t)}$ and the slack variables $\xi_i^{(t)}$.

Given $W = W^{(t)}$ at iteration t, the computation of Θ requires solving the following problem

$$\min_{\Theta} \frac{C_1}{2K}\|\Theta\|_F^2 + \frac{C_2}{N}\sum_i^N E(I_i, \hat{y}_i, \Theta, W^{(t)}) + \Delta(y_i, \hat{y}_i) - E(I_i, y_i, \Theta, W^{(t)}) \quad (13)$$

where

$$\hat{y}_i = \operatorname*{argmax}_y E(I_i, y, \Theta, W) + \Delta(y_i, y). \quad (14)$$

As stated before, the optimization problem in (13) is not convex, hence we can not guarantee that we find a global minimizer. We find an approximate solution by using an interior point method [26] applied to an approximation of a modified version of (13). The modification consists of solving the problem in (13) subject to the additional constraints

$$\begin{aligned} &E(I_i, y, \Theta, W^{(t)}) - E(I_i, y_i, \Theta, W^{(t)}) + \Delta(y_i, y) \leq \xi_i^{(t)} \\ &\forall i \in \{1, \ldots, N\} \wedge \forall y \in \{1, \ldots M\}. \end{aligned} \quad (15)$$

These additional constraints ensure that the new slack variables (after modifying Θ) are at most equal to the slack variables at the previous iteration (obtained after modifying W). The approximation consists of replacing the max-pooling function by a *soft-max* approximation to ensure differentiability. Specifically, we use the *log-sum-exponential* (LSE) approximation,

$$\max_{i=1}^{N}(z_i) \approx \frac{1}{r}\log(\sum_i^N \exp(r z_i)) \quad (16)$$

which preserves the convexity of the point-wise maximum operation. The parameter r controls the sharpness of the approximation, where larger values generate better approximations. The energy function finally becomes:

$$E(I, y, \Theta, W) \approx \sum_l^L \sum_k^K \frac{w_{y,l,k}}{r} \log(\sum_j^{N_l} \exp(r \cdot \hat{v}_{l,j}^T \theta_k)). \quad (17)$$

Te solve the approximate optimization problem we require the partial derivatives of the objective function and the partial derivatives of the constraints, both of which are straightforward to obtain from (17).

5 Experiments

We performed evaluations on 3 different datasets (Caltech 101, 15 Scene Categories, and MIT67 Indoor) and analyzed the potential performance gains delivered by the dictionary update step and classification performance compared to other methods. We obtain as main results i) state-of-the-art performance on 2 datasets when compared to other similar methods and ii) the generation of a unique dictionary of discriminative patches shared among target classes, with an order of magnitude less visual words than similar approaches.

5.1 Implementation Details

- **Feature extraction:** Images are downsized to no more than 300 pixels in each direction. Local HOG-LBP descriptors [28] are then extracted from each image using a dense grid of regions of 16x16 pixels, with a spacing of 8 pixels in each direction. We use a spatial pyramidal decomposition with 21 regions (depth 2).
- **Initial dictionary:** To obtain the initial dictionary, we sample 100.000 descriptors from training images and cluster them with *K-Means*. After this process, a linear SVM is trained for each centroid, using as positive examples the ones belonging to that centroid and as negative examples descriptors belonging to the other centroids.
- **Nonlinear optimization:** To implement the gradient descent step for the dictionary estimation, we use the interior point solver Ipopt (Interior Point OPTimizer) [26]. This solver is optimized for large scale constrained nonlinear problems as our case.

5.2 Datasets details

- **Caltech101:** This dataset is formed by 101 object categories plus a background class. We use 10 random splits of the data, keeping 30 images for training and the rest for testing.
- **15 Scene Categories:** This dataset contains 15 diferent natural scene categories. We use 10 random splits of the data, keeping this time 100 images for training and the rest for testing.
- **MIT67 Indoor:** This dataset contains 67 different indoor scene categories, with a very high intra-class variation. We use the standard evaluation procedure, using 80 images per class for training and 20 for testing.

5.3 Behavior of dictionary words

The purpose of this experiment is to visually appreciate some dictionary words before and after the dictionary update process. Figure 3 shows a group of patches from 15

Scene Categories that initially obtain high response for a dictionary word that encourages diagonal lines. The six patches on the left obtain a high score before and after the dictionary update, while the four patches on the right obtain a high score before the update, but after it obtain a very low score, using the same dictionary word.

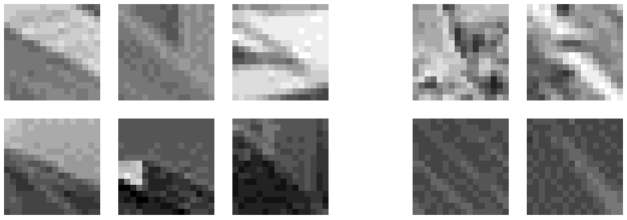

Fig. 3. The dictionary update step gradually reduces the score of noisy patches (right).

The set of patches on the left in Figure 3 show a more homogeneous appearance than the ones on the right. Despite also showing diagonal lines, the presence of noise in the rightmost patches, makes them more prone to be confused with other patches that show structures diferent than the one represented by that dictionary word. Our algorithm, using the max-pooling and the classifier weights, its able to progresively reduce the score of these patches, by updating the dictionary words accordingly.

5.4 Classification performance

As we mentioned, we test our method using three datasets: Caltech101, 15 Scene Categories, and MIT67 Indoor. Table 1 shows our first experiment designed to evaluate performance evolution as a function of the number of dictionary words. There is a ma-

	Number of Words		
Dataset	**50**	**100**	**200**
Caltech101	63.1 ± 0.8	72.0 ± 0.5	73.1 ± 0.5
15 Scenes	72.2 ± 0.5	83.7 ± 0.2	84.8 ± 0.2
MIT67 Indoor	31.2	38.3	39.9

Table 1. Performance evolution in function of number of dictionary words.

sive performance gain when dictionary size grows from 50 to 100 words. After that, the gain is less significant but clearly measurable. Unfortunately, due to memory requirements associated to our current implementation, it was not possible to test with a higher number of words.

The next experiment compares our results against methods using only a BoVW scheme, without the aid of global features [19] or object models [6], and a fixed spatial

		Dataset		
Method	**# Words**	Caltech101	15 Scenes	MIT67
Baseline	200	63.9 ± 0.6	78.1 ± 0.3	33.2
SPM [12]	400	64.6 ± 0.8	81.4 ± 0.5	-
LLC [27]	2048	**73.4**	80.5 ± 0.6	-
LCSR [21]	1024	73.2 ± 0.8	82.7 ± 0.5	-
ScSPM [31]	1024	73.2 ± 0.5	80.3	-
Max-margin [14]	5250	-	82.7 ± 0.5	-
Object Bank [13]	200	-	80.9	37.6
Reconfigurable Models [20]	200	-	78.6 ± 0.7	37.9
Discriminative Patches [23]	14070	-	-	38.1
Proposed	200	73.1 ± 0.5	$\mathbf{84.8 \pm 0.2}$	**39.9**

Table 2. Our proposed method achieves state-of-the-art performance in 2 out of 3 datasets using only 200 words.

pyramid matching scheme with at most depth 2, *i.e.* 21 pooling regions. Table 2 shows the results. We also include a baseline method in the comparison, consisting in only solving the problem in 10 with a fixed Θ.

We can observe that we achieve state-of-the-art performance in 15 Scene Categories and MIT67 Indoor, while obtaining competitive results in Caltech101. An important aspect of our results is that we use only 200 dictionary words, where other methods use more than thousand. In particular, in the case of Caltech101, we believe that our current dictionary size might be hurting performance, as the high number of categories might require a larger dictionary size for better results.

6 Conclusions and Future Work

BoVW models are commonly based on two main steps: first learning the dictionary and then learning classifiers that use this dictionary. In this work we present a novel method for jointly learning a dictionary of discriminative visual words and a top-level classifier using a multiclass max-margin approach. Our formulation is highly general an can be adapted to various spatial decompositions of images. In particular, when we compare several techniques based on a spatial pyramid matching scheme, our method achieves state-of-the-art performance on two of the three datasets considered in our experiments (15 Scenes and MIT67).

Regarding the resulting dictionary, the proposed joint learning scheme produces a strong sharing of visual words among the target classes. In practice, this allows us to use a dictionary that has an order of magnitude less visual words than previous BoVW methods, but without reducing recognition performance. As the number of target object classes increases in practical applications, word sharing will become a relevant issue, since time spent obtaining responses for different linear classifiers (at the level of dictionary words and top-level classifier) will be a major processing bottleneck.

Future work will focus on using multiscale patches to enrich our hypothesis space, thus allowing us to search for suitable dictionary words that represent more meaningful

visual structures. We also plan to improve running time and memory use, in order to be able to evaluate performance on new larger datasets [5].

References

1. T. Ahonen, A. Hadid, and M. Pietikainen. Face description with local binary patterns: Application to face recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 28(12):2037–2041, 2006.
2. Y. Boureau, F. Bach, Y. LeCun, and J. Ponce. Learning mid-level features for recognition. In *CVPR*, 2010.
3. 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.
4. 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.
5. J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.
6. P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1627–1645, 2010.
7. G. E. Hinton and S. Osindero. A fast learning algorithm for deep belief nets. *Neural Computation*, 18:2006, 2006.
8. A. Jain, L. Zappella, P. McClure, and R. Vidal. Visual dictionary learning for joint object categorization and segmentation. In *ECCV*, 2012.
9. F. Jurie and B. Triggs. Creating efficient codebooks for visual recognition. In *ICCV*, 2005.
10. A. Krizhevsky, I. Sutskever, and G. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012.
11. S. Lazebnik and M. Raginsky. Supervised learning of quantizer codebooks by information loss minimization. *PAMI*, 31(7):1294–1309, 2009.
12. 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.
13. Eric P. Xing Li-Jia Li, Hao Su and Li 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.
14. Xiao-Chen Lian, Zhiwei Li, Bao-Liang Lu, and Lei Zhang. Max-margin dictionary learning for multiclass image categorization. In *European Conference on Computer Vision (ECCV)*, ECCV'10, pages 157–170, 2010.
15. D. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004.
16. Julien Mairal, Francis Bach, Jean Ponce, Guillermo Sapiro, and Andrew Zisserman. Supervised dictionary learning. In *Advances in Neural Information Processing Systems 21*, pages 1033–1040. 2008.
17. Frank Moosmann, Bill Triggs, and Frederic Jurie. Fast discriminative visual codebooks using randomized clustering forests. In *Neural Information Processing Systems (NIPS)*, 2007.
18. J.C. Niebles, H. Wang, and F. Li. Unsupervised learning of human action categories using spatial-temporal words. *IJCV*, 79(3):299–318, 2008.

19. Aude Oliva and Antonio Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42:145–175, 2001.
20. S.N. Parizi, J.G. Oberlin, and P.F. Felzenszwalb. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2775–2782, 2012.
21. A. Shabou and H. Le-Borgne. Locality-constrained and spatially regularized coding for scene categorization. In *CVPR*, 2012.
22. D. Singaraju and R. Vidal. Using global bag of features models in random fields for joint categorization and segmentation of objects. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.
23. S. Singh, A. Gupta, and A. Efros. Unsupervised discovery of mid-level discriminative patches. In *ECCV*, 2012.
24. Josef Sivic and Andrew Zisserman. Video google: A text retrieval approach to object matching in videos. In *ICCV*, 2003.
25. I. Tsochantaridis, T. Hofmann, T. Joachims, and Y. Altun. Support vector machine learning for interdependent and structured output spaces. In *ICML*, 2004.
26. Andreas Waechter and Lorenz T. Biegler. On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming. *Mathematical Programming*, 106:25–57, 2006.
27. 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.
28. Xiaoyu Wang, T.X. Han, and Shuicheng Yan. An hog-lbp human detector with partial occlusion handling. In *IEEE International Conference on Computer Vision (ICCV)*, pages 32–39, 2009.
29. Y. Wang and G. Mori. Hidden part models for human action recognition: Probabilistic versus max margin. *PAMI*, 33(7):1310–1323, 2011.
30. J. Winn, A. Criminisi, and T. Minka. Object categorization by learned universal visual dictionary. In *ICCV*, pages 1800–1807, 2005.
31. J. Yang, K. Yu, Y. Gong, and T. Huang. Linear spatial pyramid matching using sparse coding for image classification. In *CVPR*, 2009.
32. L. Yang, R. Jin, R. Sukthankar, and F. Jurie. Unifying discriminative visual codebook generation with classifier training for object category reorganization. In *CVPR*, 2008.