Learning Semantic Visual Dictionaries: A new Method For Local Feature Encoding

Bing Shuai, Zhen Zuo, Gang Wang
School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore.
Email: {bshuai001, zzuo1, wanggang}@ntu.edu.sg

***Abstract*—In this paper, we develop a new method to learn semantic visual dictionaries for local image feature encoding. Conventional methods usually learn dictionaries from random local image patches. Different from them, we manually select a number of object classes whose visual patterns can be seen at local image patch level in complex images (Figure 1), and learn dictionaries from their thumbnail images. The benefit is that these thumbnail images have class labels, so we can cluster semantically similar images together to generate meaningful cluster centers. Some other contributions of this paper include developing an adaptation method to adapt the learned dictionaries to target datasets, and developing efficient algorithms to encode local patches with our semantic visual dictionaries. Experimental results on three benchmark datasets demonstrate the effectiveness of the proposed methods.**

***Index Terms*—Semantic Dictionary Learning, Greedy Group Sparse Coding, Scene Classification.**

I. Introduction

Building visual dictionaries for local image feature encoding is an important component for modern image classification pipelines [1], [2], [3]. The dictionary atoms are expected to encode distinctive and representative visual patterns, which can act as visual primitives to represent complicated image content. Previous methods learn dictionaries by clustering features from random image patches, which have several limitations: (1) No precise category labels can be assigned to local patches. Therefore, clustering methods may group patches which are not semantically similar together. The resulting clustering centers then cannot represent meaningful visual patterns. (2) Most cluster centers may correspond to high-frequent local patterns due to the lack of proper supervision in the clustering procedure[4]. In many cases, the visual patterns of low-frequent image patches play a significant role in discriminating different classes, for example, class *"bocce"* and *"croquet"* in UIUC-Sports dataset[5] are almost the same except several *cross* and *line* patterns appearing in *"croquet"*.

To overcome these limitations, we propose a novel idea to build semantic visual dictionaries. We observe that many simple object categories exhibit basic visual patterns that can be seen at local patch level in complex-content images. One example is shown in Figure 1. We manually select a number of object categories, which are expected to cover diverse visual patterns useful for the classification task. We resize the object-centric images to produce thumbnails, which are utilized to learn dictionaries. An interesting property of this method is that the class labels of these thumbnails are known. During the dictionary learning procedure, we only cluster thumbnails from the same object class together. As a result, the cluster centers correspond to visual patterns with semantic meanings. Besides, by carefully selecting object categories, we could also ensure that our dictionary encodes diverse patterns rather than dominant ones.

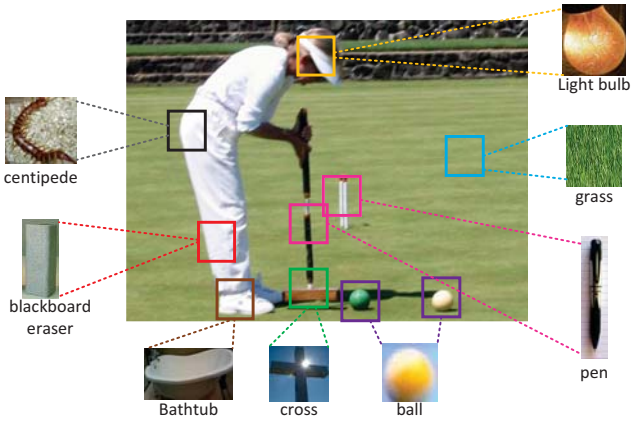


Fig. 1. Image patches in complex images can be represented by simple object categories, which exhibit basic yet important elementary visual patterns. For example, "centipede" category can approximate *"curve"* visual pattern, "pen" can be a substitute to *"line"* pattern, and "ball" carries *"circle"* pattern, etc.

Each sub-dictionary is learned from an auxiliary dataset (ImageNet), and the concatenation of them constitute our global dictionary. So it naturally has a group/block structure, in which each group corresponds to an object semantics. We exploit this group structure when encoding a local image patch by selecting only a small quantity of groups. However, it is inefficient to solve this problem through Group Lasso solver [6]. Therefore, two efficient greedy coding algorithms are proposed to generate the group sparse codes.

We further develop a domain adaption approach to adapt the learned dictionaries to fit the target database (e.g., Indoor scene 67) by learning a linear transformation matrix. On three benchmark datasets, we demonstrate that *the state-of-the-art coding method [2] can benefit from our adapted semantic dictionary.* Meanwhile, the experiments also show that our coding methods can achieve promising results based on our semantic visual dictionary.

978-1-4799-8058-1/15/$31.00 ©2015 IEEE.

II. Related Work

Dictionary Learning: Dictionary learning has received increasingly attention recently, as the performance of image classification accuracy heavily depends on the quality of dictionaries. A number of papers have been published [7], [8], [9], [10], [11], [12], [13], [14], [15] in the past several years. All of these papers learn dictionaries from random image patches. Our work is novel because we use thumbnails of object images to learn visual dictionaries, which can represent visual patterns with semantic meanings. The work that is closest to ours is to learn a discriminative dictionary for classification [7], [10], [11], [12]. They treat the labels of patches to be the same with the labels of corresponding images. However, the labels assigned to patches in this way is quite noisy. In our work, the labels for the thumbnail images are accurate and semantic meaningful, therefore more discriminative dictionary are expected to learn from them.

Using auxiliary object data for classification: The idea of using auxiliary object data to help classification or retrieval can be seen in [16], [17], [18], [19]. Different from our work, these papers build the global representation based on their responses to the pre-trained object classifiers/detectors. We leverage auxiliary object data to learn semantic dictionary to improve the quality of local feature encoding based on the assumption that local patches in images can be approximated by object-centric image thumbnail. Moreover, we empirically demonstrate that our approach works much better than [17] on several scene classification benchmarks.

III. Learning semantic visual dictionaries

We observe that many object categories exhibit basic yet important visual patterns that can be seen at local patch level in complex content images (Figure 1). It motivates us to build visual semantic dictionaries by clustering the object-centric image thumbnails. Instead of clustering random local patches, the dictionary generated from labeled thumbnails carry more discriminative and semantic information. Specifically, their class labels are known, which allows the clustering algorithms to group images from the same category, thus avoiding producing clusters whose atoms are not really semantically similar, as seen in the traditional dictionary learning methods.

We manually select 55 object categories based on the following criteria: (1) the object categories exhibit simple and elementary visual patterns; (2) the visual appearance of different object classes does not overlap and (3) some textureless background classes are also included. The selected objects include "pen", "box", "cross", "grass", etc, which are supposed to carry important and basic visual patterns. We download their images from ImageNet dataset, and bounding boxes are used to generate clean foreground object region. Next, each object region image is resized to $I \times I$ thumbnail ($I = 40$ in our case), so they can be considered as local image patches. For each object class, we learn a sub-dictionary $D^{(i)}$, and the concatenation of sub-dictionaries yields our final group-structured semantic dictionary D.

IV. Feature encoding based on semantic visual dictionary

Our semantic dictionary is organized in group structure ($D = [D^{(1)} \dots D^{(i)} \dots D^{(M)}]$), where $D^{(i)}$ contains K cluster centers $[D^{(i)}(1) \dots D^{(i)}(K)]$. Previous works [6], [20], [21] have shown that, in some cases, it's more robust to reconstruct the input signal by imposing group/block structure to its reconstruction coefficient.

Here we develop a local feature encoding approach to account for the group-structured dictionary: only a few semantic concepts (groups) are activated to reconstruct the image patch. [21] Therefore, the group sparse code β for the input y is generated via optimizing the function:

$$\begin{aligned} z &= \underset{\beta}{\operatorname{argmin}} \ ||y - D\beta||_2^2 \\ s.t.\ &||\beta||_0 \leq L \\ &||\dots||\beta^{(i)}||_1, \dots||_0 \leq T,\ 1 \leq i \leq M \end{aligned} \tag{1}$$

where β is the reconstruction coefficient of input y, $\beta^{(i)}$ denotes the code for the i-th group in the dictionary D; T and L are the maximum number of selected groups and selected atoms respectively; M is the number of groups in the group-structured dictionary D and $||\beta||_0$ is the pseudo-norm which counts the number of non-zero elements in β. This optimization problem is NP hard, which requires $\sum_j^T \binom{M}{j} \times \sum_i^L \binom{N_j}{i}$ evaluations, in which N_j is the total number of atoms in the selected $j (j \leq T)$ groups. We have to relax it to obtain a good suboptimal solution.

The first relaxation case is group Lasso [6], which only considers group-level sparsity, and it minimizes the ℓ_1/ℓ_2 norm regularization problem [6]. A better relaxation case is the sparse group Lasso [20], which enforces the sparsity over the group and within-group level reconstruction coefficients. Both of these two relaxed problems are convex to β, so it can be solved using convex optimization solvers. However, these two approximations are computationally expensive [6], [20].

To speed up the encoding modules, we approach the original NP hard problem (Equation 1) from a different perspective. Our relaxed problem is cast as a Lasso problem inside the selected few selected groups. It has the ℓ_0/ℓ_1 norm form:

$$\begin{aligned} z &= \underset{\beta}{\operatorname{argmin}} \ ||y - D\beta||_2^2 + \lambda||\beta||_1 \\ s.t.\ &||\dots||\beta^{(i)}||_1, \dots||_0 \leq T,\ 1 \leq i \leq M \end{aligned} \tag{2}$$

where λ controls the tradeoff between reconstruction error and within group sparsity. The group-structured reconstruction code β is generated through the following way: the coefficients are set to zero for the groups that are not selected, and the codes β^T for the selected T groups are obtained by solving a small-scale Lasso problem:

$$z = \underset{\beta^T}{\operatorname{argmin}} \ ||y - D_T\beta^T||_2^2 + \lambda||\beta^T||_1 \tag{3}$$

where D_T is a smaller dictionary made up atoms from T selected groups. This relaxed optimization problem is also a NP combinatorial problem, which requires $\sum_i^T \binom{M}{i} \times \mathcal{C}$

Algorithm 1: Greedy Group Sparse Coding

Input:
D: Dictionary with group structure $[G_1, G_2, \ldots, G_M]$
y: The input signal
T: Maximum number of selected groups
λ: sparse regularization term coefficient
ϵ: tolerance of reconstruction error
Output:
$\beta^{(k)}$: the group sparse coefficients

`Initialization`
`Normalize` each atom in D to be unit vector
$\beta^{(0)} = \mathbf{0}^T, g^{(0)} = \emptyset, D_O^{(0)} = \emptyset$
for k = *1...T* **do**
  Let $j^{(k)} = \text{argmax}_j \ ||D_{G_j}^T (y - D\beta^{(k-1)})||_2$
  $g^{(k)} = \texttt{Union}(g^{(k-1)},\ G_{j^{(k)}})$
  $D_O^{(k)} = \texttt{Union}(D_O^{(k-1)},\ D(:, G_{j^{(k)}}))$
  $\alpha^\star = \text{argmin}_\alpha \ ||y - D_O^{(k)}\alpha||_2^2 + \lambda||\alpha||_1$
  $\beta^{(k)} = \mathbf{0}^T,\ \beta^{(k)}(g^{(k)}) = \alpha^\star$
  if $||(y - D\beta^{(k)})||_2 \leq \epsilon$ **then**
    break
  end
end

evaluations, in which M is the number of groups and $\mathcal{C}$ is the cost of solving Equation 3. The computation cost to solve our ℓ_0/ℓ_1 problem is much less compared to group Lasso, which requires $\sum_i^L \binom{N_j}{i}$ evaluations to Equation 1. Furthermore, two efficient encoding methods are proposed to select T groups to solve it:

- **Greedy Group Sparse Coding(GGSC):** It works like Orthogonal Matching Pursuit[22] (OMP) that selects the dictionary atoms. At each iteration, it chooses the groups that best correlates to the signal residue. Algorithm 1 present the details.
- **Local-constrained Linear Group Coding(LLGC):** The T groups that most correlate with the input signal y are chosen to be the visual base for input y.

The *correlation* between the input signal (residue) y and dictionary group (sub-dictionary $D^{(i)}$) is defined as the square of ℓ_2 norm of cosine similarity of y and every atom $D_j^{(i)}$ in $D^{(i)}$. Our encoding methods are efficient as the solution of the small-scale Lasso problem is very efficient.

V. Dictionary Adaptation

The semantic dictionary is learned in a different domain without awareness of the characteristics of image patches in the target classification dataset. Therefore, we introduce a simple domain adaptation method to adapt the learned semantic dictionary to our classification task, by learning a transformation matrix W. The adaptation matrix W is learned by minimizing the overall reconstruction error on image patches from the target dataset,

$$z = \underset{W}{\text{argmin}} \ \frac{1}{2} \sum_{i=1}^{N} ||y_i - WD\beta_i||_2^2 + \lambda\Omega(\beta_i) \qquad (4)$$

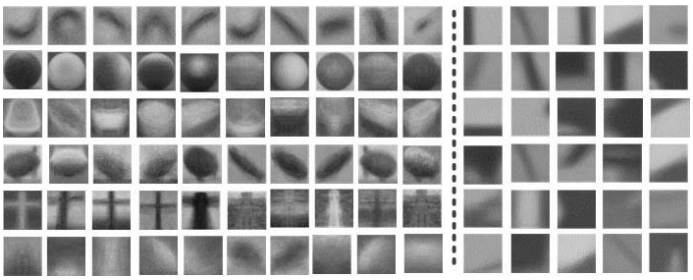


Fig. 2. Comparison of dictionary atoms. The left column presents 6 group atoms of our semantic dictionary, which exhibit semantically similar visual patterns. The right column shows 30 randomly sampled atoms from the dictionary generated through clustering local image patches.

where N is the number of patches; D is the off-learned semantic dictionary and $\Omega(\beta)$ denotes the regularization term with respect to different encoding methods (LLC, GGSC). Since the objective function is not jointly convex to W and β, we learn W using an iterative method: first, W is initialized to be an identity matrix; next, β_i is obtained via applying corresponding encoding solver; finally, given the new β_i, we update the adaptation matrix W by the gradient ΔW of the objective function (Equation 4) to W:

$$\Delta W = -\sum_{i=1}^{N} (y_i - WD\beta_i)\beta_i^T D^T \qquad (5)$$

We iterate the update procedure until it converges.

VI. Experiments and Results

Dataset: We evaluate our semantic dictionary and encoding methods on three scene classification datasets, including generic natural scene images(15-Scene), cluttered indoor images (MIT 67 Indoor Scenes), and complex event and activity images (UIUC-Sports). The classification performance is the mean of multi-way classification accuracy scores over all the classes in the dataset.

Experiment Setup: We densely sample image patches following [23], the size of which is 40×40 and the sliding stride is 8 pixels. Meanwhile, only SIFT [24] feature is used to represent the image patches [1]. All images are resized to maximum 300 pixels along the smaller axis. Finally, The image-level representation is generated by performing max pooling [2], [1], [17] operation to all the sparse codes of image patches. Besides, Spatial Pyramid Matching (SPM) [23] is also applied to enforce the spatial layout constraint.

A. Superiority of Semantic Dictionary

We verify the effectiveness of our semantic dictionary in three datasets, and the results are inspiring: the state-of-the-art coding method (LLC [2]) can benefit from our semantic visual dictionary. Table I demonstrates that LLC achieves better result on our semantic dictionary than on the dictionary generated

[1]building richer feature for image patches and using multi-scale patches can significantly improve the performance [25], but that is not our main focus.

Methods	15-Scene	UIUC-Sports	Indoor
GIST[27]	-	-	26%
DPM[28]	-	-	30.4%
Patches[29]	-	-	38%
Object Bank[17]	80.9%	76.3%	37.6%
Topic Model[30]	78%	82.5%	-
Kwitt[31]	82.3%	83%	-
LLC(RP+KSVD)[32]	$74.9\% \pm 0.48\%$	$77.5\% \pm 1.73\%$	29.2%
LLC(RP+Kmeans)[2]	$80.6\% \pm 0.72\%$	$85.2\% \pm 1.53\%$	41.2%
LLC(Semantic)	$81.0\% \pm 0.37\%$	$85.1\% \pm 1.42\%$	42.0%
LLC(ASemantic)	$81.1\% \pm 0.52\%$	$85.8\% \pm 1.57\%$	43.2%
LLC(RP+ASemantic)	$\mathbf{82.0\% \pm 0.49\%}$	$\mathbf{86.5\% \pm 1.26\%}$	**45.8%**
GGSC(Semantic)	$82.0\% \pm 0.66\%$	$85.6\% \pm 2.07\%$	43.2%
LLGC(Semantic)	$81.2\% \pm 0.42\%$	$84.4\% \pm 2.50\%$	42.0%

TABLE I
CLASSIFICATION PERFORMANCE ON THREE SCENE CLASSIFICATION DATASETS. 'RP' AND 'SEMANTIC' DENOTE DICTIONARY GENERATED FROM RANDOM PATCHES AND SEMANTIC THUMBNAIL IMAGES RESPECTIVELY, 'ASEMANTIC' REPRESENTS THE ADAPTED SEMANTIC DICTIONARY.

through clustering random image patches. The largest performance gain is observed when these representation codes are fused, demonstrating that our semantic and random patch (conventional) dictionaries are complementary.

We further investigate the performance of different encoding methods. SLEP [26] toolbox is utilized to solve the group Lasso (GLasso) and sparse group Lasso (Sparse GLasso) problem. For GGSC and LLGC, at most $T = 5$ groups are activated to generate the sparse codes. The classification accuracy and average elapsed time for every image are reported in Table II. We notice that both our two encoding alternatives achieve significant better performance than GLasso and Sparse GLasso, while in the mean time ours are much more efficient. Table I shows that GGSC also outperforms LLC, which elucidates the advantage of considering the group structure of our semantic dictionary. However, we fail to see obvious improvement over GGSC on the adapted semantic dictionary, as the adaptation may have broken the group structure of the semantic dictionary.

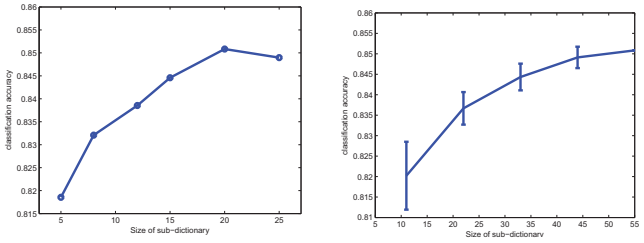
B. Discriminative power over size of sub-dictionary

We believe that each object group has a different optimal size for corresponding sub-dictionary, as the intra-class semantic variance and number of object examples in each object class varies dramatically. For example, the object semantic *"ball"* includes "basketball", "soccer", "golf ball", "bowling ball", "tennis ball" and "pooling ball" in our case, which exhibit large intra-class variance; while the object class *"hook"* has a much smaller intra-class semantic variance and very few training instances. We simply fix the size of all the sub-dictionaries to be the same ($K = |D^{(i)}|$) in our implementation.

Then we analyze how K affects the discriminative power of our semantic dictionary. On one hand, the sub-dictionary with larger size encodes finer-grained visual patterns, while on the other hand, the dimensionality of the sparse codes grows linearly with the size of sub-dictionary. As demonstrated in Figure 3, K is fixed to be 20 in our experiments as it is a good compromise between the dictionary size and discriminative power.

Methods	accuracy	time (Seconds/image)
GLasso	$82.60\% \pm 1.33\%$	19.69
Sparse GLasso	$83.62\% \pm 1.13\%$	19.28
GGSC	$\mathbf{85.60\% \pm 2.07\%}$	1.99
LLGC	$84.35\% \pm 2.50\%$	0.73

TABLE II
CLASSIFICATION ACCURACY AND TIME CONSUMPTION COMPARISON ON UIUC-SPORTS.



(a) (b)

Fig. 3. (a) classification accuracy w.r.t size of sub-dictionary. X-axis denotes the size of sub-dictionary. (b) classification performance w.r.t number of object groups that constitute the semantic dictionary (each group is composed of 20 atoms), X-axis is the number of groups. 5 rounds of randomized sampling is performed to choose the groups from 55 object groups. Statistics are based on the experiments on 8-UIUC-Sports.

C. Profitability over growing number of object groups

We envisage that the larger-size dictionary with more object semantics is going to preserve richer information, therefore may serve as a better dictionary. In this part, we verify this assumption. To simulate what will happen with growing number of groups, we randomly sample $P = 11(20\%)$, $22(40\%)$, $33(60\%)$, $44(80\%)$ and $55(100\%)$ out of $M = 55$ object classes for multiple times. The subset dictionary is generated through concatenating the corresponding P sub-dictionaries, therefore $D^{sub} = \texttt{Union}(D^{(1)} \dots D^{(P)})$, note that $D^{sub} \subseteq D$. As shown in Figure 3, the classification performance of larger-size dictionary continuously increases as more object semantics are encoded in the dictionary. Therefore, we conclude that the discriminative power of our semantic dictionary could be further boosted by adding more complementary object semantics.

VII. CONCLUSION

In this paper, we propose a new perspective to learn semantic dictionary from thumbnail images. Next, we exploit the group structure of the dictionary and develop a greedy group sparse coding algorithm to efficiently solve the original NP ℓ_0/ℓ_1 optimization problem. Finally, to make our semantic dictionary better fit the target classification task, a domain-aware adaptation algorithm is developed. The promising results on three benchmarks demonstrated the effectiveness of our proposed methods.

REFERENCES

[1] J. Yang, K. Yu, Y. Gong, and T. Huang, "Linear spatial pyramid matching using sparse coding for image classification," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 2009, pp. 1794–1801.

[2] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong, "Locality-constrained linear coding for image classification," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 3360–3367.

[3] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman, "Discriminative learned dictionaries for local image analysis," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*. IEEE, 2008, pp. 1–8.

[4] O. Boiman, E. Shechtman, and M. Irani, "In defense of nearest-neighbor based image classification," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*. IEEE, 2008, pp. 1–8.

[5] L.-J. Li and L. Fei-Fei, "What, where and who? classifying events by scene and object recognition," in *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*. IEEE, 2007, pp. 1–8.

[6] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 68, no. 1, pp. 49–67, 2006.

[7] S. Kong and D. Wang, "A dictionary learning approach for classification: separating the particularity and the commonality," in *Computer Vision–ECCV 2012*. Springer, 2012, pp. 186–199.

[8] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online dictionary learning for sparse coding," in *Proceedings of the 26th Annual International Conference on Machine Learning*. ACM, 2009, pp. 689–696.

[9] F. Bach, J. Mairal, J. Ponce, and G. Sapiro, "Sparse coding and dictionary learning for image analysis," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR). tutorial, San Francisco*, 2010.

[10] Z. Jiang, Z. Lin, and L. S. Davis, "Learning a discriminative dictionary for sparse coding via label consistent k-svd," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 2011, pp. 1697–1704.

[11] Q. Zhang and B. Li, "Discriminative k-svd for dictionary learning in face recognition," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 2691–2698.

[12] M. Yang, L. Zhang, X. Feng, and D. Zhang, "Fisher discrimination dictionary learning for sparse representation," in *Computer Vision (ICCV), 2011 IEEE International Conference on*. IEEE, 2011, pp. 543–550.

[13] H. V. Nguyen, V. M. Patel, N. M. Nasrabadi, and R. Chellappa, "Kernel dictionary learning," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*. IEEE, 2012, pp. 2021–2024.

[14] Z. Szabó, B. Póczos, and A. Lorincz, "Online group-structured dictionary learning," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 2011, pp. 2865–2872.

[15] H. Wang, C. Yuan, W. Hu, and C. Sun, "Supervised class-specific dictionary learning for sparse modeling in action recognition," *Pattern Recognition*, vol. 45, no. 11, pp. 3902–3911, 2012.

[16] G. Wang, D. Hoiem, and D. Forsyth, "Learning image similarity from flickr groups using fast kernel machines," 2012.

[17] L.-J. Li, H. Su, L. Fei-Fei, and E. P. Xing, "Object bank: A high-level image representation for scene classification & semantic feature sparsification," in *Advances in neural information processing systems*, 2010, pp. 1378–1386.

[18] L. Torresani, M. Szummer, and A. Fitzgibbon, "Efficient object category recognition using classemes," in *Computer Vision–ECCV 2010*. Springer, 2010, pp. 776–789.

[19] J. Vogel and B. Schiele, "Semantic modeling of natural scenes for content-based image retrieval," *International Journal of Computer Vision*, vol. 72, no. 2, pp. 133–157, 2007.

[20] J. Friedman, T. Hastie, and R. Tibshirani, "A note on the group lasso and a sparse group lasso," *arXiv preprint arXiv:1001.0736*, 2010.

[21] E. Elhamifar and R. Vidal, "Robust classification using structured sparse representation," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 2011, pp. 1873–1879.

[22] G. Swirszcz, N. Abe, and A. C. Lozano, "Grouped orthogonal matching pursuit for variable selection and prediction," in *Advances in Neural Information Processing Systems*, 2009, pp. 1150–1158.

[23] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, vol. 2. IEEE, 2006, pp. 2169–2178.

[24] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

[25] M. Juneja, A. Vedaldi, C. Jawahar, and A. Zisserman, "Blocks that shout: Distinctive parts for scene classification," in *Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, vol. 1, 2013.

[26] J. Liu, S. Ji, and J. Ye, *SLEP: Sparse Learning with Efficient Projections*, Arizona State University, 2009. [Online]. Available: http://www.public.asu.edu/ jye02/Software/SLEP

[27] A. Quattoni and A. Torralba, "Recognizing indoor scenes," in *Computer Vision and Pattern Recognition, 2009 IEEE Computer Society Conference on*, 2009.

[28] M. Pandey and S. Lazebnik, "Scene recognition and weakly supervised object localization with deformable part-based models," in *Computer Vision (ICCV), 2011 IEEE International Conference on*. IEEE, 2011, pp. 1307–1314.

[29] S. Singh, A. Gupta, and A. A. Efros, "Unsupervised discovery of mid-level discriminative patches," in *Computer Vision–ECCV 2012*. Springer, 2012, pp. 73–86.

[30] Z. Niu, G. Hua, X. Gao, and Q. Tian, "Context aware topic model for scene recognition," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 2743–2750.

[31] R. Kwitt, N. Vasconcelos, and N. Rasiwasia, "Scene recognition on the semantic manifold," in *Computer Vision–ECCV 2012*. Springer, 2012, pp. 359–372.

[32] M. Aharon, M. Elad, and A. Bruckstein, "K-svd: An algorithm for designing overcomplete dictionaries for sparse representation," *Signal Processing, IEEE Transactions on*, vol. 54, no. 11, pp. 4311–4322, 2006.