# Sparse Classification for Computer Aided Diagnosis Using Learned Dictionaries

Meizhu Liu, Le Lu, Xiaojing Ye, Shipeng Yu, and Marcos Salganicoff

Siemens Medical Solutions USA, Malvern, PA 19355
le-lu@siemens.com
University of Florida, Gainesville, FL 32611
mliu@cise.ufl.edu

**Abstract.** Classification is one of the core problems in computer-aided cancer diagnosis (CAD) via medical image interpretation. High detection sensitivity with reasonably low false positive (FP) rate is essential for any CAD system to be accepted as a valuable or even indispensable tool in radiologists' workflow. In this paper, we propose a novel classification framework based on sparse representation. It first builds an overcomplete dictionary of atoms for each class via $K$-SVD learning, then classification is formulated as sparse coding which can be solved efficiently. This representation naturally generalizes for both binary and multiwise classification problems, and can be used as a standalone classifier or integrated with an existing decision system. Our method is extensively validated in CAD systems for both colorectal polyp and lung nodule detection, using hospital scale, multi-site clinical datasets. The results show that we achieve superior classification performance than existing state-of-the-arts, using support vector machine (SVM) and its variants [1, 2], boosting [3], logistic regression [4], relevance vector machine (RVM) [5,6], or $k$-nearest neighbor (KNN) [7].

## 1 Introduction

Colon cancer and lung cancer are the two leading causes of cancer deaths in western population. However, these two cancers are highly preventable or "curable" if detected early. Image interpretation based cancer detection via 3D computer tomography has emerged as a common clinical practice, and many computer-aided detection tools for enhancing radiologists' diagnostic performance and effectiveness are developed in the last decade [1–4,6–8]. The key for radiologists to accept the clinical usage of a CAD system is highest possible detection sensitivity with reasonably low false positive (FP) rate per case.

CAD system generally contains two stages: *Image Processing* as extracting (sub)volumes of interest (VOI) by heuristic volume parsing, and informative feature attributes describing the underlying (cancerous) anatomic structures; *Classification* as deciding the class assignment (cancer, or non-cancer) for selected VOIs by analyzing features. VOI selection is also called *candidate generation*, or CG to rapidly identify possibly anomalous regions with high sensitivity, but low

G. Fichtinger, A. Martel, and T. Peters (Eds.): MICCAI 2011, Part III, LNCS 6893, pp. 41–48, 2011.

specificity, e.g. more than 100 candidates per scan with one to two true positives. Then dozens or hundreds of heterogeneous image features can be computed per VOI, in domains of volumetric shape, intensity, gradient, texture and even context [1, 3, 4, 7, 8]. Lastly, the essential goal for classification is achieving the best balance between high sensitivities and low false positive rates, given VOIs and associated features.

In this paper, we propose a new sparsity conducted classification framework, namely *dictionary learning as training* and *sparse coding as testing*, for CAD problems. Sparse signal representation has proved to be a very powerful tool for robustly acquiring, representing, and compressing high-dimensional signals that can be accurately constructed from a compact, fixed set of basis. The sparse representation (related to but different from subspace models of principal component analysis, independent component analysis, non-negative matrix decomposition) is effective in pattern recognition problems, and with link to biological evidence in human cortex system [9]. To the best of our knowledge, the present paper is the first reported work of exploiting sparse representation for CAD classification.

Different from the conventional *parametric* supervised classifiers of SVM, RVM, KNN, logistic regression and so on, a *nonparametric vocabulary* as set of exemplary atoms (learned as optimal rank-1 data matrix approximations) is constructed by maximizing its reconstruction power (or minimizing reconstruction error), within each positive or negative class, given the original training dataset. Then the testing or classification of a new data sample is accomplished by solving for the best approximation per vocabulary/class, under various sparsity constraints. The proposed classification method is evaluated on two large scale clinical datasets collected from multiple clinical sites across continents, for two tasks of colon polyp and lung nodule detections. Our datasets are representative, but very challenging with large within-class variations for polyp, nodule class and other anatomical structures in colon and lung volumes. The results validate that this new classification framework can significantly improve the accuracy of our baseline computer-aided detection system, using the same set of input image features, and compare favorably with other state-of-the-arts [1–4, 6–8].

## 2 A Dictionary Approach to Classification

In this section, we present the new sparsity based classification framework for both binary or multiwise classes. The framework is comprised of two steps: dictionary learning and sparse coding. Unless otherwise noted, all vectors in this paper are column vectors. Also, $\|\cdot\|_2$ represents the regular Euclidean norm, and $\|\cdot\|_0$ counts the number of nonzero components of a vector. $(\cdot;\cdot)$ denotes a vector or matrix by stacking the arguments vertically.

### 2.1 Sparse Dictionary Learning

**Problem Formulation**: Suppose that there are $N$ data samples $\{y_i \in \mathbb{R}^n : i = 1, \cdots, N\}$ of dimension $n$, and the collection of these $N$ samples forms an $n$-by-$N$

data matrix $Y = (y_1, \cdots, y_N)$ with each column as one sample vector. Our goal is to construct a representative dictionary for $Y$, in the form of an $n$-by-$K$ matrix $D = (d_1, \cdots, d_K)$, that consists of $K$ (usually $K << N$) key features $\{d_i \in \mathbb{R}^n : i = 1, \cdots, K\}$ extracted from $Y$. In the dictionary context, $d_i$ is also called an atom that represents one prototype feature in the category. This dictionary $D$ needs to be trained from $Y$, and should be capable to sparsely represent *all* the samples that are in the same category as those in $Y$. Here by sparse representation we mean that each $y_i$ can be written as a linear combination of very few atoms in $D$. In other words, we want to find a dictionary $D$ and corresponding coefficient matrix $X = (x_1, \cdots, x_N) \in \mathbb{R}^{K \times N}$ such that $y_i = Dx_i$ and $\|x_i\|_0 << K$ for all $i = 1, \cdots, N$.

The problem can be readily formulated as the following minimizations:

$$\min_{D,X} \sum_{i=1}^{N} \|x_i\|_0, \text{ subject to } \|y_i - Dx_i\|_2 \leq \epsilon, \quad i = 1, \cdots, N, \tag{1}$$

where $\epsilon > 0$ is the prescribed error tolerance of representation error. The solution $(D, X)$ of (1) yields a dictionary $D$ which extracts the main features $\{d_k : k = 1, \cdots, K\}$ from samples in $Y$, and a coefficient matrix $X$ with each column $x_i$ representing the correlations between $y_i$ and the dictionary atoms in $D$, by $\min_{D,X} \|x_i\|_0$.

**Solving $D, X$:** Since the objective function in (1) is highly nonconvex and nonsmooth, the solution is in general nontrivial. However, there are several algorithms that can be used to well approximate the solutions of (1), and numerous numerical tests demonstrated that these algorithms are very effective in practice. In this paper, we use the recently developed $K$-SVD algorithm [10], which has proved to be very robust to solve (1), by iterating exact $K$ times of Singular Value Decomposition (SVD).

Starting from an initial dictionary, $K$-SVD algorithm approaches the solution of (1) by alternating the following two steps: the minimization with respect to $X$ with $D$ fixed, and the update of atoms in $D$ using the current $X$.

The first step is called the "sparse coding" and can be formulated as

$$\min_{x_i} \|x_i\|_0, \text{ subject to } \|y_i - Dx_i\|_2 \leq \epsilon, \quad i = 1, \cdots, N, \tag{2}$$

Although (2) is in general an NP-hard problem, the solution can usually be well approximated by many pursuit algorithms. In this work, we used the default sparse coding solver in $K$-SVD algorithm called the orthogonal matching pursuit (OMP) [11].

The second step is called the "dictionary update" which modifies the atoms in $D$ one by one to better represent the data $Y$. To update $d_k$, the $K$-SVD algorithm first finds the index set $I_k = \{i : x_{ki} \neq 0\}$, which is just the set of indices of $y_i$'s who used $d_k$ in representation in the sparse coding step. Then it applies the singular value decomposition (SVD) of the error matrix

$$E_k = Y_k - D_k X_k \tag{3}$$

where $D_k$ is $D$ with $d_k$ replaced by 0. In (3), $Y_k$ and $X_k$ collect the columns with indices in $I_k$ from $Y$ and $X$, respectively. Finally, $K$-SVD substitutes $d_k$ in $D$ by the principal singular vector from the SVD of $E_k$ and modifies the coefficients accordingly. This optimization is sequentially executed for each $k = 1, \cdots, K$ while keeping all other columns $d_j$ $(j \neq k)$ fixed. Refer [10] for more details.

**Output**: The output of $K$-SVD consists of a trained dictionary $D$ that contains atoms as features extracted from $Y$, and a coefficients $X$ that records the sparse correlation or dependency of each sample $y_i$ to these atoms. This learned dictionary $D$ will be employed as a special form of classifier for CAD classification task. The dictionary $D$ is build from training data $Y$ in a data driven manner, and is capable to sparsely represent the very majority of data samples that are similar to those in $Y$.

### 2.2 Classification Using Learned Dictionaries

Our sparsity based classification framework, including the dictionary learning and classifier building, is essentially *generative*. It is able to handle both binary and multiwise classification problems.

Suppose that the training samples are given in the form of $L$ $(L \geq 2)$ categories, $\{Y^{(l)} \in \mathbb{R}^{n \times N_l} : l = 1, \cdots, L\}$, where $Y^{(l)} = (y_1^{(l)}, \cdots, y_{N_l}^{(l)})$ consists of $N_l$ training samples labeled by $l$. To design a robust classifier, we apply the $K$-SVD algorithm to (1) with $Y = Y^{(l)}$, and obtain the respective dictionary $D^{(l)}$ for each $l = 1, \cdots, L$. Now $D^{(l)}$ consists of the main exemplary atoms or features of the $l$-th category, and all samples belonging to this category can be sparsely represented by $D^{(l)}$. Furthermore, we can construct the global dictionary $D$ by concatenating all $D^{(l)}$ as follows

$$D = (D^{(1)}, D^{(2)}, \cdots, D^{(L)}) \in \mathbb{R}^{n \times N} \tag{4}$$

where $N = \sum_l N_l$. This global dictionary $D$ is used as the classifier in our tests.

In order to determine the label of a new coming sample $y$, we solve the minimization problem

$$\min_x \|x\|_0, \text{ subject to } \|y - Dx\|_2 \leq \epsilon \tag{5}$$

with the global dictionary $D$ in (4) using OMP [11]. Then we can examine the coefficient vector $x = (x^{(1)}; \cdots; x^{(L)})$ solved from (5), and classify $y$ to the $l$-th category if the nonzero components of $x$ are clustered in the $l$-th segment of $x$, i.e. $x^{(l)}$. That is, a label $l$ is assigned to $y$ if the solution $x = (x^{(1)}; \cdots; x^{(L)})$ of (5) satisfies

$$\|x^{(l)}\|_0 = \max\{\|x^{(m)}\|_0 : m = 1, \cdots, L\}. \tag{6}$$

An ambiguous situation may happen if $x^{(l')}$ contains the largest component of $x$, but has less nonzero elements when compared to $x^{(l)}$. In this case, the label assigned to $y$ is $l$ instead of $l'$ according to the criterion (6). However, it is more intuitive and reasonable to assign $y$ by the label $l'$, as the key feature of

$y$ occurs with higher weights in the $l'$-th category or $D^{(l')}$. A remedy of this is to substitutes the $\ell_0$ norm in (6) by the $\ell_1$-norm which can retain the sparsity property and take the magnitudes of the coefficients into account.

An alternative criterion for classification is to solve the per-category objective

$$\min_{z^{(l)}} \|z^{(l)}\|_0, \text{ subject to } \|y - D^{(l)} z^{(l)}\|_2 \leq \epsilon \tag{7}$$

for $l = 1, \cdots, L$ respectively, and obtain the coefficients $z^{(l)} \in \mathbb{R}^{N_l}$ for all $l = 1, \cdots, L$. Then $y$ is classified to the $l$-th category if $y$ appears to be "more" sparse with respect to $D^{(l)}$, namely,

$$\|z^{(l)}\|_0 = \min\{\|z^{(m)}\|_0 : m = 1, \cdots, L\}. \tag{8}$$

This means that $D^{(l)}$ is more capable to extract the key features, or components of $y$ than other dictionaries, indicating that $y$ should be in the category $l$.

It is worth noting that the difference between (5) and (7) leads to distinct criterion (6) and (8). The criterion (6) implies that $y$ is more similar to the contents in $D^{(l)}$ so it prefers $D^{(l)}$ when exposed to all $D = \{D^{(l)}\}$ simultaneously. On the other hand, (8) suggests that $D^{(l)}$ effectively attains the main features of $y$ and is more capable to represent $y$ sparsely when compared to other dictionaries, with the same error tolerance $\epsilon$.

## 3 Experiments

**Data:** Our colon CAD dataset contains 429 patients or 858 CT volumes (i.e., two prone/supine scans per patient), collected from multiple hospitals in the US, Canada, Asia and Europe, and acquired using Siemens, GE and Philips scanners. After the candidate generation process (briefly discussed in Section 1), we obtain 134116 data candidates, out of which 1116 samples are positives belonging to 391 real polyps because one polyp can have multiple instances appeared, and the rests are negatives. Each data sample is represented using a 96 dimensional feature vector, including geometry, shape morphology, intensity and texture cues, computed by our CAD system. Moreover, 411 positive samples are instances of 137 **flat** polyps and 705 positives belong to 254 **non-flat**, or **SP** (e.g., sessile, pedunculated and mass) polyps. Therefore, the dataset can be subdivided as two classes as negatives (-) and positives (+) or three categories, namely negatives (-1), flat polyps (+1) and non-flat polyps (+2). The lung nodule dataset was obtained from 1000 patients from multiple medical sites in different countries using various scanners. This dataset contains the information of part-solid nodules with a diameter range of 4-20mm. There are 49,094 samples after CG stage, out of which 2,531 are positive nodule instances (+) and the rest as negatives (-). Each data sample has 112 features. In the following, our experiments use 5 fold cross-validation and no data samples from the same patient are used for both training and testing.

**Standalone:** For comparison, we first train a baseline classifier using multiple instance relevance vector machine (MILRVM) [5], and its training/testing classification performances, in the form of Free-Response Operator Characteristic

(FROC) curves, are illustrated in Fig. 1 and 2. Note that these baselines achieve comparable results with state-of-the-arts [1–4, 6–8] on datasets of similar data scales). Next, using the positive and negative samples in the training dataset (i.e., $L = 2$), we learned dictionaries $\mathbf{D}_+ \in \mathbb{R}^{n\times K}$ and $\mathbf{D}_- \in \mathbb{R}^{n\times K}$ for (+/-) classes, respectively. The dictionary size $K = \beta \times n$ is normally chosen with respect to data dimension $n$, here $\beta = 4$. When $\beta > 4$, the classification performances are similar. After this, both dictionaries are concatenated into a single dictionary $\mathbf{D} \in \mathbb{R}^{n\times(L*K)}$, and the classification criterion (5) is applied. In colon dataset $n = 96$ and $K = 384$; and $n = 112$, $K = 448$ for lung. The sensitivities are calculated on per-polyp or per-nodule level, consist with multiple instance learning setting [5], and the false positive (FP) rates are reported on a per-patient level (i.e., summing FPs in two volume views) in colon and per-volume level for lung. For $L = 3$ of colon CAD, when an instance is classified either as (+1) for flats or (+2) for SP, an overall "hit" or detection will be counted. The classification results on training and testing (validation) datasets are shown in Table 1. Though sparse classification does not provide FROC curves, it has ($2\% \sim 5\%$) higher detection sensitivities than our MILRVM baselines, at low FP rates of $\leq 3$ per case in both colon and lung datasets. This is highly suitable for clinical applications.

**Table 1.** Standalone Sparsity Classification Results for Colon Polyp (L=3) and Lung Nodule Detection (L=2)

| | Colon Polyp CAD | | | | Lung Nodule CAD | |
|---|---|---|---|---|---|---|
| | FP Rate | Sensitivity | Flat Sensitivity | SP Sensitivity | FP Rate | Sensitivity |
| Training | 2.6818 | 91.12% | 84.97% | 91.79% | 2.6919 | 90.32% |
| Testing | 2.6897 | 89.68% | 79.98% | 94.20% | 2.6797 | 89.65% |

**Gated Fusion:** To build the best overall CAD system, we exploit the *three-way gated decision tree*, integrating both RVMMIL [5] and sparsity classifiers. RVMMIL assigns each data sample a probability value $\rho(+)$, of being positive class. Therefore, we design the following three gates or decision rules: (1) if $\rho(+) \geq \gamma_1$, classifying as positive; (2) if $\rho(+) \leq \gamma_2$, classifying as negative, where $\gamma_1 > \gamma_2$; (3) if $\gamma_1 > \rho(+) > \gamma_2$, employing sparse classification (L=2, or L=3). The thresholds $\gamma_1, \gamma_2$ are estimated by maximizing the decision tree classification accuracy via cross validation. Conditions $\rho(+) \geq \gamma_1$, $\rho(+) \leq \gamma_2$ indicate samples being positive or negative *with high confidence*; while $\gamma_1 > \rho(+) > \gamma_2$ refers *ambiguous* classifying data samples by RVMMIL.

Fig. 1 (Left) shows the combined model achieves $6\% \sim 8\%$, or $2.4\% \sim 3.2\%$ sensitivity improvements for colon polyp detection, in training and testing respectively, at $\sim 2.7$ FPs per patient. On the other hand, at the same sensitivities, our method can reduce the FP rates by $3 \sim 4$ per patient, with respect to training or testing. Note that $L = 3$ performs better than $L = 2$ which shows the advantage of modeling capacity for a more comprehensively generative representation, consisting of richer dictionaries $D = (D^{(1)}, D^{(2)}, \cdots, D^{(L)})$. Similarly, at least

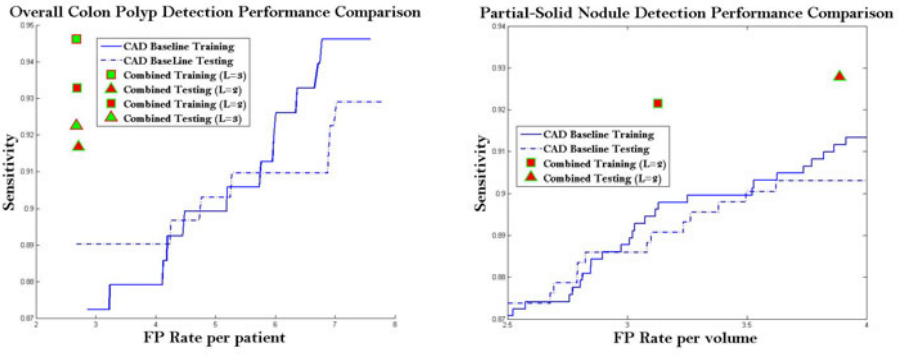


**Fig. 1.** The classification results of using our proposed method and comparison with the CAD baseline, for training and testing in the colon dataset (**Left**) and lung dataset (**Right**). CAD baselines are plotted for comparison.

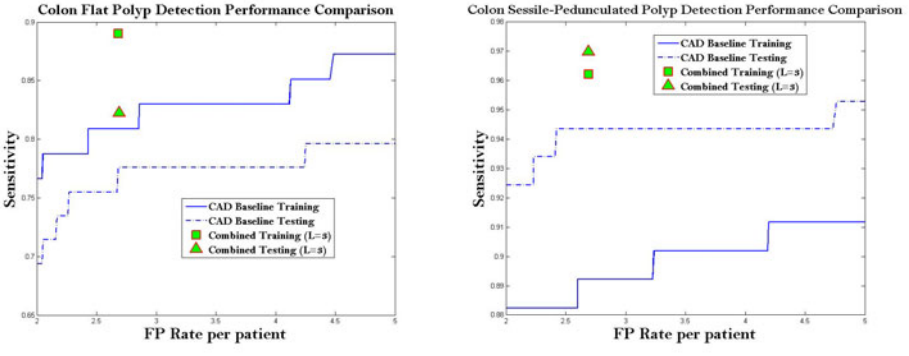


**Fig. 2.** Sensitivity vs. FP rate per patient (i.e., two volumes) for Flat polyp detection (**Left**) in training and testing datasets; and for Sessile-Pedunculated polyp detection (**Right**). CAD baselines are plotted for comparison.

$2\% \sim 3\%$ sensitivity gains are observed for lung nodule detection, corresponding to the same FP rates, in Fig. 1 (Right). These improvements are statistically significant for colon/lung cancer detections, from the already high-performed baselines. Furthermore, the sensitivities for flat or sessile-pedunculated polyps are also greatly increased (by $4\% \sim 7\%$), as shown in Fig. 2 (**Left**) and (**Right**). The three-gate combined model also consistently outperforms single RVMMIL baseline or sparsity classifier, in all scenarios.

## 4 Conclusion and Future Work

In this paper, we present a new sparse representation based classification method for computer-aided diagnosis problem, by learning an overcomplete dictionary of

exemplary atoms for each data class and adapting sparse coding criteria for effective classification. This generative formulation has the ability of modeling two or multiple classes in the same way. It can be used either as a standalone classifier, or integrated with other decision-making scheme(s). Our proposed method is validated in two CAD systems of colorectal polyp and lung nodule detection, using large scale, representative clinical datasets. The results show that we achieve superior performances than our baseline and other existing state-of-the-arts. In future work, we plan to explore how to integrate class discriminative information for dictionary learning [12], and other decision fusion structures of heterogeneous classifiers.

## References

1. Wang, S., Yao, J., Summers, R.: Improved Classifier for Computer-aided Polyp Detection in CT Colonography by Nonlinear Dimensionality Reduction. Medical Physics 35, 1377–1386 (2008)
2. Bogoni, L., Bi, J.: Lung Nodule Detection. ImageCLEF: Experimental Evaluation in Visual Information Retrieval 32, 415–434 (2010)
3. Slabaugh, G., Yang, X., Ye, X., Boyes, R., Beddoe, G.: A Robust and Fast System for CTC Computer-Aided Detection of Colorectal Lesions. Algorithms 3(1), 21–43 (2010)
4. van Ravesteijn, V., van Wijk, C., Vos, F., Truyen, R., Peters, J., Stoker, J., van Vliet, L.: Computer Aided Detection of Polyps in CT Colonography using Logistic Regression. IEEE Trans. on Med. Imag. (2010)
5. Raykar, V., Krishnapuram, B., Bi, J., Dundar, M., Rao, R.: Bayesian Multiple Instance Learning: Automatic Feature Selection and Inductive Transfer. In: ICML, pp. 808–815 (2008)
6. Mang, T., Bogoni, L., Salganicoff, M., Wolf, M.: CT Colonography: Retrospective Evaluation of the Performance of Computer-aided Detection of Colonic Polyps in Tagged and Untagged Preparation. In: ECR (2010)
7. van Ginnekena, B., Schilhama, A., Prokopb, M.: A large-scale evaluation of automatic pulmonary nodule detection in chest CT using local image features and k-nearest-neighbour classification. Med. Image Anal., 670–70 (2009)
8. van Ginneken, B.: Comparing and Combining Algorithms for Computer-aided Detection of Pulmonary Nodules in Computed Tomography Scans: The ANODE09 Study. Med. Image Anal., 707–722 (2010)
9. Vinje, W., Gallant, J.: Sparse coding and decorrelation in primary visual cortex during natural vision. Science, 1273–1276 (2000)
10. Aharon, M., Elad, M., Bruckstein, A., Katz, Y.: K-SVD: An Algorithm for Designing of Overcomplete Dictionaries for Sparse Representation. IEEE Transactions on Signal Processing 54, 4311–4322 (2006)
11. Tropp, J.: Greed is good: Algorithmic results for sparse approximation. IEEE Trans. Inf. Theory 50, 2231–2242 (2004)
12. Mairal, J., Bach, F., Ponce, J., Sapiro, G., Zisserman, A.: Learning discriminative dictionaries for local image analysis. In: Proc. IEEE CVPR (2008)