Label Consistent Multiclass Discriminative Dictionary Learning for MRI Segmentation

Oualid M. Benkarim[1], Petia Radeva[1,2], and Laura Igual[1,2⋆]

[1] Department of Applied Mathematics and Analysis, University of Barcelona, Spain
[2] Computer Vision Center of Barcelona, Spain
obenkabe7@alumnes.ub.edu,petia.ivanova@ub.edu,ligual@ub.edu

Abstract. The automatic segmentation of multiple subcortical structures in brain Magnetic Resonance Images (MRI) still remains a challenging task. In this paper, we address this problem using sparse representation and discriminative dictionary learning, which have shown promising results in compression, image denoising and recently in MRI segmentation. Particularly, we use multiclass dictionaries learned from a set of brain atlases to simultaneously segment multiple subcortical structures. We also impose dictionary atoms to be specialized in one given class using label consistent K-SVD, which can alleviate the bias produced by unbalanced libraries, present when dealing with small structures. The proposed method is compared with other state of the art approaches for the segmentation of the Basal Ganglia of 35 subjects of a public dataset. The promising results of the segmentation method show the efficiency of the multiclass discriminative dictionary learning algorithms in MRI segmentation problems.

Keywords: MRI segmentation, sparse representation, discriminative dictionary learning, multiclass classification

1 Introduction

Many clinical applications rely on the segmentation of MRI brain structures, which allows to describe, for instance, how brain anatomy changes in relation with certain brain diseases. Since manual labeling by experts is subject to inter and intra rater variability and is also a highly laborious task, an automated technique is desirable to enable the routine analysis of brain MRIs in clinical use. Despite the large number of proposed techniques [**?**, **?**, **?**], MRI segmentation still remains a challenging task due to frequent image artifacts and poor contrast between the structures to segment.

Among these techniques, atlas-based methods [**?**] are the most commonly used. They use atlases, which consist of two image volumes: one intensity image and one labeled image, to segment target images without human assistance. The segmentation turns into a registration problem. To obtain a segmentation of the

⋆ This work was partially founded by the projects TIN2012-38187-C03-01 and 2014 SGR 1219.

target image, the manual labeling of the atlas is transformed using the mapping determined during the registration; this process is called label propagation. The main drawback of this kind of techniques is that they implicitly assume that a single atlas endowed with a deformation model is a sufficiently rich representation of the whole population. Segmentation errors produced by atlas-based methods can be reduced by averaging techniques such as multi-atlas based segmentation; using several atlases to better capture the variability of target structures [**?**]. The keypoints of registration-based label propagation approaches concern the accuracy of the non-rigid registration and the fusion rules [**?**]. Recently, non-local patch-based segmentation techniques have been proposed [**?**], whose purpose is to relax the one-to-one constraint existing in non-rigid registration. This technique has two interesting properties: first, the natural redundancy of information contained in the image can be used to increase the numbers of samples considered during estimation; and second, the local intensity context (i.e., patch) can be used to produce a robust comparison of samples. The labeling of every voxel is performed by using similar image patches from coarsely aligned atlases, assigning weights to these patches according to their similarity. The final label is estimated by fusing the labels of the central voxels in the patch library.

Image similarities over small image patches may not be an optimal estimator [**?**]. In [**?**], segmentation is based on image patch reconstruction instead of similarity. A dictionary and a linear classifier are learned from the patch library of every voxel in the target image. Then, the target patch can be reconstructed by the corresponding dictionary and the label of the target voxel is estimated by the corresponding classifier. To the best of our knowledge, [**?**] is the only paper that has previously applied these techniques to subcortical structures segmentation (specifically, the Hippocampus). In this paper, we extend the MRI segmentation method in [**?**]. In particular, the proposed method is a multiclass dictionary leaning approach to simultaneously segment several subcortical brain structures. This method also incorporates a label consistent term [**?**] to impose dictionary atoms to be specialized in one given class. This can alleviate the bias produced by unbalanced patch libraries, which is the case in the boundaries of the brain structures.

The paper is organized as follows. Section 2 is devoted to review Sparse Representation and Dictionary Learning. In Section 3 we cope with the problem of MRI segmentation using these techniques and we introduce our method. Section 4 presents experimental results of our method compared with three state of the art methods. Section 5 finishes with conclusions and future work.

2 Related work

2.1 Sparse representation

Sparse representations have increasingly become recognized as providing extremely high performance for applications as diverse as image denoising [**?**] and image compression [**?**]. The aim of sparse coding is to reconstruct a signal as a linear combination of a small number of signal-atoms picked from a dictionary.

Using a dictionary $D \in \mathbb{R}^{n \times k}$, the representation of a given signal $y \in \mathbb{R}^n$ is $y = D\alpha$.

When the dictionary D is overcomplete, the linear system $y = D\alpha$ is underdetermined since $k > n$, and an infinite number of solutions (if there are any) are available for the representation problem. Hence constraints on the solution must be set. In sparse representation we are interested in the sparsest of all such solutions. As a measure of sparsity, the ℓ^0 norm is used. In general, the sparse coding problem can be formulated as:

$$\min_{\alpha} \|\alpha\|_0 \qquad \text{s. t.} \qquad \|y - D\alpha\|_2^2 \leq \varepsilon, \tag{1}$$

where α is the vector of sparse coefficients of the signal y over D, $\varepsilon > 0$ is a given error tolerance, and $\|\cdot\|_0$ is the ℓ^0 norm.

Since the combinatorial ℓ^0 norm minimization is not convex, the ℓ^1 norm minimization, as the closest convex function to ℓ^0 norm minimization, is widely employed in sparse coding, and it has been shown that both norms are equivalent if the solution is sufficiently sparse. The solution to Eq. (1) is equivalent to the solution of the following problem:

$$\min_{\alpha} \|\alpha\|_1 \qquad \text{s. t.} \qquad \|y - D\alpha\|_2^2 \leq \varepsilon. \tag{2}$$

Using the Lagrangian method, this can be rewritten as:

$$\hat{y} = \min_{\alpha} \frac{1}{2} \|y - D\alpha\|_2^2 + \lambda \|\alpha\|_1, \tag{3}$$

where $\|y - D\alpha\|_2^2$ is the data fitting term, $\|\alpha\|_1$ is the sparsity-inducing regularization, and $\lambda > 0$ is a scalar regularization parameter that balances the trade-off between reconstruction error and sparsity.

Eq. (3) can be solved efficiently by several methods such as Lasso [**?**]. However, if there is a group of variables among which the pairwise correlations are very high, then the Lasso tends to select only one variable from the group and does not care which one is selected. Therefore, it is possible to strengthen further the prediction power of Lasso. The Elastic Net (EN) method, proposed in [**?**], often outperforms Lasso, while enjoying a similar sparsity of representation:

$$\hat{\alpha} = \operatorname*{argmin}_{\alpha} \|y - D\alpha\|_2^2 + \lambda_1 \|\alpha\|_1 + \frac{\lambda_2}{2} \|\alpha\|_2^2. \tag{4}$$

In addition, the EN method encourages a grouping effect where strongly correlated predictors tend to be in or out of the model together. The elastic net is particularly useful when the number of predictors k is much bigger than the number of observations n, which is our case dealing with neuroimages.

2.2 Dictionary Learning

An overcomplete dictionary that leads to sparse representations can either be predefined or designed by adapting its content to fit a given set of signal samples.

Recent publications have shown that learning dictionaries for image representation can significantly improve tasks such as image restoration [**?**]. Concretely, given a set of signals $Y = [y_1, \cdots, y_n]$, we assume that there exists a dictionary D that gave rise to the given signal samples via sparse combinations, i.e., there exists D, so that solving Eq. (1) for each y_i gives a sparse representation α_i. Learning a dictionary with k number of atoms and with a sparsity constraint T is addressed by solving the following problem:

$$\min_{\alpha, D} \|Y - D\alpha\|_2^2 \qquad \text{s. t.} \qquad \|\alpha\|_0 \leq T. \tag{5}$$

A number of practical algorithms have been developed for learning such dictionaries like *method of optimal directions* (MOD) proposed in [**?**] and the K-SVD algorithm [**?**]. Both K-SVD and MOD are iterative approaches designed to minimize Eq. (5) by first performing sparse coding and then updating the dictionary. Other method that scales to large datasets is the online optimization algorithm for dictionary leaning proposed in [**?**].

Nevertheless, K-SVD is not suitable for classification, where the dictionary should be not only representative, but also discriminative. Hence, some supervised dictionary learning approaches incorporate classification error into the objective function to construct a dictionary with discriminative power. Zhang and Li [**?**] developed the D-KSVD algorithm that uses the labels of training data to directly incorporate a linear classifier into the basic K-SVD algorithm and finally unifies the representation power and discriminate ability to train the dictionary and classifier simultaneously. D-KSVD algorithm solves the following problem:

$$\begin{aligned} \langle D, W, \alpha \rangle = \arg\min_{D,W,\alpha} \|Y - D\alpha\|_2 + \beta\|H - W\alpha\|_2 + \eta\|W\|_2 \\ \text{s. t. } \|\alpha\|_0 \leq T, \end{aligned} \tag{6}$$

where W are the classifier parameters. Each column of H is a vector $h_i = [0, \cdots, 1, \cdots, 0]$, where the non-zero position indicates the class. So the term involving H is the classification error and $\|W\|_2$ is the regularization penalty.

Moreover, approaches such as D-KSVD consider only discriminativeness in the classifier construction, but do not guarantee the discriminativeness in the sparse representations of signals. Jiang et. al in [**?**], proposed the Label Consistent K-SVD (LC-KSVD) algorithm, which associates label information with each dictionary atom to enforce discriminability in sparse codes during the dictionary learning process. LC-KSVD solves the following problem:

$$\begin{aligned} \langle D, W, A, \alpha \rangle = \arg\min_{D,W,A,\alpha} \|Y - D\alpha\|_2 + \beta\|H - W\alpha\|_2 \\ + \lambda\|Q - A\alpha\|_2 + \eta\|W\|_2 \quad \text{s. t. } \|\alpha\|_0 \leq T, \end{aligned} \tag{7}$$

where $\|Q - A\alpha\|_2$ is the label consistent regularization term responsible for enforcing the creation of discriminative dictionaries, A is a linear transformation

matrix and $Q = [q_1, \cdots, q_N] \in \mathbb{R}^{k \times N}$ are the discriminative sparse codes of elements in Y for classification. According to [**?**], for instance, $q_i = [q_i^1, \cdots, q_i^k]^t = [0, \cdots, 1, 1, \cdots, 0]^t \in \mathbb{R}^K$ is a discriminative sparse code corresponding to a given signal $y_i \in Y$, if the non-zero values of q_i occur at those indexes where the y_i and the dictionary atom d_k share the same label.

3 Multiclass Dictionary Learning for MRI segmentation

In this section, we first review the MRI segmentation framework using Sparse Representation Classification (SRC) and Discriminative Dictionary Learning for Segmentation (DDLS) presented in [**?**]. Then, we introduce the Label Consistent Multiclass DDLS (LC-MDDLS) method for MRI segmentation, which is based on DDLS and also incorporates the *label consistency* (LC) property proposed in [**?**].

For a given target image I, we consider a training set of images previously registered to a normalized space. We select the N most similar training images based on the sum of squared intensity differences. For the segmentation of a set of subcortical structures in I, we extract a crop of the image, I_C defined by the dimensions of the union of the voxels belonging to these structures in the training images. The target voxels to segment are the ones in I_C. We define a *patch* as a bounding-box of a given size, S_p, around a target voxel. We create a patch library, P_L, from the set of N training images. As shown in figure **??**, we extract a patch for each voxel in a search window, of a previously defined size, S_w, from all training images. Subsequently, we use P_L to classify the target voxel accordingly to one of the methods presented next.

3.1 Sparse Representation based Classification

In SRC, the whole patch library is directly used as the dictionary in Eq. (4). The reconstruction error, r_j, using the coefficients α^j associated to class j is defined as

$$r_j(p_t) = \|p_t - P_L^j \hat{\alpha}^j\|. \tag{8}$$

Thereafter, the label value v_t for the target patch p_t is assigned as the class with the minimum reconstruction error over all classes:

$$v_t = \underset{j}{\operatorname{argmin}}(r_j(p_t)), \ \ \forall j = 1, \ldots, C, \tag{9}$$

where $C = 2$ is the number of classes (subcortical structure or background).

3.2 Discriminative Dictionary Learning for Segmentation

Using all training patches as the dictionary might incorporate noisy information and make the sparse coding process much more time-consuming. In contrast, DDLS learns a compact task-specific dictionary and a classifier for each target

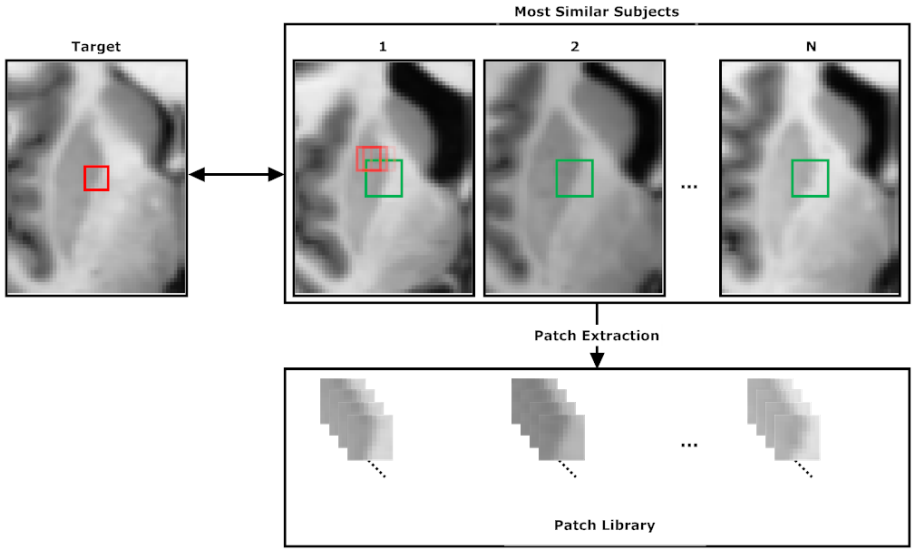

Fig. 1: Creation of the patch library for a given target patch. The red box represents the patch size and the green box corresponds to search window size.

voxel. In our case, the D-KSVD (Eq. (6)) is used. In particular, the D-KSVD algorithm uses K-SVD to find the globally optimal solution for all the parameters simultaneously, rewriting Eq. (6) as follows

$$\langle D, W, \alpha \rangle = \arg\min_{D,W,\alpha} \| \begin{pmatrix} P_L \\ \sqrt{\beta}\, H \end{pmatrix} - \begin{pmatrix} D \\ \sqrt{\beta}\, W \end{pmatrix} \alpha \|_2 + \ \eta \|W\|_2 \qquad (10)$$
$$\text{s. t. } \|\alpha\|_0 \ \leq T.$$

For labeling, since the dictionary is small enough, the sparse representation $\hat{\alpha}_t$ of the target patch is computed using Lasso. The class label vector h_t for the target voxel is estimated by the learned classifier as follows

$$h_t = \hat{W}_t \hat{\alpha}_t. \qquad (11)$$

The index of the largest element in h_t is assigned as the label of the target voxel:

$$v_t = \operatorname*{argmax}_j \ h_t(j). \qquad (12)$$

3.3 Label Consistent Multiclass Discriminative Dictionary Learning for Segmentation (LC-MDDLS)

In LC-MDDLS, we use a multiclass approach to learn a classifier for all structures simultaneously. Consequently, H will have as many rows as structures to

be segmented (including the background). In this way, learned classifiers will be richer than the ones learned with DDLS. Furthermore, DDLS lacks the ability to handle unbalanced libraries (i.e., more patches from one class than another). This problem is important in our case and highly recurrent in voxels near structure boundaries. Consider a target voxel in a structure boundary, due to inter-variability of atlases, extracted patches to create the library might not correspond to the target structure, therefore the patch library might have a higher number of patches belonging to other structures (or background) than the correct one. This imbalance will be transmitted to the learning process producing classifiers with poorer performance. Consider, for instance, a library of 100 patches, where only 5 patches belong to a specific class. Traditional dictionary learning can achieve a good overall reconstruction error without accurately representing these 5 patches; however, LC-MDDLS enforces the representation of all patches in the library as it uses discriminative sparse codes.

LC-MDDLS learns a single discriminative dictionary and a multiclass linear classifier simultaneously for each target voxel. Thence, learned dictionaries will have good representational power, and enforce better discrimination capabilities. Eq. (7) is used as the objective function and, as proposed in [**?**] D, W, and A are initialized before solving Eq. (7): We use K-SVD to learn an intermediate dictionary D_j for each group of patches in the patch library whose class is j. Then all intermediate dictionaries are combined to initialize D. We assign a label j to each atom in D based on the intermediate dictionary D_j it corresponds to and will remain fixed. On the other hand, W and A are initialized using multivariate ridge regression. The dictionary learning process here is similar to the one used in DDLS, although we need to add the label consistent term into the equation:

$$\langle D, W, \alpha \rangle = \arg\min_{D,W,\alpha} \| \begin{pmatrix} P_L \\ \sqrt{\beta}\, H \\ \sqrt{\lambda}\, Q \end{pmatrix} - \begin{pmatrix} D \\ \sqrt{\beta}\, W \\ \sqrt{\lambda}\, A \end{pmatrix} \alpha \|_2 \quad \text{s. t. } \|\alpha\|_0 \leq T. \quad (13)$$

After dictionary learning, the labeling procedure remains the same as the one used in DDLS (Eq. (**??**) and Eq. (**??**)).

4 Experiments

This section is devoted to present the experiments whose objective is to segment the Basal Ganglia, composed of Accumbens, Caudate, Pallidum and Putamen structures.

Dataset. The dataset consists of 35 control subjects and their corresponding segmentations, made public by the MICCAI 2012 challenge[3]. They were all right handed and include 13 males and 22 females. Their ages ranged from 19 to 90

[3] `https://masi.vuse.vanderbilt.edu/workshop2012`

with an average of 32.4 years old. Dataset images consist of a de-faced T1-weighted structural MRI dataset and associated manually labeled volume with one label per voxel.

Evaluation measure. All the experiments were evaluated by computing the Dice coefficient between a reference A and an automated segmentation B:

$$\kappa(A, B) = \frac{2|A \cap B|}{|A| + |B|} \tag{14}$$

Experimental settings. For comparison, we consider as baseline methods: (1) the atlas-driven subcortical segmentation method included within the Freesurfer Software Suite[4], (2) SRC and (3) DDLS[5].

To speed up the dictionary learning phase, we have used the same sampling strategy proposed in [**?**]. Instead of creating a dictionary for each target voxel, dictionaries are created each $n = 3$ voxels. Segmentation of target voxels with no learned dictionary is performed using the dictionaries of the 6 nearest voxels for which we have computed the dictionaries.

Regarding the parameters, we have set $N = 10$, $K = 100$, $S_p = 5 \times 5 \times 5$ and $S_w = 3 \times 3 \times 3$. Using larger patch size or search window size for the whole Basal Ganglia is computationally expensive. Finally, a leave-one-out procedure was used in our validation strategy.

Computational time. Experiments were carried out using a four core Intel Core i7-2630QM processor at 2.0 GHz with 4 GB of RAM. To segment the Basal Ganglia using SRC, DDLS and LC-MDDLS took around 18, 24, 17 minutes per subject, respectively (excluding the learning step in DDLS and LC-MDDLS).

Results. Table **??** contains the obtained average Dice overlaps for each of the Basal Ganglia structures and the whole Basal Ganglia (last column). As it can be seen, LC-MDDLS outperforms the rest of the methods in all the structures, specially in the Accumbens, being statistically significant with 10% significance level.

SRC and DDLS were used to separately segment each structure of the Basal Ganglia as proposed in [**?**]. For this reason, they present several important issues with respect to LC-MDDLS, as illustrated in figure **??**:

1. SRC produces holes: voxels that lie sufficiently far from the boundary and, thus, clearly belong to the structure at hand are labeled as background. However, DDLS and LC-MDDLS are more robust against this problem because

[4] FreeSurfer Software Suite is an open source package for processing and analyzing (human) brain MRI images developed at the Martinos Center for Biomedical Imaging by the Laboratory for Computational Neuroimaging.

[5] SPAMS optimization toolbox (`http://spams-devel.gforge.inria.fr`) [**?**] was used in the learning step.

	Caudate	Accumbens	Pallidum	Putamen	Basal Ganglia
FS	0.82	0.552	0.741	0.786	0.725
SRC	0.869	0.758	0.828	0.876	0.833
DDLS	0.865	0.744	0.855	0.901	0.841
LC-MDDLS	0.873	0.764	0.866	0.906	0.852

Table 1: Average Dice overlaps for Basal Ganglia structures.

of the intermediate dictionary learning process, where noisy information contained in the library is discarded.

2. Under-segmentations: SRC and DDLS segmentation results are, most of the time, smaller than they should be. This indicates that SRC and DDLS are not that accurate in boundaries, where intensities do change. This problem is also present in LC-MDDLS results, although segmentations are quite better.

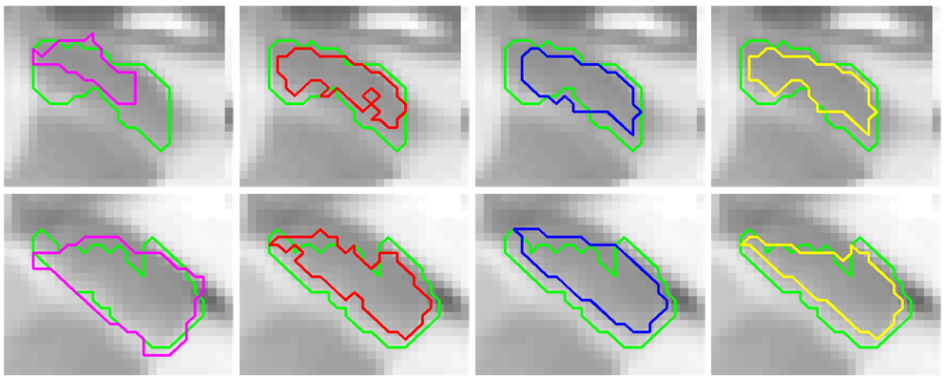

Fig. 2: Segmentation results of Accumbens (top) and Caudate (bottom) structures using (from left to right) FreeSurfer, SRC, DDLS and LC-MDDLS. Ground-truth segmentations are green.

5 Conclusions and future work

In this paper, we presented the LC-MDDLS method to perform segmentation of brain MRI subcortical structures. LC-MDDLS extends DDLS to segment multiple structures at the same time and also exploits discriminativeness in sparse codes in order to achieve dictionary atoms specialized in one given class, which can smooth the impact of unbalanced libraries. The evaluation on Basal Ganglia structures segmentation of a public dataset demonstrates the good accuracy and robustness of these methods. Particularly, LC-MDDLS provided the highest overlap compared with FreeSurfer, SRC and DDLS methods.

Although accurate segmentation results were achieved by LC-MDDLS, there are several aspects that may improve its performance. Future research might focus on adapting the dictionary learning procedure to use a weight matrix within the reconstruction loss term to balance the importance of the different classes. Another improvement can be the application of hierarchical dictionary learning techniques to model dependencies between dictionary elements.

References

1. K. Babalola, B. Patenaude, P. Aljabar, J. Schnabel, D. Kennedy, W. Crum, S. Smith, T. Cootes, M. Jenkinson, and D. Rueckert. *An evaluation of four automatic methods of segmenting the subcortical structures in the brain.* Neuroimage 47 (4), 2009.
2. P. Aljabar, R. Heckemann, A. Hammers, J. Hajnal, and D. Rueckert. *Multi-atlas based segmentation of brain images: Atlas selection and its effect on accuracy.* NeuroImage 46 (3), pp. 726-738, 2009.
3. B. Scherrer, F. Forbes, C. Garbay, and M. Dojat. *Fully bayesian joint model for MR brain scan tissue, structure segmentation.* MICCAI 2008, pp. 1066-1074, 2008.
4. R. Wolz, P. Aljabar, D. Rueckert, R. Heckemann, and A. Hammers. *Segmentation of subcortical structures and the hippocampus in brain mri using graph-cuts and subject-specific a-priori information.* IEEE International Symposium on Biomedical Imaging: From Nano to Macro (ISBI). pp. 470-473, 2009.
5. F. Rousseau, P. Habas, and C. Studholme. *A supervised patch-based approach for human brain labeling.* IEEE Trans. on MI, 30 (10), 2011.
6. P. Coupé, J. Manjón, V. Fonov, J. Pruessner, M. Robles, and D. Collins. *Patch-based segmentation using expert priors: Application to hippocampus and ventricle segmentation.* Neuroimage 54 (2), pp. 940-954, 2011.
7. H. Wang, and P. Yushkevich. *Dependency prior for multi-atlas label fusion.* ISBI: From Nano to Macro (ISBI), pp. 892-895, 2012.
8. T. Tong, R. Wolz, P. Coupé, J. V. Hajnal, and D. Rueckert. *Segmentation of MR images via discriminative dictionary learning and sparse coding: Application to hippocampus labeling.* NeuroImage 76, pp. 11-23, 2013.
9. M. Elad and M. Aharon. *Image denoising via sparse and redundant representations over learned dictionaries.* IEEE Trans. on IP, 15 (12), 2006.
10. O. Bryt, and M. Elad. *Compression of facial images using the K-SVD algorithm.* IEEE Trans. on IP, 19 (4), 2008.
11. R. Tibshirani. *Regression shrinkage and selection via the lasso.* Journal of the Royal Statistical Society. Series B, pp. 267-288, 1996.
12. H. Zou, and T Hastie. Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B 67 (2), pp. 301-320, 2005.
13. K. Engan, S. O. Aase, and J. H. Husoy. *Frame based signal compression using method of optimal directions (MOD).* IEEE Intern. Symp. Circ. Syst., 1999.
14. M. Aharon, M. Elad, and A. M. Bruckstein. *The K-SVD: An algorithm for designing of overcomplete dictionaries for sparse representations.* IEEE Trans. SP, 54 (11), 2006
15. Q. Zhang, and B. Li. *Discriminative K-SVD for dictionary learning in face recognition.* CVPR, pp. 2691-2698, 2010.
16. Z. Jiang, Z. Lin, and L. Davis. *Learning a discriminative dictionary for sparse coding via label consistent k-svd.* CVPR, pp. 1697-1704, 2011.
17. J. Mairal, F. Bach, J. Ponce, and G. Sapiro. *Online Dictionary Learning for Sparse Coding.* International Conference on Machine Learning, Montreal, Canada, 2009.