Supervised Block Sparse Dictionary Learning for Simultaneous Clustering and Classification in Computational Anatomy

Erdem Varol and Christos Davatzikos

University of Pennsylvania, USA

Abstract. An important prerequisite for computational neuroanatomy is the spatial normalization of the data. Despite its importance for the success of the subsequent statistical analysis, image alignment is dealt with from the perspective of image matching, while its influence on the group analysis is neglected. The choice of the template, the registration algorithm as well as the registration parameters, all confound group differences and impact the outcome of the analysis. In order to limit their influence, we perform multiple registrations by varying these parameters, resulting in multiple instances for each sample. In order to harness the high dimensionality of the data and emphasize the group differences, we propose a supervised dimensionality reduction technique that takes into account the organization of the data. This is achieved by solving a supervised dictionary learning problem for block-sparse signals. Structured sparsity allows the grouping of instances across different independent samples, while label supervision allows for discriminative dictionaries. The block structure of dictionaries allows constructing multiple classifiers that treat each dictionary block as a basis of a subspace that spans a separate band of information. We formulate this problem as a convex optimization problem with a geometric programming (GP) component. Promising results that demonstrate the potential of the proposed approach are shown for an MR image dataset of Autism subjects.

1 Introduction

Computational Anatomy (CA) employs statistical methods in order to analyze and model anatomical structures across individuals. Typical CA approaches include Voxel Based Analysis (VBA) [2] and high dimensional pattern-classification [7]. These are complementary techniques and suffer from different limitations.

On the one hand, VBA employs mass univariate linear statistical tests on voxel values in order to identify regional individual differences. The simplicity of the statistical models limits its ability to capture multivariate relationships. On the other hand, high dimensional pattern-classification is able to recover multivariate relationships that characterize group differences while accurately classifying individuals. Nonetheless, it requires a dimensionality reduction step in order to cope with the challenges due to the high dimensional small sample size data that are typical in medical imaging.

P. Golland et al. (Eds.): MICCAI 2014, Part II, LNCS 8674, pp. 446–453, 2014.

A common assumption behind all CA techniques is that the data are optimally brought in correspondence through a registration process. However, the optimality of the spatial normalization is evaluated through measures that usually reflect the intensity agreement of the voxels. While these criteria are relevant in the case of image matching, they are potentially insufficient, or even irrelevant, in the case of group analysis. As a consequence, the choice of registration parameters (*i.e.*, regularization weight) may act as confounding factor for the subsequent statistical analysis.

Motivated by this observation, we propose a novel approach for computational anatomy that is insensitive to the pre-processing step (*i.e.*, registration). In order to limit the influence of registration parameters, we expand the space where the statistical analysis takes place by performing multiple registrations for varying degrees of regularization. In this space, group analysis is performed by simultaneous clustering and classification. The proposed approach is formulated as supervised dictionary learning [9, 3] endowed with a block-sparsity inducing norm [5, 6] in order to harness the underlying structure of the data and enhance the discriminative power of the estimated elements.

The remainder of this paper is divided into 3 sections. Section 2 details the proposed method. Section 3 presents empirical results obtained using an Autism Spectrum Disorder dataset. Section 4 provides a brief discussion of the implications of this work and possible applications in the context of computational anatomy.

2 Method

Motivation. Suppose that the data set comprises of multiple instances per sample where each instance potentially conveys a new band of information. Then our goal is to accurately model the entire dataset in the span of a dictionary such that different blocks of the dictionary capture the information contained in these bands. Furthermore, we want the dictionary, and specifically, each band of the dictionary to be discriminative with respect to the observed labels. The visualization of this goal is illustrated in figure 1.

Supervised Block Sparse Dictionary Learning. To introduce the formulation, first we define the variables. Let $\mathbf{X} \in \mathbb{R}^{d \times n}$ denote the spatially normalized data stored in a tall matrix. Let $\mathbf{y} \in \mathbb{R}^{n}$ denote the labels. The aim is to find an appropriate lower dimensional representation $\mathbf{C} \in \mathbb{R}^{p \times n}$ where each row corresponds to loading coefficients for the bases of the lower dimensional space. The bases can also be called the atoms of a dictionary $\mathbf{D} \in \mathbb{R}^{d \times p}$ that we aim to learn. Furthermore, we want the dictionary to have sub-blocks $\mathbf{D} = [\mathbf{D}_1 | \ldots | \mathbf{D}_{n_b}]$ such that the corresponding loading coefficients can discriminate the labels $\mathbf{y}$ using a different discriminative model $\{\mathbf{w}_1, \ldots, \mathbf{w}_{n_b}\}$ for each block.

In a graphical sense, the problem of supervised dictionary learning can be described as minimizing a joint energy with 5 variables 1) $\mathbf{D}$: the dictionary, 2) $\mathbf{C}$: the loading coefficients, 3) $\mathbf{W}$: the discriminative model 4) $\mathbf{X}$: observed signal

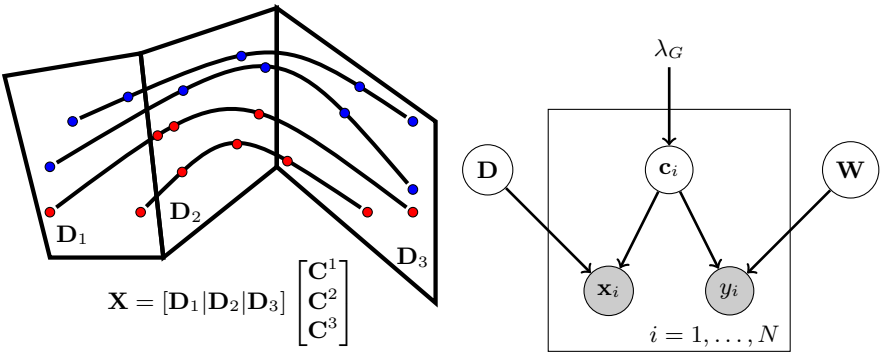

Fig. 1. Left: A graphical depiction of subspace clustering for multiple instance data. Black curves indicate the domain of instances for an independent subject and red points indicate the sampled instances. Dictionary blocks $\mathbf{D}_1, \mathbf{D}_2, \mathbf{D}_3$ indicate the basis of the subspaces where clusters of data reside and where separate classifiers discriminate. **Right:** The graphical model of supervised block sparse dictionary learning. Gray denotes observed variable. $\mathbf{D}$ is the dictionary, $\mathbf{c}_i$ are loading coefficients, $\mathbf{W}$ are the discriminative parameters, $(\mathbf{x}_i, y_i)$ is the sample/label pair, λ_G is the hyperparameter that controls block sparsity

and 5) $\mathbf{y}$: observed labels. The dependencies of these variables is the following: $\mathbf{D}$ and $\mathbf{C}$ generate $\mathbf{X}$ and $\mathbf{C}$ and $\mathbf{W}$ discriminate $\mathbf{y}$. The graphical model can be seen in figure 1.

While the graphical model provides the dependency structure between the pertinent random variables, it is necessary to define a measure of goodness of fit for potential choices of these random variables in order to formulate an optimization program to solve for the hidden variables $\mathbf{D}, \mathbf{C}, \mathbf{W}$. If we let $\mathbf{c}_i^\ell$ denote the ℓth block of the loading coefficients for sample i, we can define the joint energy of these variables as the following term $\mathcal{E}$:

$$\mathcal{E}(\mathbf{x}_i, \mathbf{c}_i, \mathbf{D}, \mathbf{W}, \mathbf{b}, y_i) \triangleq \|\mathbf{x}_i - \mathbf{D}\mathbf{c}_i\|_2^2 + \sum_{\ell=1}^{n_b} \|\mathbf{c}_i^\ell\|_2 (\lambda_G + \lambda_D \epsilon_{i,\ell}) \tag{1}$$

The terms above can be as explained as such:

- $\|\mathbf{x}_i - \mathbf{D}\mathbf{c}_i\|_2^2$ is the data reconstruction error.
- $\epsilon_{i,\ell} = \max\{0, 1 - y_i(\mathbf{w}_\ell^T \mathbf{c}_i^\ell + b_\ell)\}$ is the place holder for the hinge loss function
- $\sum_{\ell=1}^{n_b} \|\mathbf{c}_i^\ell\|_2 (\lambda_G + \lambda_D \epsilon_{i,\ell})$ is a joint term for enforcing block sparsity and minimizing classification loss
- λ_G and λ_D are generative and discriminative penalty parameters, respectively.

The main novelty in our work is the term $\sum_{\ell=1}^{n_b} \|\mathbf{c}_i^\ell\|_2 (\lambda_G + \lambda_D \epsilon_{i,\ell})$. This term diverges from the typical sparse dictionary learning regularization term $\lambda \|\mathbf{c}_i\|_1$ in 3 ways:

1. $\sum_{\ell=1}^{n_b} \|\mathbf{c}_i^\ell\|_2$ is the mixed $L_{2/1}$ norm of the coefficients broken up into n_b blocks. This term enforces sparsity in block selection rather than atom selection which adds additional structure. It implicitly clusters data along subspaces that are spanned by the atoms in each block [5][6].
2. If the basis block ℓ is not used to represent subject i: $\|\mathbf{c}_i^\ell\|_2 \approx 0$ then $\epsilon_{i,\ell}$ is penalized less in the objective. $\epsilon_{i,\ell}$ is the margin violation term for the subspace spanned by dictionary block ℓ. This way the corresponding subspace specific hyperplane $\mathbf{w}_\ell$ is not affected by this sample.
3. If $\epsilon_{i,\ell}$ is large then the block-sparsity of $\mathbf{c}_\ell^i$ is penalized more: objective aims to not represent sample i from the subspace spanned by dictionary block ℓ because it is poorly discriminated there.

Given this energy function for each sample, the overall objective that we aim to minimize for a given dataset is:

$$\begin{aligned} \underset{\mathbf{D},\mathbf{C},\mathbf{W},\mathbf{b},\epsilon}{\text{minimize}} \quad & \|\mathbf{X} - \mathbf{D}\mathbf{C}\|_F^2 + \sum_{i=1}^{n}\sum_{\ell=1}^{n_b} \|\mathbf{c}_i^\ell\|_2(\lambda_G + \lambda_D\epsilon_{i,\ell}) + \lambda_M \sum_{\ell=1}^{n_b} \|\mathbf{w}_\ell\|_2^2 \qquad (2) \\ \text{subject to} \quad & y_i(\mathbf{w}_\ell^T \mathbf{c}_i^\ell + b_\ell) \geq 1 - \epsilon_{i,\ell} \quad \forall i, \ell \\ & \|\mathbf{d}_k\|_2 \leq 1 \quad \forall k \\ & \epsilon_{i,\ell} \geq 0 \quad \forall i, \ell \end{aligned}$$

This is a difficult optimization to perform since like most dictionary learning and clustering tasks; the objective is not convex, specifically with respect to $\mathbf{D}, \mathbf{C}$. In addition, the term $\epsilon_{i,\ell}\|\mathbf{c}_i^\ell\|_2$ constitutes a geometric programming (GP) form. Nevertheless, the formulation is block-convex in $\mathbf{C}$ and $\mathbf{D}, \mathbf{W}, \mathbf{b}, \epsilon$ and we may perform an iterative procedure to obtain a local minimum. The iterative optimization procedure is described in Algorithm 1.

The overview of the algorithm is as follows: The input samples $\mathbf{X}$, labels $\mathbf{Y}$ and hyperparameters that set number of blocks in the dictionary (n_b), the number of atoms per block (n_a) and penalty terms for block sparsity λ_G, discrimination loss penalty λ_D and model complexity λ_M are initially specified. All model parameters $\mathbf{W}$ and $\mathbf{b}$ are set to 0, initially. The initial dictionary $\mathbf{D}$ is set to a random Gaussian matrix with normalized columns. Then iteratively $\{\mathbf{C}\}$ and $\{\mathbf{D}, \mathbf{W}, \mathbf{b}, \epsilon\}$ are optimized. The gradients for each variable are omitted due to space limitations.

Once $\mathbf{D}, \mathbf{W}, \mathbf{b}$ are trained, during test time, an unobserved sample is classified using the following rule:

$$y^\star = \underset{y\in\{-1,+1\}}{\arg\min} \{\min_{\mathbf{c}} \quad \mathcal{E}(\mathbf{x}^\star, \mathbf{c}, \mathbf{D}, \mathbf{W}, \mathbf{b}, y)\} \qquad (5)$$

In words, the predicted label $y^\star$ is one which the minimum energy stated in equation (1) can be obtained after optimizing for the loading coefficients.

Since the energy $\mathcal{E}$ can be interpreted as the negative joint log probability of the graphical model, our framework readily handles the multiple instance data in the following sense:

Algorithm 1. Block sparse supervised dictionary learning (**BS-SDL**)

Input: $\mathbf{X} \in \mathbb{R}^{d \times n}$, $\mathbf{Y} \in \{-1, +1\}^n$ (training signals); n_a (number of atoms per block); n_b (number of blocks); $\lambda_G, \lambda_D, \lambda_M$ (parameters)
Output: $\mathbf{D} \in \mathbb{R}^{d \times (n_b n_a)}$ (dictionary); $\mathbf{w}_1, \ldots, \mathbf{w}_{n_b} \in \mathbb{R}^{n_a}, b_1, \ldots, b_{n_b} \in \mathbb{R}$ (parameters)
Initialization: Set $\mathbf{D}$ to a random Gaussian matrix with normalized columns. Set $\mathbf{w}_1, \ldots, \mathbf{w}_{n_b}, b_1, \ldots, b_{n_b}$ to zero. Set $\epsilon_{i,\ell} = 1$ for $i = 1, \ldots, n \quad \ell = 1, \ldots, n_b$
Loop: Repeat until convergence (or a fixed number of iterations)
• *Supervised block-sparse coding:* Fix $\mathbf{D}, \mathbf{W}, \mathbf{b}, \epsilon$, optimize w.r.t. $\mathbf{C}$

$$\mathbf{C}^\star = \arg\min_{\mathbf{C}} \|\mathbf{X} - \mathbf{D}\mathbf{C}\|_F^2 + \sum_{i=1}^{n} \sum_{\ell=1}^{n_b} \|\mathbf{c}_i^\ell\|_2 (\lambda_G + \lambda_D \epsilon_{i,\ell}) \quad (3)$$
$$\text{subject to} \quad y_i(\mathbf{w}_\ell^T \mathbf{c}_i^\ell + b_\ell) \geq 1 - \epsilon_{i,\ell} \quad \forall i, \ell$$

• *Dictionary and parameters update:* Fix $\mathbf{C}$, optimize w.r.t. $\mathbf{D}, \mathbf{W}, \mathbf{b}, \epsilon$

$$\mathbf{D}^\star, \mathbf{W}^\star, \mathbf{b}^\star, \epsilon^\star = \arg\min_{\mathbf{D}, \mathbf{W}, \mathbf{b}, \epsilon} \|\mathbf{X} - \mathbf{D}\mathbf{C}\|_F^2 + \sum_{i=1}^{n} \sum_{\ell=1}^{n_b} \|\mathbf{c}_i^\ell\|_2 (\lambda_G + \lambda_D \epsilon_{i,\ell}) + \lambda_M \sum_{\ell=1}^{n_b} \|\mathbf{w}_\ell\|_2^2 \quad (4)$$
$$\text{subject to} \quad y_i(\mathbf{w}_\ell^T \mathbf{c}_i^\ell + b_\ell) \geq 1 - \epsilon_{i,\ell} \quad \forall i, \ell$$
$$\|\mathbf{d}_k\|_2 \leq 1 \quad \forall k$$
$$\epsilon_{i,\ell} \geq 0 \quad \forall i, \ell$$

$$y^\star = \arg\min_{y \in \{-1,+1\}} \{\min_{\{\mathbf{c}_j\}} \sum_j \mathcal{E}(\mathbf{x}_j^\star, \mathbf{c}_j, \mathbf{D}, \mathbf{W}, \mathbf{b}, y)\} \quad (6)$$

where $\{\mathbf{x}_j\}$ denotes the set of multiple instances belonging to the test subject.

3 Experiments and Results

Data: We validated our framework using data from a dataset comprising of Autism Spectrum Disorder (ASD) patients and controls. Autism dataset images were T1-weighted MR scans from a 3T scanner, acquired sagitally using volumetric 3D MPRAGE with 0.8 mm × 0.8 mm in plane resolution and 0.9 mm thick sagittal slices. The ASD dataset included 206 subjects (105 patients / 101 controls) representing both sexes and no age matching.

Pre-processing: The acquired images were all bias corrected, skull stripped and had cerebellum removed. The resulting brain images were then segmented into white matter (WM), gray matter (GM) and ventricle (VN). The images were then registered to a template brain using registration algorithm in [10] using varying smoothness penalty terms from $\lambda = 0$ to $\lambda = 1$ to obtain 11 instances per independent subject. The resulting deformation field and the segmentation results were used to obtain WM, GM, VN tissue density maps [4], which were then used for the subsequent analysis. In order to reduce computational burden,

the dimensionality of the data was reduced to 1000 (which amounts to $\approx 95\%$ of the energy) by the random matrix projections method described in [8].

Parameter Selection: As in all generative models, an important issue is hyper parameter selection. In our model, we have 5 hyperparameters: 1) λ_G: generative penalty, 2) λ_D: discriminative loss penalty, 3) λ_M: discriminative model complexity penalty, 4) n_b: number of blocks/clusters, 5) n_a: number of atoms per block.

Cross validating to select all of these parameters is a daunting task, therefore we limited our cross validation to smaller set of values for each hyperparameter obtained by the following heuristics: Since the number of instances per each sample was 11, it could be assumed that there are at most 11 possible bands of information. We limited the number of blocks to be half of this value: $n_b \in \{1, 2, 3, 4, 5\}$. We limited the number of atoms per block to $n_a \in \{10, 20, 40, 80\}$ to control overfitting as suggested in [3]. We fixed the $\lambda_M = 1$. λ_D and λ_G were cross validated from possible values of $\{0, 1, 2, 8\}$. A robust parameter combination that we found among these values was $\lambda_G = 2$, $\lambda_D = 8$. All $n_b \geq 3$ achieved a plateau performance given $n_a \geq 20$. We used $n_b = 3$ and $n_a = 20$.

Classification Results: We compared the classification performance of our method with both linear support vector machine (**lin-SVM**) and Gaussian kernel support vector machine (**rbf-SVM**). The input data for these methods was both single instances (denoted by *sing.*) obtained by choosing the best registration parameter by cross validation and also multiple instances (denoted by *mult.*). The multiple instance data was classified using two protocols. The first approach is a variant of multiple instance SVM [1] that was adapted for neuroimaging datasets in [11] (denoted by **ensemble-MI-SVM**). The second approach consists of naively concatenating all instances of a subject in a single vector (denoted by *naive-concat.*) and using these to train the **lin-SVM** and **rbf-SVM**. All of these methods underwent cross validation to determine the best kernel width and hinge loss penalty. All classification was done by separating the data in two parts and training on the half of data and testing on the other half. The splits were randomly permuted 100 times to obtain a confidence interval. Our method is denoted as block sparse supervised dictionary learning (**BS-SDL**). The results are given in Table 1.

As it can be seen, our method performs comparably well to some of the state of art methods for both single instance learning and multiple instance learning. Although **rbf-SVM** has a slightly higher mean AUC than our method, a t-test has shown that the difference is not significant. In addition, the contribution of our work is that **BS-SDL** discriminates as well as **rbf-SVM** in a dimension much lower (60) than the input dimension of **rbf-SVM** (1000). This shows that our model effectively performs feature selection. Furthermore, the block dictionaries learned shed insight into how the data clusters.

Clustering Results: Although we did not have clinical labels to assess the quality of clustering, inspecting the loading coefficients learned for our training samples demonstrates that enforcing structured sparsity indeed does cluster the

Table 1. Mean area under the ROC curve (AUC) and standard deviation for 100 permutations of 50% training and 50% testing

Method	Mean AUC	St.Dev
BS-SDL	0.7131	0.0410
*sing.***lin-SVM**	0.6819	0.0509
*sing.***rbf-SVM**	0.7200	0.0492
*mult.***lin-SVM**.*naive-concat.*	0.6762	0.0527
*mult.***rbf-SVM**.*naive-concat.*	0.6744	0.0529
*mult.***ensemble-MI-SVM**	0.6735	0.0516

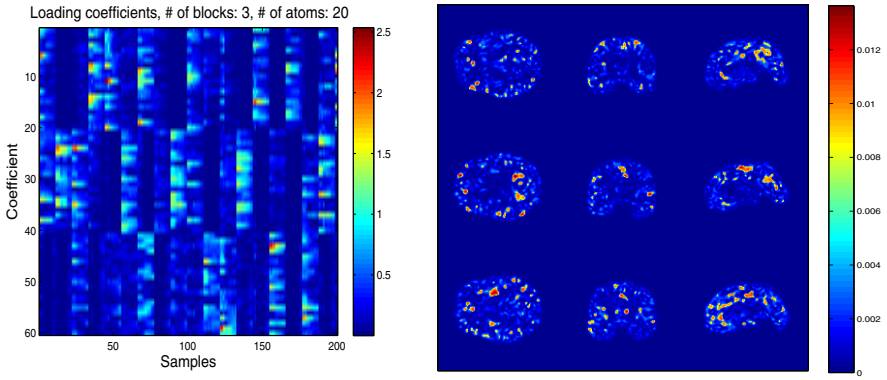

Fig. 2. Left: The learned block sparse loading coefficients **C**, **Right**: Most utilized gray matter atoms for each block (rows) of the learned dictionary **D**. Note: Images are absolute values for visual clarity

data in some way. In figure 2, the block sparsity of the loading coefficients **C** can be observed along with some of the atoms that comprise the corresponding dictionary blocks. We intend to do further analysis in future work to assess the interpretation of this clustering action.

4 Discussion

In this work, we have presented a novel dictionary learning method that is both discriminative and has clustering capabilities. We achieved this by promoting formation of independent blocks within the learned dictionary that only operate on a subset of data. In addition, adaptively penalizing the block sparsity of the loading coefficients by the classification loss allows the formation of discriminative dictionary blocks. We used this methodology to address an important issue in computational anatomy: the selection of pre-processing parameters. In particular, our method allows to circumvent the potentially biased pre-processing parameter selection step by being able to operate on multiple instances generated by pre-processing the data under various parameter settings. It is of importance

to note that although we demonstrate the effectiveness of our model using instances obtained by registrations at varying levels of smoothness, our framework is able to handle multiple instances generated by other processes such as segmentation, selection of varying registration templates or even selection of different registration algorithms. Another advantage of our work is that it is not limited to only multiple instanced or single instanced data as it can readily perform classification on both. In the future, we would like validate our method on additional datasets where heterogeneity is more evident to assess our clustering performance. In conclusion, perhaps the most significant strength of our model is that it is competitive in terms in classification performance with respect to some state of the art methods. Furthermore, we achieve this performance using far fewer input dimensions due to dictionary construction.

References

[1] Andrews, S., Tsochantaridis, I., Hofmann, T.: Support vector machines for multiple-instance learning. In: Advances in Neural Information Processing Systems, pp. 561–568 (2003)
[2] Ashburner, J., Friston, K.J.: Voxel-based morphometry - the methods. Neuroimage 11(6), 805–821 (2000)
[3] Batmanghelich, N.K., Taskar, B., Davatzikos, C.: Generative-discriminative basis learning for medical imaging. IEEE Transactions on Medical Imaging 31(1), 51–69 (2012)
[4] Davatzikos, C., Genc, A., Xu, D., Resnick, S.M.: Voxel-based morphometry using the ravens maps: methods and validation using simulated longitudinal atrophy. NeuroImage 14(6), 1361–1369 (2001)
[5] Eldar, Y.C., Kuppinger, P., Bolcskei, H.: Block-sparse signals: Uncertainty relations and efficient recovery. IEEE Transactions on Signal Processing 58(6), 3042–3054 (2010)
[6] Elhamifar, E., Vidal, R.: Sparse subspace clustering: Algorithm, theory, and applications. IEEE Transactions on Pattern Analysis and Machine Intelligence 35(11), 2765–2781 (2013)
[7] Fan, Y., Shen, D., Gur, R.C., Gur, R.E., Davatzikos, C.: Compare: classification of morphological patterns using adaptive regional elements. IEEE Transactions on Medical Imaging 26(1), 93–105 (2007)
[8] Halko, N., Martinsson, P.G., Tropp, J.A.: Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. SIAM Review 53(2), 217–288 (2011)
[9] Mairal, J., Bach, F., Ponce, J.: Task-driven dictionary learning. IEEE Transactions on Pattern Analysis and Machine Intelligence 34(4) (2012)
[10] Ou, Y., Sotiras, A., Paragios, N., Davatzikos, C.: Dramms: Deformable registration via attribute matching and mutual-saliency weighting. Medical Image Analysis 15(4), 622–639 (2011)
[11] Varol, E., Gaonkar, B., Davatzikos, C.: Classifying medical images using morphological appearance manifolds. In: 2013 IEEE 10th International Symposium on Biomedical Imaging (ISBI), pp. 744–747. IEEE (2013)