Neighbourhood Approximation Forests

Ender Konukoglu, Ben Glocker, Darko Zikic, and Antonio Criminisi

Microsoft Research Cambridge

Abstract. Methods that leverage neighbourhood structures in high-dimensional image spaces have recently attracted attention. These approaches extract information from a new image using its "neighbours" in the image space equipped with an application-specific distance. Finding the neighbourhood of a given image is challenging due to large dataset sizes and costly distance evaluations. Furthermore, automatic neighbourhood search for a new image is currently not possible when the distance is based on ground truth annotations. In this article we present a general and efficient solution to these problems. "Neighbourhood Approximation Forests" (NAF) is a supervised learning algorithm that approximates the neighbourhood structure resulting from an arbitrary distance. As NAF uses only image intensities to infer neighbours it can also be applied to distances based on ground truth annotations. We demonstrate NAF in two scenarios: i) choosing neighbours with respect to a deformation-based distance, and ii) age prediction from brain MRI. The experiments show NAF's approximation quality, computational advantages and use in different contexts.

1 Introduction

Computational methods that leverage available datasets for analyzing new images show high accuracy and robustness. Among these methods one class that has lately shown significant potential is *neighbourhood*-based approaches. These approaches formulate the set of all images as a high-dimensional space equipped with an application-specific distance. They then utilize the neighbourhood structure of this space for various tasks. The underlying principle is that neighbouring images, in other words images that are similar with respect to the distance, provide valuable and accurate information about each other. Therefore, when analyzing a new image one can propagate information from its neighbours.

Neighbourhood-based approach, as a general framework, has recently been applied in different contexts. Patch-based techniques [7] and multi-atlas based methods [11] utilize it for segmenting medical images. Nonlinear "manifold"-based methods, which are used in different applications [10,20], also rely on the neighbourhood-based approach, i.e. the neighbourhood structure is preserved during the low-dimensional embedding and subsequent analyses in the low-dimensional space are based on this structure.

One problem in neighbourhood-based approaches, which currently limits their use, is determining the close neighbours of a new image within an existing

N. Ayache et al. (Eds.): MICCAI 2012, Part III, LNCS 7512, pp. 75–82, 2012.

dataset. In theory, to determine this neighbourhood one should compute the distances between the new image and all the other images. However, depending on the nature of the distance and the size of the training set this exhaustive search can be computationally very expensive or even impossible. For instance, in multi-atlas based segmentation one would register a new image to all other images to determine its neighbours and propagate labels based on this. The cost of this exhaustive search is high due to computational times of nonlinear registration. Similar problems exist in "manifold"-based techniques, as also pointed out in [2]. In case the distance is defined with respect to ground truth annotation not available for the new image then exhaustive search becomes impossible.

Besides exhaustive search, currently used techniques for finding the neighbourhood of a new image is either through heuristic search strategies [1,7] or K-means like approaches such as multi-template constructions [16,3]. Heuristic strategies are based on application specific rules, therefore not flexible. K-means like approaches have a trade-off in choosing the number of centroids, i.e. too many will result in a computational bottleneck and too few will not correctly reflect the neighbourhood structure. In manifold techniques, some methods find the manifold coordinates of a new image without reconstructing the embedding, [5,14]. However, these methods also rely on computing all distances. Lastly, if a set of low-dimensional features that describes the neighbourhood structure is known then quantization [13] and hashing [19,15] techniques create short binary codes from these features for fast image retrieval. Construction of the initial low-dimensional features still remains an open problem though.

In this article, we present "Neighborhood Approximation Forests" (NAF), a general supervised learning algorithm for approximating an arbitrary neighbourhood structure using image intensities. The main principle of NAF is to learn a compact representation that can describe the neighbourhood structure of a high-dimensional image space equipped with a user-specified distance. For a new image, NAF predicts its neighbourhood within an existing dataset in an efficient manner. We first define the general framework of neighbourhood-based approaches and detail the proposed algorithm. In the experiments we apply NAF to two applications. First we treat the problem of determining the closest neighbours of a new image within a training set with respect to the amount of deformation between images. This experiment demonstrates the prediction accuracy of NAF compared to the real neighbourhood structure and shows the computational advantages. In the second application we devise a simple neighbourhood-based regression method powered by NAF to solve the "toy" problem of age prediction using brain MRI. This experiment demonstrates the use of NAF on an image space where the neighbourhood relation is determined by a continuous and non-image based meta information. Results show high regression accuracies achieved by NAF compared to the values reported in the literature.

2 Neighbourhood Approximation Forests

Neighbourhood-based approach (NbA) is a general framework that is applied for various image analysis tasks. The underlying principle is to extract information

from an image using other "similar" images within a dataset with ground truth, i.e. *training set*. NbA formulates the set of all images as a high-dimensional space $\mathcal{I}$, where each point $I \in \mathcal{I}$ denotes an image. The dataset with ground truth is a finite subset within this space $\mathbf{I} = \{I_p\}_{p=1}^{P} \in \mathcal{I}$. The space $\mathcal{I}$ is equipped with a distance $\rho(I, J)$ that quantifies a similarity between images, which is application dependent. For an image I the set of k most similar images in $\mathbf{I}$ is then defined as the neighborhood $\mathbf{N}_\rho^k(I)$, i.e. k images with the lowest distance to I. To analyse a new image $J \notin \mathbf{I}$, one needs to determine $\mathbf{N}_\rho^k(J)$ within $\mathbf{I}$ to be able to use NbA. This is challenging because the computation of $\rho(\cdot, \cdot)$ between J and all images in $\mathbf{I}$ can be expensive or even not possible. In the following we describe a learning algorithm to approximate $\mathbf{N}_\rho^k(J)$ that overcomes these challenges.

Our approach relies on the hypothesis that the neighbourhood structure constructed by $\rho(\cdot, \cdot)$ can be approximated using compact image descriptors derived from intensity information. Consequently, using these descriptors, for a new image J we can approximate its neighborhood $\mathbf{N}_\rho^k(J)$ within $\mathbf{I}$ without the need to evaluate $\rho(\cdot, \cdot)$. Neighborhood Approximation Forests (NAF) is a supervised algorithm that learns such descriptors for arbitrary $\rho(\cdot, \cdot)$. It is a variant of random decision forests [6,8], i.e. an ensemble of binary decision trees, where each tree is an independently learned predictor of $\mathbf{N}_\rho^k(J)$ given J. As all supervised learning algorithms NAF has two phases: training and prediction. Below we explain these phases and then demonstrate NAF in Section 3.

Predicting neighbourhood with a single tree: We represent each image I using a set of intensity-based features $\mathbf{f}(I) \in \mathbb{R}^Q$ of possibly high dimensions, which can be as simple as intensity values at different points. These features have no prior-information on $\rho(\cdot, \cdot)$. For a new image J, each tree T predicts J's neighbours within a training set $\mathbf{I}$ by applying a sequence of learned binary tests to a subset of its entire feature vector $\mathbf{f}_T(J) \in \mathbb{R}^q,\ q < Q$ and $\mathbf{f}_T(J) \subset \mathbf{f}(J)$. Each binary test in the sequence depends on the result of the previous test. This whole process is represented as a binary decision tree [4], where each test corresponds to a branching node in the tree. Starting from the root node s_0 the image J traverses the tree taking a specific path and arrives at a node with no further children, a leaf-node. The path and the final leaf-node depend on the feature vector $\mathbf{f}_T(J)$ and the binary tests at each node.

Each leaf-node stores the training images (or simply their indeces) $I_n \in \mathbf{I}$ which traversed T and arrived at that node. So, at the leaf-node J arrives there is a subset of training images which have taken the same path as J and therefore share similar feature values based on the applied tests. This subset of training images, $\mathbf{N}_{T(\rho)}(J)$, is the neighbourhood of J predicted by T. The subscript $T(\rho)$ denotes the tree's dependence on $\rho(\cdot, \cdot)$, which we explain in the training part.

Approximating neighborhood with the forest: The forest F is composed of multiple independent trees with independent predictions. Each tree works with a different subset of features $\mathbf{f}_T(J) \subset \mathbf{f}(J)$ focusing on a different part of the feature space. We compute the ensemble forest prediction by combining the independent tree predictions. This combination process computes the approximate affinity of J to each I_n by $\mathbf{w}_F(J, I_n) \triangleq \sum_{\forall T \in F} \mathbf{1}_{\mathbf{N}_{T(\rho)}(J)}(I_n)$, where $\mathbf{1}_A(x)$

is the indicator function (we note that [9] uses a similar construction for a different purpose: defining a neighborhood structure) The forest prediction of $\mathbf{N}_{\rho}^{k}(J)$ is simply the k training images with the largest $\mathbf{w}_F(J, I_n)$ values. We denote this set with $\mathbf{N}_{F(\rho)}^{k}(J)$. Once again the subscript denotes the ρ of the forest.

Training: In order to learn the structure of a tree we use the training set $\mathbf{I}$ and the distances $\rho(I_n, I_m)$ for each image pair in $\mathbf{I}$. Our goal is to find the sequence of binary tests on image features that sequentially partition $\mathbf{I}$ into the most spatially compact subsets with respect to $\rho(\cdot, \cdot)$. Assuming $\mathbf{I}$ is a representative dataset, the learned binary tests would then successfully apply to other images.

Given a node s and the set of training images at it, $\mathbf{I}_s$, we first define branching of s via the binary test and the partitioning of $\mathbf{I}_s$ into two as

$$t_s(I_n; m, \tau) \triangleq \begin{cases} I_n \in \mathbf{I}_{s_R}, \text{ if } f_T^m(I_n) > \tau, \\ I_n \in \mathbf{I}_{s_L}, \text{ if } f_T^m(I_n) \leq \tau, \end{cases} \quad \forall I_n \in \mathbf{I}_s \tag{1}$$

where f_T^m denotes the m^{th} component of $\mathbf{f}_T(I_n)$, $\tau \in \mathbb{R}$, and s_L and s_R are the children of s. At every node we would like to optimize the parameters m and τ to obtain the most compact partitioning of $\mathbf{I_s}$. To do this we define spatial compactness of a set $\mathbf{A}$ with respect to $\rho(\cdot, \cdot)$ as

$$C_\rho(\mathbf{A}) \triangleq \frac{1}{|\mathbf{A}|^2} \sum_{I_i \in \mathbf{A}} \sum_{I_j \in \mathbf{A}} \rho(I_i, I_j), \tag{2}$$

where $|\mathbf{A}|$ denotes the size of the set and $C_\rho(\mathbf{A})$ its *cluster size*. Using $C_\rho(\cdot)$ we can formulate the gain in compactness a specific set of parameters yields with

$$G(\mathbf{I}_s, m, \tau) \triangleq C_\rho(\mathbf{I}_s) - \frac{|\mathbf{I}_{s_R}|}{|\mathbf{I}_s|} C_\rho(\mathbf{I}_{s_R}) - \frac{|\mathbf{I}_{s_L}|}{|\mathbf{I}_s|} C_\rho(\mathbf{I}_{s_L}), \tag{3}$$

where the weights $|\mathbf{I}_{s_R}|/|\mathbf{I}_s|$ and $|\mathbf{I}_{s_L}|/|\mathbf{I}_s|$ avoid constructing too small partitions. Using this formulation we determine the best possible binary test at node s with the following optimization problem

$$(m_s, \tau_s) = \arg_{m,\tau} \max \; G(\mathbf{I}_s, m, \tau). \tag{4}$$

In practice we do not take into account all m in the above optimization problem but choose a small random subset of the components of $\mathbf{f}_T(\cdot)$ at each node as is commonly done in decision forests [8]. The optimization over τ though is done through exhaustive search.

For each tree we start from its root node setting $\mathbf{I}_{s_0} = \mathbf{I}$. We then sequentially determine the binary tests using Eqn. 4 and add new nodes to the tree. We continue this process and grow the trees. The growth process is terminated at a node when i) we can no longer find a test that creates a more compact partitioning than the one in the node, i.e. $\forall (m, \tau), G < 0$, ii) the number of training images within the node is too small or iii) we reach at the maximum allowed depth and stop due to computational cost considerations.

3 Experiments

In this section we demonstrate NAF on two different applications. Our aim is to analyze NAF in different experimental setups and for different image spaces. We also highlight the application-specific components that can be changed to use NAF in different contexts. For both experiments we use 355 T1 weighted brain MR images from the publicly available OASIS dataset [12]. These images are skull stripped, histogram equalized and aligned to a common reference frame via affine registration. The resolution of each image is $1 \times 1 \times 1\ mm^3$.

A. Choosing the Closest Images for Non-linear Registration: In the first application we focus on predicting the neighbourhood of a new image J within a dataset $\mathbf{I}$ with respect to the amount of deformation between images. We predict images in $\mathbf{I}$ that need the least amount of deformation to nonlinearly align them to J. This is a relevant problem for large cohort studies and multi-atlas based segmentation methods. Our aim in this experiment is to demonstrate the quality of NAF's predictions compared to the real neighbourhoods for this highly nonlinear problem. The application specific and experimental details are given below along with results and discussions.

$\rho(\cdot,\cdot)$: We measure the amount of deformation between two images using the distance $\rho(I,J) \triangleq \int_{\Omega_I} \log|\mathrm{Jac}(\Phi_{I\to J})|\, d\Omega_I + \int_{\Omega_J} \log|\mathrm{Jac}(\Phi_{J\to I})|\, d\Omega_J$, where Ω_I is the domain of I, $\Phi_{I\to J}$ is the deformation mapping I to J, i.e. $\Phi_{I\to J} \circ I = J$, and $\mathrm{Jac}(\cdot)$ is the Jacobian determinant. We use the diffeomorphic demons algorithm [18] for determining each deformation.

Dataset: The first 169 images are used in training and the rest 186 for testing.

Features: We randomly choose $Q = 10000$ pairs of voxels in the reference frame. Then we smooth each image with an averaging kernel of size $12 \times 12 \times 12\ mm^3$. The feature vector for each image consists of the intensity differences between the pairs of voxels in the smooth version of the image.

NAF details: Using the training set we train a NAF of 1500 trees, each of maximum depth 6. Minimum number of allowable training images for a node is set to 7 beyond which we stop growing the tree. Each tree is constructed using a random subset of the entire feature vector of size $q = 1000$. For each test image J we predict its neighbourhood, $\mathbf{N}^k_{F(\rho)}(J)$, for different values of $k = 1, 3, 5, 7, 10$.

Evaluation and Results: For each test image J, we evaluate the quality of $\mathbf{N}^k_{F(\rho)}(J)$ by comparing it to the real neighbourhood $\mathbf{N}^k_{\rho}(J)$ using the following ratio

$$\varrho_J(\mathbf{N}^k_{F(\rho)}(J)) \triangleq \frac{\sum_{I\in\mathbf{N}^k_{F(\rho)}(J)} \rho(I,J)}{\sum_{I\in\mathbf{N}^k_{\rho}(\mathbf{J})} \rho(I,J)} \geq 1, \tag{5}$$

which measures how close the images in $\mathbf{N}^k_{F(\rho)}(J)$ to J compared to the ones in $\mathbf{N}^k_{\rho}(J)$. In Table 1 we provide the mean values and standard deviations of $\varrho_J(\mathbf{N}^k_{F(\rho)}(J))$ computed over 186 test images for different k. These values can be best interpreted in comparison with the ranges $\varrho_J(\cdot)$ can take for each k. In

order to present these ranges, for each test image J and each k we randomly chose 2000 subsets within the training set. We denote each of these subsets by $\mathbf{N}^k_{\mathbf{r}}(J)$. We then computed $\varrho_J(\mathbf{N}^k_{\mathbf{r}}(J))$ values and present the mean and standard deviations for these random subsets (computed over 186×2000 subsets for each k) in Table 1. Results given in Table 1 demonstrate that NAF predictions are indeed very close to the real neighbourhoods in terms of their distances to J. Especially in comparison with $\varrho_J(\mathbf{N}^k_{\mathbf{r}}(J))$ we notice that $\varrho_J(\mathbf{N}^k_{F(\rho)}(J))$ values are within the lowest part of the entire range of $\varrho_J(\cdot)$. We further plot in Figures 1(a) and (b) the normalized histograms for $\varrho_J(\mathbf{N}^k_{\mathbf{r}}(J))$ and $\mathbf{N}^k_{F(\rho)}(J)$ for $k = 1$ and $k = 7$. Comparing these histograms we see that the distribution of $\mathbf{N}^k_{F(\rho)}(J)$ is more concentrated close to one and it lies in the lower frequency region of the distribution for $\varrho_J(\mathbf{N}^k_{\mathbf{r}}(J))$. The difference is even more pronounced for $k = 7$, i.e. choosing multiple neighbours, which is more relevant for most applications such as multi-atlas based segmentation. Lastly in Figures 1(c)-(e) we show two sets of examples (different rows) where NAF predicts a different closest neighbour than the real one. However, visually the test image and the predicted neighbour are very similar.

Computation Times: For each test image NAF took at maximum 10.2 seconds to predict the neighbourhood with a C++ implementation on an Intel Xeon® at 2.27 GHz. Exhaustive search requires 169 nonlinear registrations which took on the average 1.9 hours for each test image.

B. Age Regression from Brain MR Scans: In the second application we focus on a high-dimensional image space equipped with a distance based on non-image based meta information: subject age. We devise an image-based regression algorithm powered by NAF to predict the age of a subject using the MR image. Our aim is to demonstrate the use of NAF for this type of applications and also quality of the predicted neighbourhood through an analysis end result.

$\rho(\cdot,\cdot)$: The distance of the image space is $\rho(I, J) = |\text{age}(I) - \text{age}(J)|$, where age$(I)$ denotes the subject's age with image I and $|\cdot|$ is the absolute value.

Dataset: We use the 355 images and perform leave-one-out tests.

Features: We randomly choose $Q = 10000$ voxels in the reference frame and use the intensity values taken from the images smoothed as in the previous case.

NAF details: Most details of NAF are the same as the previous case. The only differences is this time the maximum tree depth is 12 and we use 700 trees.

Evaluation and Results: In this application we evaluate NAF's results by comparing the real age of the test subject with the prediction obtained using the neighbourhood predicted by NAF. For each test image J we predict the age of the subject by taking the average age in $\mathbf{N}^{15}_{F(\rho)}(J)$. Figure 1(f) plots the predicted age vs. actual age for all 355 tests. The resulting correlation is reasonable high with a r-value $= 0.93$ ([17] reports slightly lower values for a slightly smaller dataset). We observe that NAF is able to approximate an informative image neighbourhood for a new image that is useful for the regression analysis.

Table 1. Top row: mean and standard deviations for the ratios of total distance from $\mathbf{N}^k_{F(\rho)}(J)$ to J and from $\mathbf{N}^k_{\rho}(J)$ to J, see Eqn. 5. Bottom row: presents the range of $\varrho_J(\cdot)$ within the training set by providing same values for random subsets of the training set. NAF predictions are very close the real neighbourhood considering the range of $\varrho_J(\cdot)$.

k	1	3	5	7	10
$\varrho_J(\mathbf{N}^k_{F(\rho)}(J))$	1.05 ± 0.04	1.05 ± 0.02	1.04 ± 0.02	1.04 ± 0.02	1.04 ± 0.01
$\varrho_J(\mathbf{N}^k_{\mathbf{r}}(J))$	1.20 ± 0.07	1.18 ± 0.06	1.18 ± 0.06	1.17 ± 0.06	1.16± 0.06

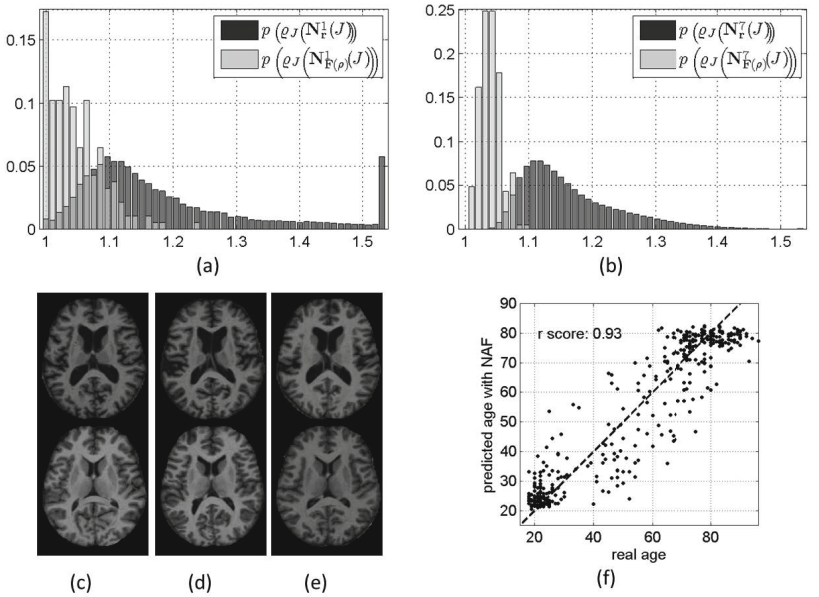

Fig. 1. Experiment A:(a,b) Normalized histograms of $\varrho_J(\mathbf{N}^k_{F(\rho)}(J))$ (light) and $\varrho_J(\mathbf{N}^k_{\mathbf{r}}(J))$ (dark) for $k = 1$, 7 respectively. NAF predictions are concentrated close to one and lie in the low frequency region of the distribution for $\varrho_J(\mathbf{N}^k_{\mathbf{r}}(J))$ (c)-(e) Two tests (different rows) where NAF suggests a different closest image than the real one: (c) the test image, (d) real closest (e) NAF prediction. Note that images are very similar visually. **Experiment B**:(f) Image-based regression for age prediction by NAF using $\mathbf{N}^{15}_{F(\rho)}(J)$. Note the high correlation $r = 0.93$.

4 Conclusion

We proposed an algorithm for solving one of the critical problems common to all neighborhood-based approaches for image analysis: approximating the neighborhood of a new image within a training set of images with respect to a given distance. The algorithm is general and can be applied to various tasks that utilize different distance definitions, as shown in the experiments. Furthermore, as the method is based on the framework of random decision forests the computation times are fast. We believe that applications such as multi-atlas registration and 'manifold'-based techniques can benefit from the proposed algorithm.

References

1. Aljabar, P., Heckemann, R., Hammers, A., Hajnal, J., Rueckert, D.: Multi-atlas based segmentation of brain images: Atlas selection and its effect on accuracy. Neuroimage 46(3), 726–738 (2009)
2. Aljabar, P., Wolz, R., Rueckert, D.: Manifold Learning for Medical Image Registration, Segmentation, and Classification. In: Machine Learning in Computer-Aided Diagnosis: Medical Imaging Intelligence and Analysis. IGI Global (2012)
3. Allassonnire, S., Amit, Y., Trouv, A.: Towards a coherent statistical framework for dense deformable template estimation. J. R. Stat. Soc.: Series B 69 (2007)
4. Amit, Y., Geman, D.: Shape quantization and recognition with randomized trees. Neural Computation 9 (1997)
5. Bengio, Y., Paiement, J., Vincent, P., Delalleau, O., Le Roux, N., Ouimet, M.: Out-of-sample extensions for lle, isomap, mds, eigenmaps, and spectral clustering. In: NIPS, vol. 16 (2004)
6. Breiman, L.: Random forests. Machine Learning 45 (2001)
7. Coupe, P., Manjon, J.V., Fonov, V., Pruessner, J., Robles, M., Collins, D.L.: Patch-based segmentation using expert priors: application to hippocampus and ventricle segmentation. Neuroimage 54 (2011)
8. Criminisi, A., Shotton, J., Konukoglu, E.: Decision forests: A unified framework for classification, regression, density estimation, manifold learning and semi-supervised learning. NOW Publishing: Foundations and Trends 7 (2012)
9. Gray, K.R., Aljabar, P., Heckemann, R.A., Hammers, A., Rueckert, D.: Random Forest-Based Manifold Learning for Classification of Imaging Data in Dementia. In: Suzuki, K., Wang, F., Shen, D., Yan, P. (eds.) MLMI 2011. LNCS, vol. 7009, pp. 159–166. Springer, Heidelberg (2011)
10. Hamm, J., Ye, D.H., Verma, R., Davatzikos, C.: GRAM: A framework for geodesic registration on anatomical manifolds. Med. Image Anal. 14 (2010)
11. Jia, H., Yap, P.T., Shen, D.: Iterative multi-atlas-based multi-image segmentation with tree-based registration. Neuroimage 59 (2012)
12. Marcus, D., Wang, T., Parker, J., Csernansky, J., Morris, J., Buckner, R.: Open access series of imaging studies (oasis): cross-sectional mri data in young, middle aged, nondemented, and demented older adults. J. of Cog. Neuroscience 19 (2007)
13. Nister, D., Stewenius, H.: Scalable recognition with a vocabulary tree. In: CVPR, vol. 2, pp. 2161–2168 (2006)
14. Niyogi, X.: Locality preserving projections. In: NIPS, vol. 16 (2004)
15. Norouzi, M., Fleet, D.: Minimal loss hashing for compact binary codes. In: ICML (2011)
16. Sabuncu, M.R., Balci, S.K., Shenton, M.E., Golland, P.: Image-driven population analysis through mixture modeling. IEEE Trans. Med. Imaging 28 (2009)
17. Sabuncu, M.R., Van Leemput, K.: The Relevance Voxel Machine (RVoxM): A Bayesian Method for Image-Based Prediction. In: Fichtinger, G., Martel, A., Peters, T. (eds.) MICCAI 2011, Part III. LNCS, vol. 6893, pp. 99–106. Springer, Heidelberg (2011)
18. Vercauteren, T., Pennec, X., Perchant, A., Ayache, N.: Diffeomorphic demons: Efficient non-parametric image registration. NeuroImage 45 (2009)
19. Weiss, Y., Torralba, A., Fergus, R.: Spectral hashing. In: NIPS (2008)
20. Wolz, R., Aljabar, P., Hajnal, J.V., Hammers, A., Rueckert, D., Weiner, M.W.: LEAP: learning embeddings for atlas propagation. Neuroimage 49 (2010)