Published in final edited form as:

Mach Learn Med Imaging. 2016 ; 10019: 313–321. doi:10.1007/978-3-319-47157-0_38.

Fast Neuroimaging-Based Retrieval for Alzheimer's Disease Analysis

Xiaofeng Zhu, Kim-Han Thung, Jun Zhang, and Dinggang She

Department of Radiology and BRIC, University of North Carolina at Chapel Hill, Chapel Hill, USA

Abstract

This paper proposes a framework of fast neuroimaging-based retrieval and AD analysis, by three key steps: (1) *landmark detection*, which efficiently extracts landmark-based neuroimaging features without the need of nonlinear registration in testing stage; (2) *landmark selection*, which removes redundant/noisy landmarks via proposing a feature selection method that considers structural information among landmarks; and (3) *hashing*, which converts high-dimensional features of subjects into binary codes, for efficiently conducting approximate nearest neighbor search and diagnosis of AD. We have conducted experiments on Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset, and demonstrated that our framework could achieve higher performance than the comparison methods, in terms of accuracy and speed (at least 100 times faster).

1 Introduction

Recent studies have demonstrated that neuroimaging data are useful for neurodegenerative disease analysis [2,7]. For example, Magnetic Resonance (MR) image is able to show different soft tissues in the brain with good contrast, and thus gives us important information about brain atrophy resulting from neurodegeneration. In addition, MR image is also non-invasive and safe. As a result, studies have been using MR images to devise biomarkers for AD (and its related disease stages) identification [8,10,17].

Conventional methods of AD diagnosis using neuroimaging data (*e.g.*, MR images), like ROI-based methods [16,17] and voxel-based methods [3], involve a time-consuming nonlinear registration process to register all the images into the same common space, before performing further analysis. For example, Zhang *et al.* [11] reported that the time cost of nonlinear registration in the seminal works, such as HAMMER [6] and FSL [13], are around 30 min and at least 3 min, respectively. Such time cost makes the conventional methods impractical for realtime neuroimaging analysis in clinical study. In addition, rather than predicting categorical labels, the clinician is also interested in retrieving similar subjects of each new subject, to assist diagnosis. This is unfortunately unaccounted for by the conventional methods, which only provide the diagnosis label prediction.

In this paper, we overcome the aforementioned issues by proposing a framework of fast neuroimaging-based retrieval and AD analysis, which consists of three key elements: (1) *landmark detection* [12], which avoids time-consuming process of nonlinear registration, as needed in conventional ROI-based and voxel-based methods, in testing stage. We first identify the discriminative landmarks in the MR images, via group comparison in the training set. Then, we construct a regression-forest-based landmark detection model to efficiently predict the landmarks of testing MR images. In this way, the neuroimaging features of testing MR images can be extracted without involving the process of non-linear registration and thus enabling to conduct real-time neuroimaging-based retrieval and AD analysis. (2) *landmark selection*, which removes redundant/noisy landmarks by proposing a novel feature selection method. We first arrange the feature vector of each landmark side by side to obtain a feature matrix for each subject. Then, we consider structural information among landmarks and possible noise in the feature vector of each landmark, to devise a novel landmark selection model. (3) *hashing*, which converts the high-dimensional feature vector of each subject into a short binary code. After landmark selection, the features of the selected landmarks within one subject are concatenated into a long vector, where the dimension is still too high (more than 20,000 in our experiments) to train a stable classifier. Hashing is able to convert long-dimensional vector into short binary codes while preserving their neighborhood relationship. By representing each subject by a short binary code vector, we can effectively retrieve the neighboring MR images of a testing image for AD retrieval and analysis.

2 Proposed Framework

Figure 1 shows the flowchart of the proposed framework, which consists of 2 stages, *i.e.*, training stage and testing stage, respectively. We describe details in the following.

2.1 Landmark Detection

We employ the method in [11] to define AD-related landmarks in MR images and learn a regression-forest-based landmark detection model. We first nonlinearly register all the training MR images to a template image, where the voxels with their corresponding local morphological features significantly different between two groups (*i.e.*, AD and normal control (NC)) are identified as landmarks. Then, we map the landmarks in the template to the training images (through the deformation fields estimated by nonlinear registrations), to obtain the landmarks of all the training images. Using the detected landmarks, we construct a regression-forest-based landmark detection model, where the Low-Energy Pattern (LEP) features [12] of the subject voxels and their corresponding displacements to the detected landmarks are used as input and output of the model, respectively. Note that all these time-consuming processes are done offline. In testing stage, we use the learnt landmark detection model to find the landmarks of testing MR images, which is extremely fast, as no time-consuming process (*i.e.*, either nonlinear registration or regression-forest training) is involved.

2.2 Landmark Selection

After landmark detection, Zhang *et al.* [11] uses the morphological (*i.e.*, LEP) features of all the landmarks to train a classifier (*i.e.*, Support Vector Machine (SVM)) for predicting categorical labels of subjects. As the number of landmarks is about 1741, and the number of features for each landmark is 50, the total number of features for each subject is very high, *i.e.*, over 80,000, while there are only few hundreds of subjects available. Obviously, there is a “High Dimension Low Subject Size” (HDLSS) issue, which makes conventional machine learning techniques like SVM unstable. Moreover, as the landmarks are detected based on group comparison, which does not consider correlations among landmarks, for such large number of landmarks, it is highly possible that the detected landmarks are redundant or not discriminative to the classifier. Thus, we perform landmark selection to remove redundant and noisy landmarks.

To this end, we propose a novel landmark selection method that integrates subspace learning and 2D feature selection in a unified framework. Firstly, we represent each subject as a 2D feature matrix, by concatenating the LEP features of all the landmarks in one subject side by side, rather than concatenating all the features into a long vector (as in [11]). In this way, we preserve the spatial information among landmarks, which helps us in identifying redundant (*e.g.*, by checking the correlation among landmarks) and noisy landmarks (*e.g.*, by checking the discriminative power of the landmarks in classification task). We then convert the high-dimensional LEP features of the landmarks into target-like (*i.e.*, clinical-status-like) features, and jointly select the discriminative landmarks of all the subjects.

Let us denote $\mathbf{X}_i \in \mathbb{R}^{l \times d}$ as the feature matrix of the i -th subject, where l is the number of landmarks, and d is the dimension of LEP features for each landmark. The corresponding clinical status of i -th subject is denoted as $\mathbf{y}_i \in \mathbb{R}^{c \times 1}$, where c is the number of possible clinical statuses. Assuming linear relationship between $\mathbf{X}_i$ and $\mathbf{y}_i$ and n number of subjects, we formulate our landmark detection algorithm as

$$\min_{\mathbf{S}, \mathbf{R}} \sum_{i=1}^n \|\mathbf{y}_i^T - \mathbf{s}_i^T \mathbf{X}_i \mathbf{R}\|_F^2 + \gamma \|\mathbf{S}\|_{2,1}, \text{ s.t., } \mathbf{R}^T \mathbf{R} = \mathbf{I}, \quad (1)$$

where $\mathbf{S} = [\mathbf{s}_1, \dots, \mathbf{s}_n] \in \mathbb{R}^{l \times n}$ and $\mathbf{R} \in \mathbb{R}^{d \times c}$ are coefficient matrices, and γ is the tuning parameter. $\mathbf{s}_i \in \mathbb{R}^{l \times 1}$ is a sparse vector, which is used to select landmarks for i -th subject, and linearly combines the selected landmarks into a single vector (*i.e.*, $\mathbf{s}_i^T \mathbf{X}_i$ is a d dimensional row vector). To select common landmarks for all the subjects, we use group sparsity

regularizer (*i.e.*, $l_{2,1}$ -norm) on $\mathbf{S}$, *i.e.*, $\|\mathbf{S}\|_{2,1} = \sum_i \sqrt{\sum_j s_{ij}^2}$. The group sparsity regularizer (*i.e.*, the second term in Eq. (1)) encourages all-zero-value rows in $\mathbf{S}$, to jointly select discriminative landmarks for all the subjects. In addition, we would like to select landmarks not based on the original high-dimensional (and possibly noisy) feature space spanned by the landmarks, but rather based on a low-dimensional and target-like feature space, via the transformation matrix $\mathbf{R}$. Specifically, $\mathbf{R}$ maps a high-dimensional feature vector of each

landmark (*i.e.*, a d -dimensional row vector of $\mathbf{X}_i$) into a c -dimensional target-like feature space ($\mathbf{X}_i \mathbf{R} \in \mathbb{R}^{1 \times c}$). It is noteworthy that the inequality $c \ll d$ always holds in AD study. With this, the orthogonal constraint on $\mathbf{R}$ enables to output unrelated low-dimensional vectors, thus actually conducting subspace learning.

After optimal condition is met, an index of selected landmarks can be obtained through all-non-zero rows of $\mathbf{S}$.

2.3 Hashing Construction

After landmark selection, we represent each subject using a long feature vector, which is the concatenation of LEP features of the selected landmarks. The dimension of such a feature vector is still too high to construct an effective and efficient classification model. In addition, unlike conventional neuroimaging analysis methods that only focus on classification task, we are also interested in efficiently retrieving similar subjects, which can assist the clinician to conduct AD analysis. Both of these issues can be addressed through hashing, which involves converting high-dimensional data into short binary codes [14,15], while preserving the neighborhood relationship among subjects in the original data space as much as possible. Specifically, the neighboring subjects of each testing subject can efficiently be found by first representing them with short binary codes and then comparing their binary codes (*i.e.*, calculating the Hamming distance between this testing subject and its neighbors) in the memory of PC. We can also obtain the diagnosis label of this testing subject with the majority voting rule on the labels of its neighboring subjects.

To this end, we employ the hashing framework in [18], which includes two key steps, *i.e.*, probabilistic representation step and hash function learning step.

In the probabilistic representation step, we employ a probabilistic method to represent high-dimensional data using low-dimensional probabilistic representation. This step is aimed to resolve the HDLSS issue while preserving the local structures of the data, which are the most important factors in constructing effective hashing method [18]. In the hash function learning step, we encode the probabilistic representations (of all the subjects) using a dictionary, and then binarize the resulting coding coefficients as final hash codes. As described in [18], the dictionary and the binarization threshold are learnt jointly.

Probabilistic Representation—Let $\mathbf{x}_i$ denotes the long feature vector of the i -th subject after landmark selection, and $\tilde{\mathbf{X}} = [\mathbf{x}_1 \cdots \mathbf{x}_n]$ is the feature matrix. We first partition $\tilde{\mathbf{X}}$ into m clusters and then compute the probability of each subject belonging to each cluster by calculating the Euclidean distance between the subject and the centroid of the cluster. Considering the fact that a subject $\mathbf{x}_i$ can only be members of several neighboring clusters, the Probabilistic Representation (PR) of a subject, $\mathbf{p}_i = [p_{i,1} \cdots p_{i,k} \cdots p_{i,m}]$ ($p_{i,k}$ denotes the probability of i -th subject belonging to k -th cluster), should be sparse. Therefore, we only preserve the first few largest probabilities in $\mathbf{p}_i$, and set the remaining minor probabilities to 0. Finally, we denote the PR of $\mathbf{x}_i$ after sparsity thresholding as $\tilde{\mathbf{p}}_i = [\tilde{p}_{i,1} \cdots \tilde{p}_{i,m}]$. $\tilde{\mathbf{p}}_i$ is a sparse vector, and it can well characterize the original spatial location of $\mathbf{x}_i$ in $\tilde{\mathbf{X}}$. In this way, similar subjects should have similar PR vectors, and therefore $\tilde{\mathbf{p}}_i$ is a good representation of $\mathbf{x}_i$ for effective hashing.

Hash Function Learning—Let $\tilde{\mathbf{P}} = [\tilde{\mathbf{p}}_1, \tilde{\mathbf{p}}_2, \dots, \tilde{\mathbf{p}}_n] \in \mathbf{R}^{m \times n}$ be the set of PRs of all training subjects in $\tilde{\mathbf{X}}$. The hash function can then be learnt from $\tilde{\mathbf{P}}$ by using:

$$\min_{\Phi, \Lambda} \|\tilde{\mathbf{P}} - \Phi \Lambda\|_F^2 + \lambda \sum_{i=1}^n \|\alpha_i - \mu\|^2, \text{ s.t. } \|\phi_j\|^2 = 1 \quad (2)$$

where $\Phi \in \mathbf{R}^{m \times a}$ is called the encoding dictionary and each atom $\phi_j \in \mathbf{R}^m$ in Φ has unit ℓ_2 -norm; $\Lambda = [\alpha_1, \dots, \alpha_n] \in \mathbf{R}^{a \times n}$ is the coding matrix of $\tilde{\mathbf{P}}$ using dictionary Φ , with α_i being

the coding vector of $\tilde{\mathbf{p}}_i$; $\mu = \frac{1}{n} \sum_{i=1}^n \alpha_i$ is the mean of all coding vectors; λ is a tuning parameter; and a is the number of atoms or the code length.

We alternatively optimize Φ and Λ to solve Eq. (2). First, we randomly initialize Φ and solve for Λ to yield the following analytical solution to each α_i in Λ :

$$\alpha_i = \mathbf{T} \tilde{\mathbf{p}}_i + \beta \quad (3)$$

where $\begin{cases} \mathbf{T} = (\Phi^T \Phi + \lambda \mathbf{I})^{-1} \Phi^T \\ \beta = \lambda (\Phi^T \Phi + \lambda \mathbf{I})^{-1} (\Phi^T \Phi)^{-1} \Phi^T \mathbf{z} \end{cases}$, $\mathbf{I}$ is the identity matrix, and $\mathbf{z} = \frac{1}{n} \sum_{i=1}^n \tilde{\mathbf{p}}_i$ is the mean of all PR vectors $\tilde{\mathbf{p}}_i$. Equation (3) defines our hash functions; that is, the real-valued hash code of $\tilde{\mathbf{p}}_i$ is obtained by projecting it onto $\mathbf{T}$ with a shift β .

With all α_i available, the mean vector $\mu = \frac{1}{n} \sum_{i=1}^n \alpha_i$ can be readily computed. We then binarize α_i as follows:

$$\begin{cases} \mathbf{b}_i(j) = 1 & \text{if } \alpha_i(j) \geq \mu_i(j) \\ \mathbf{b}_i(j) = 0 & \text{if } \alpha_i(j) < \mu_i(j) \end{cases} \quad (4)$$

That is, the mean value is used as the threshold for binarization. Note that in our model, the mean μ is actually optimized jointly with dictionary Φ , which directly determines the hash transform matrix $\mathbf{T}$ and shift β . Arranging all the binary codes together will form the hash matrix of PR matrix $\tilde{\mathbf{X}}$, i.e., $\mathbf{B} = [\mathbf{b}_1, \dots, \mathbf{b}_n]$.

2.4 Testing Stage

Given a testing subject $\hat{\mathbf{x}}$, we detect its landmarks using the learnt landmark detection model, remove redundant/noisy landmarks via keeping the indices of selected landmarks, extract LEP features from the selected landmarks, compute the corresponding PR vector and its real-valued code using Eq. 3, and binarize the code using Eq. (4). Finally, the Hamming distance between the binarized code of $\hat{\mathbf{x}}$ and $\mathbf{B}$ is computed to find the nearest neighboring

subjects of $\hat{\mathbf{x}}$. The associated labels of the neighboring subjects are then ensembled to obtain the categorical label of $\hat{\mathbf{x}}$, via the majority voting rule.

3 Experiments

To evaluate the proposed framework, we conducted experiments using two disjoint subsets of Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset. Specifically, we used 199 ADs and 229 age-matched NCs from ADNI-1 as training data, while used 159 ADs and 201 age-matched NCs from ADNI-2 as testing data. To obtain more robust results, we employed 10 repetitions of 10-fold cross validation on the training (ADNI-1) data, to use only subset (*i.e.*, 9 out of 10 folds) of the training data to train the model in each experiment, and testing the model using testing (ADNI-2) data. In each experiment, we conducted nested 5-fold cross-validations, in the grid searching range of $\{10^{-5}, \dots, 10^5\}$, to determine the parameter values of the models (*e.g.*, γ in Eq. (1), and λ in Eq. (2)).

We only evaluate the proposed landmark selection model and the hashing method of our framework, as the landmark detection model used in this work is the same as [11]. Using the landmark detection model in [11], we obtained 1500 landmarks for each MR image, and 50-dimensional LEP features for each landmark, to conduct the following experiments.

3.1 Experimental Results of Landmark Selection

To evaluate our landmark selection model, we conducted landmark selection (via Eq. (1)) using the training set via 10-fold cross-validation for 10 times. Using the results of this experiment, we knew how frequent each landmark was selected.

By setting different thresholds for this frequency values, we can select different number of landmarks. We then used LEP features for all the selected landmarks to train classifiers and reported the results in Fig. 2, in terms of average classification accuracy and testing cost (in time). The numbers in the horizontal axis of Fig. 2, from right to left, are respectively corresponding to the number of landmarks which is at least 100 % (13 landmarks), 95 % (43 landmarks), 90 % (116 landmarks), 85 % (482 landmarks), and 80 % (784 landmarks) being selected in the aforementioned experiment using Eq. (1). From Fig. 2, we found that testing cost is proportional to the number of landmarks used, while the classification accuracy is maximum when around 482 to 784 landmarks are used. Considering the tradeoff between accuracy and testing cost, we chose to use 482 landmarks to represent each subject for the next step.

3.2 Experimental Results of Hashing

In this section, we evaluated our hashing method by comparing its AD diagnosis performance with other state-of-the-art hashing methods, such as Anchor Graph Hashing (AGH) [5], Iterative Quantization (ITQ) [1], Spectral Hashing (SpH) [9], and Locality-Sensitive Hashing (LSH) [4]. Figure 3 shows that our hashing method consistently outperforms other comparison hashing methods, in terms of classification accuracy, using different number of bit lengths.

Figure 3 also includes kNN and SVM results using all the original landmark features (*i.e.*, 75,000-dimensional LEP features), which are 73.89 % and 75.83 % respectively for classification accuracy, and 12.86 (seconds) and 3.05 (seconds) respectively for testing cost. We found that the classification accuracies of three hashing methods (*i.e.*, AGH, ITQ and proposed) with moderate code length (*i.e.*, from 28 bits to 52 bits) are generally better than kNN and SVM, while they are faster than them for more than 100 times.

We also checked whether the top 10 nearest neighbors of all testing subjects have the same diagnosis labels with testing subjects. The results showed that most hashing methods (including ours, AGH, ITQ, and SpH) with code length 36, on average achieved around 79.05 % of accuracy, compared to 68.29 % for kNN.

4 Conclusion

In this paper, we propose a framework of fast neuroimaging-based retrieval and AD analysis, which includes three main steps, *i.e.*, landmark detection, landmark selection, and hashing construction, respectively. Experimental results using the ADNI dataset showed that the proposed framework achieved higher classification accuracy and faster retrieval results, compared to traditional classification methods, such as kNN and SVM. In our future work, we will design new hashing techniques to further improve the performance of neuroimaging-based retrieval and AD analysis diagnosis.

Acknowledgments

This work was supported in part by NIH grants (EB006733, EB008374, EB009634, MH100217, AG041721, AG042599). Xiaofeng Zhu was supported in part by the National Natural Science Foundation of China under grants 61573270 and 61263035.

References

1. Gong Y, Lazebnik S, Gordo A, Perronnin F. Iterative quantization: a procrustean approach to learning binary codes for large-scale image retrieval. *IEEE Trans Pattern Anal Mach Intell.* 2013; 35(12):2916–2929. [PubMed: 24136430]
2. Huang L, Jin Y, Gao Y, Thung K, Shen D. Alzheimer's Disease Neuroimaging Initiative: Longitudinal clinical score prediction in Alzheimers disease with soft-split sparse regression based random forest. *Neurobiol Aging.* 2016; 46:180–191. [PubMed: 27500865]
3. Hutton C, De Vita E, Ashburner J, Deichmann R, Turner R. Voxel-based cortical thickness measurements in MRI. *Neuroimage.* 2008; 40(4):1701–1710. [PubMed: 18325790]
4. Kulis B, Grauman K. Kernelized locality-sensitive hashing for scalable image search. *ICCV.* 2009:2130–2137.
5. Liu W, Wang J, Kumar S, Chang SF. Hashing with graphs. *ICML.* 2011:1–8.
6. Shen D, Davatzikos C. HAMMER: hierarchical attribute matching mechanism for elastic registration. *IEEE Trans Med Imaging.* 2002; 21(11):1421–1439. [PubMed: 12575879]
7. Thung K, Wee C, Yap P, Shen D. Neurodegenerative disease diagnosis using incomplete multi-modality data via matrix shrinkage and completion. *NeuroImage.* 2014; 91:386–400. [PubMed: 24480301]
8. Thung K, Wee C, Yap P, Shen D. Identification of progressive mild cognitive impairment patients using incomplete longitudinal MRI scans. *Brain Struct Funct.* 2015:1–17. [PubMed: 24248427]
9. Weiss Y, Torralba A, Fergus R. Spectral hashing. *NIPS.* 2009:1753–1760.
10. Zhang C, Qin Y, Zhu X, Zhang J, Zhang S. Clustering-based missing value imputation for data preprocessing. *INDIN.* 2006:1081–1086.

11. Zhang J, Gao Y, Gao Y, Munsell B, Shen D. Detecting anatomical landmarks for fast Alzheimer's disease diagnosis. *IEEE Trans Med Imaging*. 2016; PP(99):1.
12. Zhang J, Liang J, Zhao H. Local energy pattern for texture classification using self-adaptive quantization thresholds. *IEEE Trans Image Process*. 2013; 22(1):31–42. [PubMed: 22910113]
13. Zhang Y, Brady M, Smith S. Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm. *IEEE Trans Med Imaging*. 2001; 20(1): 45–57. [PubMed: 11293691]
14. Zhu X, Huang Z, Cheng H, Cui J, Shen HT. Sparse hashing for fast multimedia search. *ACM Trans Inf Syst*. 2013; 31(2):9.
15. Zhu X, Huang Z, Shen HT, Zhao X. Linear cross-modal hashing for efficient multimedia search. *ACM Multimedia*. 2013:143–152.
16. Zhu, X., Suk, HI., Shen, D. A novel multi-relation regularization method for regression and classification in AD diagnosis. In: Golland, P.Hata, N.Barillot, C.Hornegger, J., Howe, R., editors. *MICCAI 2014 LNCS*. Vol. 8675. Springer; Heidelberg: 2014. p. 401–408.p. 51
17. Zhu X, Suk HI, Wang L, Lee SW, Shen D, et al. A novel relational regularization feature selection method for joint regression and classification in AD diagnosis. *Med Image Anal*. 2015
18. Zhu X, Zhang L, Huang Z. A sparse embedding and least variance encoding approach to hashing. *IEEE Trans Image Process*. 2014; 23(9):3737–3750. [PubMed: 24968174]

Fig. 1.

The flowchart of the proposed framework of fast neuroimaging-based AD analysis. The blue arrows and the red arrows, respectively, represent training stage and testing stage. It is noteworthy that training stage involves three main sequential steps, *i.e.*, landmark detection, landmark selection, and hashing. (Color figure online)

Fig. 2.

The classification accuracy (left) and testing cost (right), using different number of selected landmarks, based on k Nearest Neighbors (kNN) and Support Vector Machine (SVM) classification method.

Fig. 3.
The classification accuracy (left) and testing cost (right), of all methods.