Alzheimer's Disease Diagnosis using Landmark-based Features from Longitudinal Structural MR Images – *Supplementary Materials*

Jun Zhang, *Member, IEEE,* Mingxia Liu, *Member, IEEE,* Le An, Yaozong Gao, Dinggang Shen*, *Senior Member, IEEE*

In what follows, we present additional experimental results. Specifically, we first visualize the discriminative landmarks in MCI vs. HC classification and then analyze the learned bag-of-words for spatial feature extraction. Finally, we investigate the dependence of the proposed method on the landmark detection accuracy.

A. Anatomical Landmarks Identified in MCI vs. HC Classification

Besides the anatomical landmarks identified from training subjects in AD vs. NC classification in Fig. 1 and Fig. 2 in the main text, we further plot the landmarks discovered from the training subjects in MCI vs. HC classification, and show them in Fig. S1.

B. Lanmark Detection Accuracy

We further report the accuracy achieved by the landmark detection algorithm [1] adopted in this study, by using manually annotated landmarks. Specifically, we manually annotate 20 landmarks (see Fig. S2) for MR images based on two criteria. First, landmarks are placed at locations that can generally be identified on every individual in the study. Second, the landmarks are scattered throughout the entire brain in different tissues. In our experiment, we use two-fold cross validation to evaluate the detection performance.

In this group of experiments, we compare the landmark detection method adopted in this study with two other landmark detection methods based on (a) affine registration, and (b) classification forest. In the affine registration based method, we linearly align the testing image to all training images using FLIRT in FSL package [2]. Thus, each aligned training image provides a potential landmark position. Finally, we take the mean value of all potential position as the landmark position.

TABLE S1
LANDMARK DETECTION RESULTS ACHIEVED BY THREE DIFFERENT METHODS FOR 20 MANNUAL ANNOTATED LANDMARKS.

Method	Mean Error (mm)
Affine registration based method	3.98 ± 3.37
Classification forest based method	2.65 ± 1.82
Method proposed in [1]	2.41 ± 1.42

In the classification-forest-based landmark detection method, we train a classification forest model for each landmark individually. Specifically, in the training stage, for each landmark, we define a cubic ROI around it and extract patch-based morphological features for all sampled voxels within this cubic ROI ($30 \times 30 \times 30$). We mark the landmark and its neighbors within a $5 \times 5 \times 5$ cubic as positives and the others are negatives to train the classification forest. In the testing stage, we first use the affine registration to get the initial position of the landmark, and then we can define a small ROI around the initialization for the prediction. It means that we consider only the voxels within the ROI. In our implementation, the ROI is defined as a cubic patch centered at the initial position. The side length of the patch is 3 times larger than the standard deviation of this landmark across training images. Finally, the mean position of the classification results is regarded as the landmark location.

The landmark detection results are shown in Table S1. As shown in Table S1, the result of affine registration based method has a relatively large mean detection error and a large standard deviation, since no local information is further considered. The classification forest based method improves detection performance significantly, but both the detection error and standard deviation are still larger than that obtained for the proposed method.

C. Classification Results with Noisy Landmarks

Second, to study the influence of the landmark detection accuracy on the performance of diagnosis, we further design another synthetic experiment. Specifically, we randomly add additional errors to the positions of the detected landmarks. In our experiment, the noise we add follows the uniform distribution. The experimental results are reported in Fig. S3. As shown in Fig. S3, when the error for landmarks increases

∗ Corresponding author.
Jun Zhang, Mingxia Liu, Le An, Yaozong Gao, and *Dinggang Shen are with Department of Radiology and BRIC, University of North Carolina, Chapel Hill, NC, USA (Emails: xdzhangjun@gmail.com, mingxia_liu@med.unc.edu, le_an@med.unc.edu, yzgao@cs.unc.edu, dgshen@med.unc.edu).
Yaozong Gao is also with Department of Computer Science, University of North Carolina, Chapel Hill, NC, USA.
*Dinggang shen is also with Department of Brain and Cognitive Engineering, Korea University, Seoul 02841, Republic of Korea.

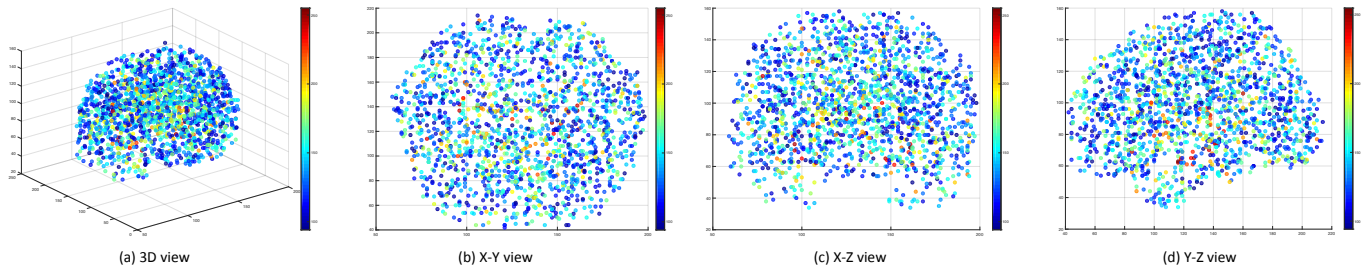

Fig. S1. 3D Illustration of the identified landmarks in MCI vs. HC classification with MCI and HC subjects from ADNI-1. The color illustrates the corresponding p-value in group comparison.

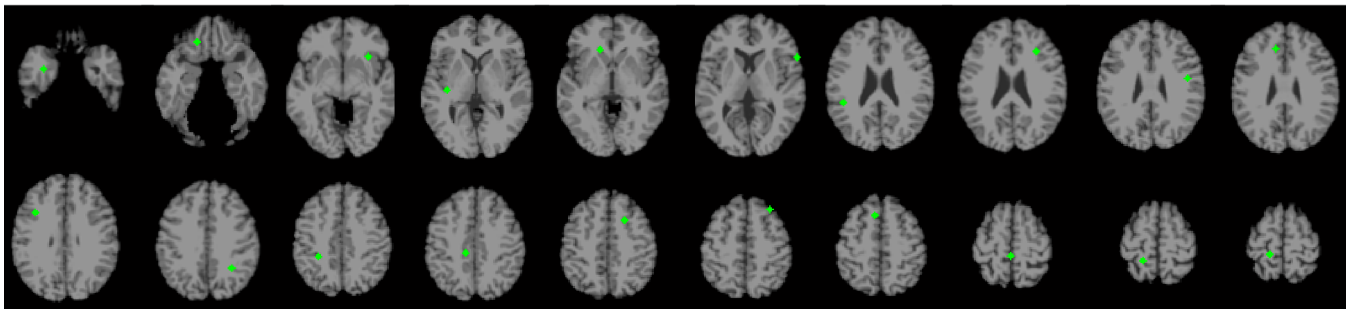

Fig. S2. Illustration of manually annotated landmarks.

from $0\,mm$ to $3\,mm$, the classification accuracy decreases from 88.3% to 86.8% for AD vs. HC classification and from 79.2% to 77.9% for MCI vs. HC classification, respectively. Figure S5 shows that, using Spatial+Longitudinal features, we can obtain reasonable classification accuracy when the error smaller than $5\,mm$. That is because our features are extracted based on local patches that can be relatively robust to the small landmark errors. Moreover, the method using longitudinal features is more sensitive to landmark noise, in comparison to the method using spatial features. It may be because that those longitudinal changes are subtle. However, when generating the Jacobian map, TPS interpolation can rectify some inaccurate landmarks. Therefore, our longitudinal features can also achieve reasonable classification accuracy when the landmark error is not that large. This result indicates that the proposed method is robust to the potential bias of landmarks and is able to obtain accurate diagnosis using the landmarks even with certain noise.

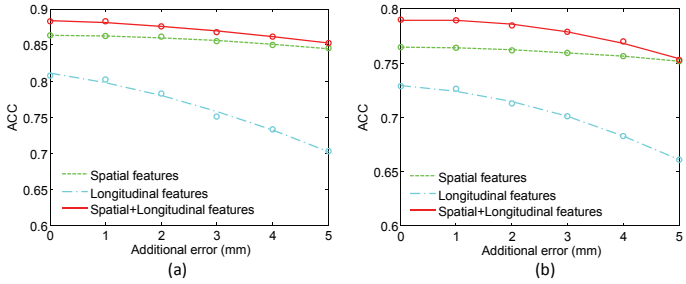

Fig. S3. The classification accuracy achieved by the proposed method with respect to different level of additional landmark detection errors in the classification tasks of (a) AD vs. HC and (b) MCI vs. HC.

D. Learned Visual Words for Spatial Features

In this paper, we extract HOG features to represent local patch appearance. Our visual words are the clustering centers of those HOG features. Specifically, for generating 3D HOG features, we use 9 orientations, $2 \times 2 \times 2$ cells, and a size of $8 \times 8 \times 8$ for each cell. Therefore, the dimensionality of 3D HOG features is 72. In the bag-of-words strategy, the number of clustering centers is set to 50. Therefore, we have 50 visual words for each landmark. In Fig. S4 and Fig. S5, we have reported a part of visual words in both tasks of AD vs. HC classification and MCI vs. HC classification, respectively. From these figures, one could observe that those learned bag-of-words for different landmarks have different distributions. Thus, these visual words can provide supplementary information, which could further promote the final classification performance.

REFERENCES

[1] J. Zhang, Y. Gao, Y. Gao, B. C. Munsell, and D. Shen, "Detecting anatomical landmarks for fast Alzheimer's disease diagnosis," *IEEE Transactions on Medical Imaging*, vol. 35, no. 12, pp. 2524–2533, 2016.

[2] M. Jenkinson, P. Bannister, M. Brady, and S. Smith, "Improved optimization for the robust and accurate linear registration and motion correction of brain images," *NeuroImage*, vol. 17, no. 2, pp. 825–841, 2002.

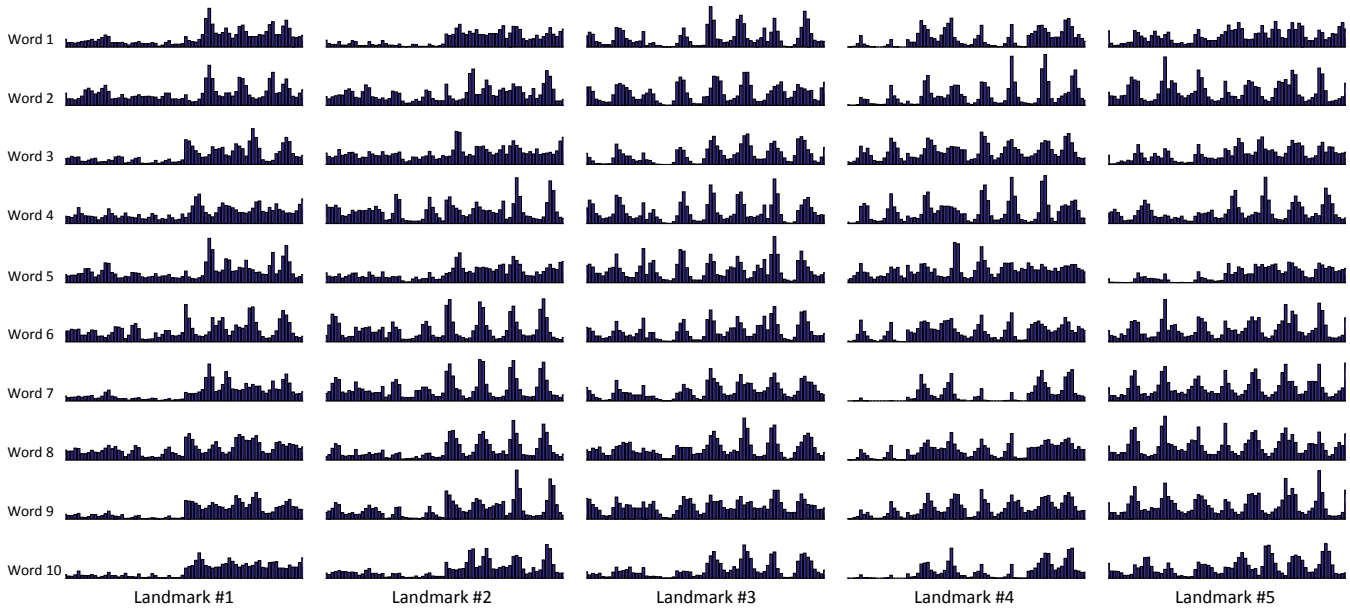

Fig. S4. Example visual words learned in the task of AD vs. HC classification. Each column corresponds to the 10 randomly selected visual words for the specific landmark.

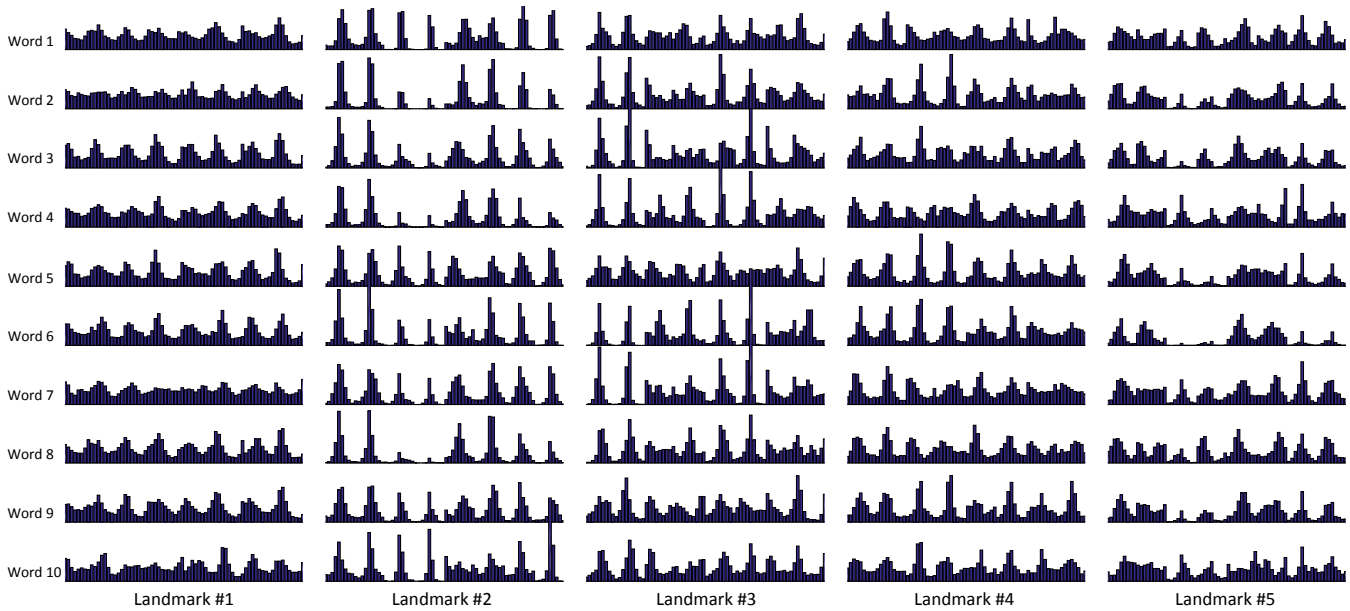

Fig. S5. Example visual words learned in the task of MCI vs. HC classification. Each column corresponds to the 10 randomly selected visual words for the specific landmark.