HHS Public Access

Author manuscript

Hum Brain Mapp. Author manuscript; available in PMC 2016 February 08.

Published in final edited form as:

Hum Brain Mapp. 2014 January ; 35(1): 38–52. doi:10.1002/hbm.22149.

Feed-forward Hierarchical Model of the Ventral Visual Stream Applied to Functional Brain Image Classification

DB Keator[1,2], **JH Fallon**[2], **A Lakatos**[2], **CC Fowlkes**[1], **SG Potkin**[2], and **A Ihler**[1] **for the Alzheimer's Disease Neuroimaging Initiative**

[1]Department of Computer Science, University of California, Irvine, CA. USA

[2]Department of Psychiatry and Human Behavior, University of California, Irvine, CA. USA

Abstract

Functional brain imaging is a common tool in monitoring the progression of neurodegenerative and neurological disorders. Identifying functional brain imaging derived features that can accurately detect neurological disease is of primary importance to the medical community. Research in computer vision techniques to identify objects in photographs have reported high accuracies in that domain, but their direct applicability to identifying disease in functional imaging is still under investigation in the medical community. In particular, Serre et al. (Serre, et al. 2005) introduced a biophysically inspired filtering method emulating visual processing in striate cortex which they applied to perform object recognition in photographs. In this work, the model described by Serre et al. is extended to 3D volumetric images to perform signal detection in functional brain imaging (PET, SPECT). The filter outputs are used to train both neural network and logistic regression classifiers and tested on two distinct datasets: ADNI Alzheimer's disease 2-deoxy-D-glucose (FDG) PET and National Football League players Tc99m HMPAO SPECT. The filtering pipeline is analyzed to identify which steps are most important for classification accuracy. Our results compare favorably with other published classification results and outperform those of a blinded expert human rater, suggesting the utility of this approach.

Corresponding Author: David B. Keator, University of California, Irvine, Department of Computer Science, Department of Psychiatry and Human Behavior, Irvine Hall, rm. 163 - ZOT 3960, Irvine, CA. 92697, Ph: 949-824-7870, dbkeator@uci.edu.
Alexander Ihler, University of California, Irvine, Department of Computer Science, Donald Bren Hall 4066, Irvine, CA. 92697, Ph: 949-824-3645, ihler@ics.uci.edu
Charless C. Fowlkes, University of California, Irvine, Department of Computer Science, Donald Bren Hall 4076, Irvine, CA. 92697, Ph: 949-824-6945, fowlkes@ics.uci.edu
Steven G. Potkin, University of California, Irvine, Department of Psychiatry and Human Behavior, 5251 California Ave., suite 240, Irvine, CA. 92617, Ph: 949-824-8061, sgpotkin@uci.edu
James H. Fallon, University of California, Irvine, Department of Psychiatry and Human Behavior, 5251 California Ave., suite 240, Irvine, CA. 92617, Ph: 949-856-3111, jfallon@es.nacs.uci.edu
Anita Lakatos, University of California, Irvine, Department of Psychiatry and Human Behavior, 5251 California Ave., suite 240, Irvine, CA. 92617, Ph: 949-824-4423, alakatos@uci.edu

**Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.ucla.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.ucla.edu/wp-content/uploads/how_to_apply/ADNI_Authorship_List.pdf

Keywords

Object recognition; Gabor filters; template matching; classification; brain imaging; PET; SPECT; ADNI; Alzheimer's Disease; NFL

INTRODUCTION

Significant progress has been made in the diagnostic decision-making processes and in predicting the onset and the course of brain disorders (Kantarci and Jack Jr 2003; Lovestone 2010; Rachakonda, et al. 2004; Roe, et al. 2011). The traditional endpoint diagnosis, clinical measurements and cognitive tests used in clinical trials have proved to be informative but have their own limitations in accurately quantifying the progression of brain disorders in an unbiased and objective manner (Borroni, et al. 2007; Knopman, et al. 2001). Advances in brain imaging technologies have enabled researchers to investigate and test novel biomarkers that could serve either as diagnostic tools to aid clinical decision-making or as surrogates, reflecting disease progression and underlying disease pathology (Biomarkers Definitions Working Group, 2001). Accordingly, there is a growing body of evidence in the literature showing that structural and functional brain imaging can be valuable tools for predicting and classifying gradually progressive neurological and psychiatric disorders such as Alzheimer's disease (AD) (Drzezga 2009; Kawachi, et al. 2006; Mosconi, et al. 2006; Nordberg, et al. 2010; Tartaglia, et al. 2011). Although both **PET and MRI** imaging modalities have been found to be discriminative in various neurological disorders, there is disagreement in the community about which are most sensitive for particular disorders. Specifically, differences in sensitivity and specificity of structural Magnetic Resonance Imaging (MRI) and 2-deoxy-D-glucose (FDG) Positron Emission Tomography (PET) features in the prediction of early AD has been debated in the literature with no clear consensus (De Santi, et al. 2001; Mosconi, et al. 2006). Nevertheless, AD research studies evaluating the diagnostic and predictive value of regional specific glucose metabolic rate and volume changes suggest the greater reliability of FDG PET over MRI in discriminating AD from subjects with intact and mild cognitive impairment (De Santi, et al. 2001; Kawachi, et al. 2006; Mosconi, et al. 2006). However, De Santi and Mosconi indicate image post-processing influences the outcome of discriminative analyses and subsequently, their predictive value.

Although advances in imaging have enabled researchers to visually inspect both functional and structural brain scans of disease, it is often difficult for the human observer to identify the subtle differences in the brain images that are often necessary for reliable disease classification. Furthermore, visual identification of brain diseases by a human observer is time consuming and error prone. Automated image analysis algorithms that can reliably discriminate the diseased from the healthy brain are preferred because they save time, are generally less prone to errors, are not influenced by rater bias **or** inter-rater differences in neuroanatomical expertise, and can identify subtle statistical correlations in the data. For preventative and longitudinal studies in large populations, automated image analysis is critically important to evaluate the data. To achieve automated and reliable image analysis

and classification, we can use computer vision techniques that are designed to extract information from images.

Object recognition in images and video is an active area of research in the computer vision community. Finding objects is fundamentally related to pattern recognition where the presence of unique patterns of colors, edges, and/or textures are consistent with a particular class of object. Probabilistic models are particularly well suited for recognition problems because they provide a structured approach to modeling uncertainty and can be less sensitive to noise in the data. Object recognition systems often consist of a feature extraction component and a classifier. The feature extractor is used to identify properties of the objects that are most important in discriminating one object from another. The features along with a labeled training set are then used to train a classifier to map the features into a class label for each object the detection system is built to recognize. Although the overall process is simple, there are many subtleties in real world applications of detection systems such as object illumination, scale, occlusion, and orientation that affect accuracy. Most often we have a small set of images representing the objects to be recognized and do not have exhaustive examples at all possible scales, orientations, illuminations, etc. The challenge is therefore to find a feature space that avoids irrelevant variations in the objects and instead captures the most discriminating characteristics (Forsyth and Ponce 2002).

One source of inspiration for engineering such invariant features is the primate visual system, which performs object detection robustly across a huge range of viewpoints, illuminations and occlusions. One very successful method, the Scale Invariant Feature Transform (SIFT) proposed by Lowe (Lowe 1999) uses features with partial invariance to local variations in scale and illumination, similar to the receptive fields of the neurons in the inferior temporal cortex, an area important for object recognition in primates. Serre et al. introduced a filtering method whose hierarchical architecture was designed specifically to emulate visual processing in the cat and primate striate cortex. They applied this method to detecting objects in photographs and reported high success rates from a few training examples. Mutch et al. (Mutch and Lowe 2006) reported similar performance results using a similar filtering scheme that scaled the input images instead of the filters as was done in the Serre work.

Similar to object recognition in photographs, for automated image-based diagnosis, it is necessary to ignore some classes of variation across healthy individuals while identifying other specific variations which are indicative of disease state. Differences in ligand uptake in the brain measured by functional brain imaging modalities such as FDG PET and Tc99m HMPAO Single Photon Emission Tomography (SPECT) result in spatially smooth patterns of differing intensities which can be used to differentiate a disease group from healthy subjects. Similarly, precise morphology/anatomy may vary among individuals requiring some degree of local scale and orientation invariance. Based on this insight, we extend the neurologically-inspired filtering model described by Serre et al. to signal detection in functional brain imaging. To evaluate how well the Serre feature model works in capturing disease patterns in the human brain, the model is extended to 3D volumetric space and signal detection differentiation in functional brain imaging. The hierarchical filtering pipeline is analyzed to identify which steps are most important for classification accuracy

and the filter outputs are used to train both neural network (NN) and logistic regression (LR) classifiers. Two distinct and previously published datasets are tested using this feature extraction and classification method: (1) Alzheimer's Disease Neuroimaging Initiative (ADNI) AD FDG PET scans sampled at baseline, 12 month, and 24 month time-points versus the study specific age-matched healthy comparison (HC) subjects (Mueller, et al. 2008); (2) a Tc99m HMPAO SPECT National Football League (NFL) dataset versus study specific age-matched HC subjects (Amen, et al. 2011). The AD classification results are further compared against a blinded expert human rater (co-author J.H. Fallon), providing a baseline measure of how well a human counterpart can recognize disease in the same dataset.

METHODS

2.1 Filtering and Feature extraction

The image filtering pipeline consists of a series of alternating steps of simple filtering (S layers) and complex filtering (C layers) layers briefly summarized here and discussed in detail in subsequent sections. The first simple layer (S1) outputs respond to oriented edges at different spatial scales and orientations (section 2.2). Spatial scales in this context refer to the underlying spatial distribution of the signal in the images. Filters with larger spatial scales will respond to larger (spatially) image signals. S1 layer filters are separated into "bands" where each band is composed of two similar spatial scales as shown in table 1, rows 1 and 2. The first complex layer (C1) combines the outputs from the S1 layer at different scales but within orientations, providing scale invariance (section 2.3). The complex layers pool the simple layer outputs using a max operator, where the strongest simple layer output drives the complex layer output. The second simple layer (S2) matches the detections from the C1 layer against healthy subjects in a template matching framework where higher scores indicate a closer match (sections 2.3.1 and 2.4). The second complex layer (C2) combines the outputs from template matching scores across orientations gaining invariance to orientation (section 2.5).

2.2 S1 Layer

The S1 layer is computed by applying sixteen orientated 3D Gabor filters at orientations $\theta \in \{0, \pi/4, \pi/2, 3\pi/4\}$, $\varphi \in \{0, \pi/4, \pi/2, 3\pi/4\}$, and wavelength λ to each brain scan in the dataset. A Gabor filter is a linear filter whose impulse response is a harmonic function multiplied by a Gaussian function:

$$G(x,y,z)=\frac{1}{(2\pi)^{3/2}\sigma^3}exp\left(-\frac{1}{2}\left(\frac{x^2+y^2+z^2}{\sigma^2}\right)\right)cos\left(\frac{2\pi}{\lambda}\left(xcos(\theta)sin\left(\phi\right)+ysin(\theta)sin\left(\phi\right)+zcos(\phi)\right)\right) \quad (1)$$

The cosine term in equation (1) controls the harmonic component through the λ wavelength parameter. The variables x, y, and z are the spatial variables defining the spatial extent of the filter. The standard deviation σ describes the size of the Gaussian envelope. The orientation of the filter is represented by variables θ and φ, where φ orients the filter in the x–y plane and θ is the orientation from the positive z axis. For a detailed description of 3D Gabor filters, refer to Bau et al. (Bau and Healey 2009; Bau, et al. 2008). Frequency and orientation

representations of the filter are similar to those of the human visual system. The original Serre method performed Gabor filtering in 2D, consistent with the image matrix of photographs. In this work, the Gabor filtering was performed in 3D and applied using filter sizes, sigmas, and lambdas over a series of eight bands. The parameters of each band are listed in Table I, rows 2–4. The filter sizes and parameters were kept essentially the same as the Serre work, but the spatial extents of the bands were decreased in order to make the features more sensitive to small activation differences in functional brain imaging. The relative proportions between sizes across the bands remained the same. The voxel sizes of the functional brain imaging data used in this study were $2mm^3$ per voxel (see section Materials/Methods for a detailed description of the test data). The smallest filter size in the Serre work (7 pixels) if directly applied as 7 voxels would be unlikely to respond to small differential signals that could be discriminative in the context of functional imaging and disease. To avoid missing small signals, the lowest filter band was set to 3 voxels. An example of the AD PET scan slices filtered with the 3D Gabor functions are shown in Figure I. Oriented signals are indeed differentially selected by the filters, consistent with our hypothesized responses of the filters when applied to functional brain imaging data.

2.3 C1 Layer

The C1 layer combines incident S1 units of the same θ and φ orientations, creating tolerance to size and shift within Gabor filter orientation. Complex cells in the hierarchical visual cortex model have larger receptor fields than the S1 layer (Serre, et al. 2005). To operationalize this relationship, the S1 layer volumes are filtered with a max operator over Gabor filter scales (Table I, row 1(filter)), but within each orientation band (columns of Table I). Max filtering is a nonlinear image processing technique where the value at each voxel in the filtered image is the maximum of the input image voxels in a local neighborhood defined by the filter size. The filter size over which the maximums are calculated depends on the Gabor filter size (shown in Table I, row 4 (max grid)). Gabor filters with larger spatial scales will respond more strongly to larger (spatially) signals in the images at the same θ and φ orientations, therefore, the corresponding max filter sizes should be tuned accordingly. These operations are performed for each Gabor orientation and for each band resulting in 16*8 volumes, representing maximums over scales but within orientations. Due to the large numbers of voxels in the volumes and thus the large numbers of max operations over increasing neighborhoods, we used the algorithm developed by Van Herk et al. (Van Herk 1992) to efficiently compute the maximums over neighborhoods for each voxel in the S1 layer volumes. The method requires only a small number of operations per voxel to compute the maximums and lowers the computational time of this stage of the processing pipeline.

2.3.1 C1 Layer Training Patches—Template matching is a common approach to object recognition in computer vision systems. It is a technique which matches image regions to stored representative templates using a specific scoring function (Brunelli 2009). In this work, representative templates were collected on a random subset of hold-out healthy subjects to be used in the subsequent S2 layer template-matching step. Ten randomly selected hold-out training images were chosen for template extraction. Templates were extracted randomly across these training images and from random locations within the

images but constrained to fall within the boundaries of user specified regions of interest (see section 3.1). The regions from which templates are randomly sampled are completely user defined and could be chosen based on some a-priori hypothesis or from the literature. Selecting templates from specific regions of interest in the brain is similar to learning that a car is characterized by particular features in spatial locations, e.g. rides on four tires, has doors on the sides, a hood on the front, etc.

Operationally, the user selects regions of interest and the number of features prior to pipeline execution. We uniformly divide the number of random locations across the number of regions of interest. To generate the random voxel locations within a region of interest, we use an atlas labelmap, which assigns a numerical code to each atlas region. Each atlas region is therefore defined by all the contiguous voxels in the labelmap volume that have equal numerical codes. From this information, we can find the cube containing this region. We then use rejection sampling: drawing a random point uniformly within the cube, we accept it if it falls within the bounded region; otherwise we reject and try again. This process continues until the required number of locations has been found for each region. In our experiments, 50 or 100 templates were chosen to describe the low level representation of the brain images. We chose the two sets such that we had a reasonable number of templates per region of interest selected and so we could evaluate the dependence of the classification results on the number of feature scores used. The original Serre work suggests a modest dependence of performance on the number of feature scores used. For each selected template location, 5^3, 9^3, 13^3, and 17^3 voxel “patches” were extracted from each of the 16 Gabor filtering orientations and bands from the C1 layer of the ten randomly selected hold-out healthy subject training images. These patches are simply contiguous sets of voxels of differing spatial extents (5^3, 9^3, 13^3, and 17^3) centered on the template location and effectively give the vision system a “memory” of image feature examples from the functional brain images of healthy subjects.

2.4 S2 Layer

The S2 layer corresponds to the template-matching phase of the pipeline. For each C1 image in the test dataset and for each template patch collected from the hold-out healthy subject data, we compare the Gaussian radial basis function score shown in equation (2) for each band independently. The S2 unit’s response depends on the Euclidean distance between the test dataset patch (X) and the stored prototype patch (*P*) sampled at the same location, scale, and orientation. If the functional activity profile in the test data is identical to the stored template patch, the score equals 1 whereas if the differences from the stored template patch are large, the score approaches 0. The parameter γ normalizes for different patch sizes ($n \in \{5, 9, 13, 17\}$) when computing the score in equation (2). The parameter γ is fixed to $(n/5)^3$ where n is the patch size and the denominator is the smallest patch size. The parameter σ in equation (2) is the uncertainty or variance in the stored prototype patch (*P*). This parameter was set to 1 in all experiments. Alternatively, it could be set to the empirical variance of the training prototype patches discussed in section 2.3.1.

$$F\left(X_{\theta,\phi}, P_{\theta,\phi}\right) = exp\left(-\frac{\left\|X_{\theta,\phi} - P_{\theta,\phi}\right\|^2}{2\sigma^2\gamma}\right) \quad (2)$$

2.5 C2 Layer

The final layer in the pipeline computes the maximum response of the S2 layer scores from all bands and orientations for each prototype template. The final feature sets therefore consist of 50 or 100 shift and scale invariant scores (i.e., for 50 and 100 prototype patches) that are subsequently used for classification. Conceptually, for each test image, for each prototype template patch sampled from a brain region, we are using the score that indicates the best match between the test image and a healthy subject regardless of signal size and orientation. We expect that subjects with neurological disorders will match less well with the healthy subjects and thus have a lower score. The final size of the feature vector therefore depends only on the number of patches extracted during training and not on the number of voxels in the full three-dimensional brain image. This allows the user to balance the number of template patches sampled during patch selection (i.e. number of features) and the number of subjects available in the dataset. Flexibility in choosing the number of features provides insulation from classifier over-fitting, which can occur if the number of features greatly exceeds the number of examples.

3. Evaluation

We used two datasets to evaluate the approach. Both are functional imaging datasets but distinctly different modalities. We selected these datasets to evaluate the generality of this approach and its application to distinctly different neurological abnormalities.

3.1 The Alzheimer's Disease Neuroimaging Initiative (ADNI)

ADNI was launched in 2003 by the National Institute on Aging (NIA), the National Institute of Biomedical Imaging and Bioengineering (NIBIB), the Food and Drug Administration (FDA), private pharmaceutical companies, and non-profit organizations as a $60 million, 5-year public–private partnership. The primary goal of ADNI is to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, such as cerebrospinal fluid (CSF) markers, APOE status and full-genome genotyping via blood sample, and clinical and neuropsychological assessments can be combined to measure the progression of mild cognitive impairment (MCI) and AD. Determination of sensitive and specific markers of very early AD progression is intended to: (1) aid in the development of new treatments, (2) increase the ability to monitor their effectiveness, and (3) reduce the time and cost of clinical trials. The principal investigator of the initiative is Michael W. Weiner, M.D., of the Veteran's Affairs Medical Center and University of California, San Francisco. ADNI is the result of efforts of many co-investigators from a broad range of academic institutions and private corporations, and participants have been recruited from over 50 sites across the U.S. and Canada. ADNI participants range in age from 55 to 90 years and include approximately 200 cognitively normal elderly followed for three years, 400 elderly with MCI followed for three years, and 200 elderly with early AD followed for

two years. Participants are evaluated at baseline, 6, 12, 18 (for MCI only), 24, and 36 months (AD participants do not have a 36 month evaluation). Baseline and longitudinal follow-up structural MRI scans are collected on the full sample and 11C-labeled Pittsburgh Compound-B (11C-PIB) and FDG PET scans are collected on a subset every 6–12 months (for study details see http://www.adni-info.org). A subset of these data were published in (Mueller, et al. 2008) and (Langbaum, et al. 2009) was used in this analysis.

3.2 AD Dataset

The dataset used in this study consisted of 154 baseline FDG PET scans acquired as part of the ADNI study and published in (Mueller, et al. 2008) and (Langbaum, et al. 2009). There were 82 HC subjects (Mini-Mental State Exam (MMSE) 28.6±1.1; Age 75.1±9.6 yrs) and 72 AD subjects (MMSE 23.2±3.5; Age 75.1±11.2 yrs) from the baseline ADNI sample used for this study. The 12m and 24m ADNI samples contained a subset of the baseline dataset due to subject dropout. The 12m sample included 72 HC subjects (MMSE 29.2±1.2; Age 77.5±8.4 yrs) and 61 AD subjects (MMSE 20.9±4.9; Age 75.4±11.8 yrs). The 24m sampled included 68 HC subjects (MMSE 28.6±3.7; Age 76.0±10.2 yrs) and 33 AD subjects (MMSE 18.4±6.1; Age 74.6±15.2 yrs). The acquisition protocol consisted of collecting six five-minute frames 30–60 minutes post ^{18}FDG-injection. During the uptake period subjects were asked to rest comfortably in a dimly lit room with their eyes open. The collected frames were registered to the first frame (acquired at 30–35 min post-injection) and averaged to yield a single 30 minute average PET image in “native” space. The image matrix, field of view, and resolution of the datasets from participating sites were then matched by the ADNI group. The images were spatially normalized to the MNI atlas using SPM8 software (2007) resulting in image matrices of 79 × 95 × 68 voxels in x, y, and z dimensions respectively with isotropic 2^3mm voxel sizes. The Automated Anatomical Labeling (AAL) atlas was used to constrain the region of interest selection based on the anatomical parcellations available in the atlas (Tzourio-Mazoyer, et al. 2002). The AAL atlas used to define the region of interest boundaries is consistent with the space defined by the MNI atlas.

Coordinates for template patch sampling and S2 layer matching scores were constrained to fall within regions identified in the literature to be affected by AD (see sections 2.3.1 and 2.4). Delacourte et al. identified stages of AD neurofibrillary degeneration in patients of various ages and different cognitive statuses (Delacourte, et al. 1999). Further, Langbaum et al. (Langbaum, et al. 2009) identified regions of reduced metabolic rates in AD. Regions included the cingulate cortex, parietal and temporal lobes, among others. For this study, we chose AAL atlas regions (bilateral): anterior and posterior cingulum, temporal lobes (middle), hippocampus, amygdala, thalamus, frontal and orbital cortices (superior and middle), temporal pole (superior, middle, inferior), and parietal lobe (inferior) as being consistent with published findings on potentially discriminative regions.

3.3 NFL Dataset

The NFL dataset used in this study consisted of 162 technetium-99m hexamethylpropyleneamine oxide (Tc99m HMPAO) SPECT scans acquired for a study evaluating the impact of playing American football by Amen et al. (Amen, et al. 2011). There were 83 HC (Age 41.7±17.8 yrs) and 79 NFL (Age: 57.5±11.5 yrs) subjects. Subjects

were injected with an age/weight appropriate dose of Tc99m HMPAO and performed the Conners' Continuous Performance test II for 30 minutes during uptake. All subjects completed the task and were subsequently scanned on a high-resolution Picker Prism 3000 triple-headed gamma camera with fan beam collimators. The original reconstructed image matrices were 128×128×29 voxels with sizes of 2.16mm ×2.16mm × 6.48mm. The images were spatially normalized to the MNI atlas using SPM8 software (2007) resulting in image matrices of 79 × 95 × 68 voxels in x, y, and z dimensions respectively with isotropic 2^3mm voxel sizes. Images were smoothed using an 8mm FWHM isotropic Gaussian kernel. The pre-processing steps were identical to the previously published work by Amen et al. In the previously published work, a subset of the HC dataset was used and matched on gender and race. For this work, all subjects were used regardless of race and gender.

Coordinates for template patch sampling and S2 layer matching scores (see sections 2.3.1 and 2.4) were constrained to fall within regions identified in Amen et al. as the top discriminating regions for the NFL group. To our knowledge, the Amen study was the first brain imaging study evaluating NFL players and as such, the regions were picked based only on that publication. For this study, we used AAL atlas regions (bilateral): anterior and posterior cingulum, frontal pole, hippocampus, amygdala, and temporal pole (middle and inferior).

3.4 Ethics

The NFL study was approved by each of the participating sites' Institutional Review Boards (IRBs) and complied with the Code of Ethics of the World Medical Association (Declaration of Helsinki). Written informed consent was obtained from all participants after they had received a complete description of the studies.

The ADNI data was previously collected across 50 research sites. Study subjects gave written informed consent at the time of enrollment for imaging and genetic sample collection and completed questionnaires approved by each participating sites' Institutional Review Board (IRB).

3.5 Feature Sets

In order to identify which components of the feed-forward hierarchical model implemented in this study were most important in correct classification, three separate feature sets were computed. The FTM (Gabor filter + template match) feature set is the result of the full hierarchical pipeline as described in section 2. In order to understand the effect of the Gabor filtering, the TM (template matching) dataset was created using the same procedures outlined in section 2 without Gabor filtering. More precisely, the dataset consists of selecting template patches from the un-filtered images (neither S1 nor C1 layers) and performing the computations in the S2 and C2 layers. To evaluate the effect of template matching, the AP (average patch) feature set consists of simply averaging the voxels in the neighborhood around the prototype patch locations selected in section 2.3.1, across the various filter sizes (Table I, row 2) and taking the maximum response.

To compare the feature sets of the hierarchical model with more typical data reduction techniques, the maximum group difference (MaxT) and data reduction (DR) sets were

computed. The MaxT feature set is computed by performing a typical voxel-wise independent, 2 sample t-test in the SPM8 software. The resulting SPM(t) maps were then thresholded at $p<0.01$ (AD baseline), $p<0.001$ (AD 12m), $p<0.001$ (AD 24m), and $p<1e-6$ (NFL) and corrected for multiple comparisons using the family-wise error rate (FWE) correction. Probability thresholds were chosen to limit the number of voxels in the resulting t-score maps such that similar numbers of voxels were obtained for each data set (~3K points). The absolute values of the resulting t-scores were ranked and the data from the top 50 and 100 locations were then sampled from each subject and used for classification (MaxT). The DR feature set used all the locations found in the group difference maps, discussed above, after probability thresholding (~3K points), sampling the original data at those locations (~3K points) for each subject. The resulting N×K matrix (N subjects, K sampling locations) was mean-centered for each column K and run through principal components analysis (PCA). Each subject's data was then projected onto the eigenvectors of the top 50 and 100 largest eigenvalues from the PCA decomposition giving a low dimensional representation with 50 or 100 feature scores that were subsequently used for classification. The top 50 and 100 largest eigenvectors were chosen so that the projected dataset contained 50 and 100 scores per subject, consistent with the number of feature scores calculated from the full feed-forward hierarchical model.

3.6 Classification

Classification was done using both a multilayered perceptron neural network (NN) and a logistic regression (LR) classifier to understand the dependence of the results on the classifier chosen (Hall, et al. 2009). Each classifier was trained separately on the same datasets to compare the performance of the simpler logistic regression classifier, able to find linear decision boundaries, with the neural network classifier, able to model more complex nonlinear functions. The neural network was constructed with one hidden layer (hidden layer nodes = (#features + #classes)/2) and trained with a learning rate of 0.3 and a momentum of 0.2. For each classifier, ten-fold cross validation was used. The dataset was divided in each fold into training and testing subsets. The classifier was trained using the training subset and tested on the testing subset. This process was repeated ten times. Areas under the receiver operating characteristic (ROC-AUC) curves were computed from the probability of class membership of the testing data from each of the trained classifiers. The full filtering pipeline ROC-AUC curves were statistically compared to each of the alternative methods for each dataset and classifier using the DeLong et al. (DeLong, et al. 1988) method of comparing areas under correlated ROC curves as implemented in the pROC package (Robin, et al. 2011). To compute 95% confidence intervals and statistics, the data was resampled 2000 times, stratified by group membership.

To compare the classifier results on the baseline AD dataset with the visual ratings of neuroanatomist and co-author J.H. Fallon, true positive (TP) and false positive (FP) rates were calculated. To calculate the TP and FP rates, the probability of class membership from the trained classifiers for each testing subset data point, in each fold, was computed. The data point was assigned to the class with the largest probability. The TP rate was the proportion of examples in the testing subsets that were classified as class AD, among all testing examples that were originally labeled as class AD. The TP rate is the average across

all folds. The FP rate was the proportion of examples in the testing subsets that were classified as class AD, but were originally labeled as the alternative class, among all testing examples which are not of class AD. The FP rate is the average across all folds. The TP and FP results for Dr. Fallon were computed from his designation of either AD or healthy control for each of the baseline data compared to the original class labels.

RESULTS

To summarize the performance of each classifier, the ROC-AUC results for the Alzheimer's disease (AD) baseline, 12m, and 24m datasets are shown in Figures IIa, IIb, and IIc respectively for 50 and 100 feature datasets and both logistic regression and neural network classifiers. The confidence intervals for each ROC-AUC and statistical comparisons of the full filtering pipline (FTM) with each of the other methods for all classifiers and datasets are shown in Tables IIa, IIb, and IIc. The FTM method outperformed the other methods in terms of ROC-AUC in 80% of the tests, and was statistically better in 35%. No other method was statistically better than FTM; although, the PCA data reduction strategy (DR) in the 50-feature, baseline AD, logistic regression classifier was close ($p<0.064$). Overall, the neural network classifier generally outperformed the logistic regression classifier in ROC-AUC. Further, the FTM method was statistically better than all other methods in 46% of the neural network classification experiments compared to 25% using the logistic regression classifier, suggesting a benefit of using the more sophisticated classifier with the FTM method. There was a small, non-significant, increase on average in ROC-AUC over all the classifiers in the results using the larger 100 feature datasets. Overall performance of the FTM trained classifiers were consistent with other published classification results (see Discussion) using the ADNI dataset, with maximum ROC-AUC at baseline of 0.962±0.025 (neural network, 100 feature), at 12m of 0.837±0.073 (neural network, 100 feature), and at 24m of 0.878±0.070 (neural network, 100 feature).

Neuroanatomist and co-author J.H. Fallon was given the baseline AD dataset images in transaxial, coronal, and sagittal orientations, without the diagnosis and given no practice set of normal or ADs to examine prior to the analysis, and asked to classify the scans as either AD or HC. These results are only available for the baseline AD data due to the significant effort in manually rating so many scans. Dr. Fallon achieved a true/false positive rate for AD of 0.718/0.380 and for the HC group of 0.671/0.244 as shown in Table IV. The FTM classifier performed better in both true/false positives for both AD and HC groups while also outperforming the maximum group difference (MaxT) and data reduction (DR) methods, further suggesting the potential utility of this approach.

The AUC results for the NFL group are shown in Figure III for 50 and 100 feature datasets and both logistic regression and neural network classifiers. The confidence intervals for each ROC-AUC and statistical comparisons of the FTM with each of the other methods for all classifiers and datasets are shown in Table III. Interestingly, unlike the AD dataset, the FTM method did not dominate the others, outperforming the other methods in 44% of the tests and was statistically better in only one. Alternatively, the MaxT method consistently outperformed the others in terms of ROC-AUC and was statistically better than the FTM method in three out of four comparisons. We speculate this result is related to specific brain

functional changes accompanying repeated head injuries evident in the NFL dataset (see Discussion). Overall performance of the FTM classifier was still quite good with maximum ROC-AUC of 0.939±0.037/0.145 (logistic regression, 100 features). Unlike the AD experiments, the neural network classifier did not outperform the logistic regression classifier for the FTM dataset but did for the best performing MaxT dataset.

DISCUSSION

The overall classification results suggest the biophysically inspired feed-forward hierarchical model used in these experiments is sensitive to differences in functional brain imaging data. Both AD and NFL classification experiments showed impressive ROC-AUC rates using a method not specifically tuned for these imaging modalities. The full filtering pipeline (FTM) results are consistent with published classification rates for the ADNI AD data set using brain imaging; although, most reported results use a mix of structural and functional imaging features. For example, Hinrichs et al. used the ADNI dataset in a spatially augmented boosting framework and reported an ROC-AUC of 0.8716 when using just FDG PET (Hinrichs, et al. 2009).

A benefit of using logistic regression classifiers is the clear interpretation of which features are most informative for classification. For baseline AD classification, the four most informative patches (highest weights) were sampled from AAL atlas regions right hippocampus and superior temporal lobes left and right while the posterior cingulate, a region commonly associated with disease progression, ranked fourth. For 12m AD classification the most informative patches were sampled from frontal superior right, frontal superior orbital left, and the temporal pole superior right. For 24m AD classifications the most informative patches were sampled from the frontal superior right, temporal pole mid left, and hippocampus left. It is interesting that the frontal lobe was not included as a top discriminating location in the baseline data set but was in both the 12m and 24m data, consistent with well-known structural changes in AD disease progression. We also evaluated the performance of the FTM features using ROIs that specifically did not include those selected in section 3.2. The results were on average 10–15% lower in ROC-AUC for baseline AD than those reported in the results section, suggesting this method is sensitive to region of interest selection. Therefore we suspect the filtering pipeline could be used to test competing hypotheses about specific regions of interest implicated in disease. The top three most informative patches from the features evaluated using ROIs that specifically did not include those selected in section 3.2, were sampled from AAL atlas regions frontal inferior orbital left, insula right, and occipital middle right. Other informative patches for AD included the supramarginal right, lingual right and frontal inferior operculum left. Interestingly, the frontal inferior orbital, operculum, and the supramarginal gyrus are all associated with AD in the literature suggesting the classification results are still picking up on areas related to the disease (Espasy and Jacobs 2010; Grignon, et al. 1998).

Overall, the average patch (AP) feature set outperformed the template matching (TM) feature set, suggesting no compelling benefit of template matching without Gabor filtering in this application. The utility of oriented Gabor filtering and template matching in deriving the feature set was most evident in AD classification. This trend was not observed for the

NFL classification experiments. Why would oriented filtering improve classification rates in AD and not the NFL data set? It is well known in the literature that structural changes in AD follow an anatomical trajectory starting in entorhinal cortex and hippocampus, then moving to temporal and parietal lobes, and finally affecting the frontal lobes in late stage AD (Braak and Braak 1997; Hua, et al. 2008; Thompson, et al. 2003). These structural changes should be reflected in corresponding functional changes. In addition, the accumulation of amyloid plaques between nerve cells in the brain is known to be a hallmark of AD. Both the structural changes and plaques may be altering the functional brain imaging derived signal in orientation, scale, and localized spatial extent due, in part, to brain plasticity and compensation.

Alternatively, the full filtering pipeline might not perform as well in data sets with widespread, global functional changes observed in the NFL data. Indeed the manuscript by Amen et al. reports “significant decreases in regional cerebral blood flow were seen across the whole brain”. The comparison feature sets MaxT and DR should perform well in that setting because they rely on group differences and maximal variation. It is possible that the FTM method performs better in settings with more localized functional differences. The NFL dataset differed from the AD dataset in both imaging modality (SPECT vs. PET) and uptake conditions (continuous performance test vs. rest), which could contribute to the differences in classifier performance. We suspect modality is not a factor as the feature scores used in classification are modality neutral. Lower resolution imaging systems may contribute to lower true positive rates if the regions of interest are small in size, despite the model’s attempt to mitigate this effect using filter sizes of differing spatial scales. Regardless of how well the filtering method does, if the discriminating feature of a disease is too small to be accurately measured by the imaging device, performance of the classification system will undoubtedly suffer. The benefit of this method is that it uses information across spatial scales, orientations, and locations in the volumes to calculate the matching scores used for subsequent classification and should therefore be less reliant on any one discriminating feature. The uptake task will contribute to the functional signals and should be taken into account when selecting the regions of interest to calculate feature scores (section 2.3.1). Choosing regions that are absolutely not affected by the disease will decrease the discriminative power of the method. Alternatively, if the number of subjects in the dataset is high and there is no fear of classifier overfitting, choosing many regions, some known to be related to the disease and/or task and others whose relationship is unknown could provide interesting insight into whether the unknown regions are contributing to classification accuracy. Further, because the features of the dataset are computed separately from the classifier, one could choose to sample some features from all brain regions and either perform regularization in the classifier or choose a classification model that is less sensitive to overfitting (e.g. support vector machines). Each of these decisions should be made relative to the particular dataset and illness being studied.

CONCLUSIONS

In general our volumetric variant of the hierarchical feed-forward model originally proposed by Serre et al. for detecting objects in photographs performed quite well on the functional brain imaging data sets used in this study. In fact, it outperformed both the comparison

methods and the human counterpart at detecting AD in the FDG PET ADNI data set. The method is very general and does not rely on particular imaging modalities. It could be used on many spatial maps commonly computed in diagnostic and research imaging studies. Furthermore, there is evidence that it could be used to test hypotheses about regions implicated in disease. In conclusion, models designed in the computer vision community for object recognition and tracking in images of natural scenes may indeed have applications in detecting and tracking disease progression in human functional brain imaging with minimal modifications.

Acknowledgments

This research was supported in part by Grant Number 1 UL1 RR024150 from the National Center for Research Resources (NCRR), a component of the National Institutes of Health (NIH), and the NIH Roadmap for Medical Research. Its contents are solely the responsibility of the authors and do not necessarily represent the official view of NCRR or NIH. Information on NCRR is available at http://www.ncrr.nih.gov/. Information on Reengineering the Clinical Research Enterprise can be obtained from http://nihroadmap. The authors wish to thank the patients and healthy volunteers who participated in the studies supplying the data for this methods evaluation. The research was further supported by grants to the Functional Imaging Biomedical Informatics Research Network (FBIRN U24-RR021992, National Center for Research Resources), and the Biomedical Informatics Research Network (1 U24 RR025736-01). Data collection and sharing for the NFL dataset used in this study was made available by Daniel Amen, M.D. and the Amen Clinics Inc. Data collection and sharing for the AD dataset used in this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI; Principal Investigator: Michael Weiner; NIH grant and supplement). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering (NIBIB), and through generous contributions from the following: Pfizer Inc., Wyeth Research, Bristol-Myers Squibb, Eli Lilly and Company, GlaxoSmithKline, Merck & Co. Inc., AstraZeneca AB, Novartis Pharmaceuticals Corporation, Alzheimer's Association, Eisai Global Clinical Development, Elan Corporation plc, Forest Laboratories, and the Institute for the Study of Aging, with participation from the U.S. Food and Drug Administration. Industry partnerships are coordinated through the Foundation for the National Institutes of Health. The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Disease Cooperative Study at the University of California, San Diego. ADNI data are disseminated by the Laboratory of Neuro-Imaging at the University of California, Los Angeles. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript (R Grant Number UL1 RR025774).

References

1. Friston, KJ.; Ashburner, J.; Kiebel, SJ.; Nichols, TE.; Penny, WD., editors. Statistical Parametric Mapping: The Analysis of Functional Brain Images. Academic Press;
2. Amen D, Newberg A, Thatcher R, Jin Y, Wu J, Phillips B, Keator D, Willeumier K. Impact of Playing Professional American Football on Long-Term Brain Function. Journal of Neuropsychiatry and Clinical Neurosciences. 2011; 23(1)
3. Bau TC, Healey G. Rotation and scale invariant hyperspectral classification using 3D Gabor filters. 2009:73340B.
4. Bau TC, Sarkar S, Healey G. Using three-dimensional spectral/spatial Gabor filters for hyperspectral region classification. 2008:69660E.
5. Borroni B, Premi E, Di Luca M, Padovani A. Combined biomarkers for early Alzheimer disease diagnosis. Current medicinal chemistry. 2007; 14(11):1171–1178. [PubMed: 17504137]
6. Braak H, Braak E. Staging of Alzheimer-related cortical destruction. International Psychogeriatrics. 1997; 9(S1):257–261. [PubMed: 9447446]
7. Brunelli, R. Template matching techniques in computer vision: Theory and practice. John Wiley & Sons Inc; 2009.
8. De Santi S, de Leon MJ, Rusinek H, Convit A, Tarshish CY, Roche A, Tsui WH, Kandil E, Boppana M, Daisley K, et al. Hippocampal formation glucose metabolism and volume losses in MCI and AD. Neurobiology of aging. 2001; 22(4):529–539. [PubMed: 11445252]

9. Delacourte A, David JP, Sergeant N, Buee L, Wattez A, Vermersch P, Ghozali F, Fallet-Bianco C, Pasquier F, Lebert F, et al. The biochemical pathway of neurofibrillary degeneration in aging and Alzheimer's disease. Neurology. 1999; 52(6):1158. [PubMed: 10214737]
10. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics. 1988; 44(3):837–45. [PubMed: 3203132]
11. Drzezga A. Diagnosis of Alzheimer's disease with [18F] PET in mild and asymptomatic stages. Behavioural Neurology. 2009; 21(1):101–115. [PubMed: 19847049]
12. Espasy A, Jacobs D. Frontal Lobe Syndromes. eMedicine Specialties. Behavioral Neurology and Dementia ed. 2010
13. Forsyth, DA.; Ponce, J. Computer vision: a modern approach. Prentice Hall; Professional Technical Reference: 2002.
14. Grignon Y, Duyckaerts C, Bennecib M, Hauw JJ. Cytoarchitectonic alterations in the supramarginal gyrus of late onset Alzheimer's disease. Acta Neuropathol. 1998; 95(4):395–406. [PubMed: 9560018]
15. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. The WEKA data mining software: An update. ACM SIGKDD Explorations Newsletter. 2009; 11(1):10–18.
16. Hinrichs C, Singh V, Mukherjee L, Xu G, Chung MK, Johnson SC. Spatially augmented LPboosting for AD classification with evaluations on the ADNI dataset. Neuroimage. 2009; 48(1): 138–149. [PubMed: 19481161]
17. Hua X, Leow AD, Parikshak N, Lee S, Chiang MC, Toga AW, Jack CR Jr, Weiner MW, Thompson PM. Tensor-based morphometry as a neuroimaging biomarker for Alzheimer's disease: an MRI study of 676 AD, MCI, and normal subjects. Neuroimage. 2008; 43(3):458–469. [PubMed: 18691658]
18. Hubel DH, Wiesel TN. Receptive fields of single neurones in the cat's striate cortex. The Journal of Physiology. 1959; 148(3):574. [PubMed: 14403679]
19. Hubel DH, Wiesel TN. Integrative action in the cat's lateral geniculate body. J Physiol. 1961; 155:385–98. [PubMed: 13716436]
20. Hubel DH, Wiesel TN. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. J Physiol. 1962; 160:106–54. [PubMed: 14449617]
21. Hubel DH, Wiesel TN. Binocular interaction in striate cortex of kittens reared with artificial squint. J Neurophysiol. 1965a; 28(6):1041–59. [PubMed: 5883731]
22. Hubel DH, Wiesel TN. Receptive Fields and Functional Architecture in Two Nonstriate Visual Areas (18 and 19) of the Cat. J Neurophysiol. 1965b; 28:229–89. [PubMed: 14283058]
23. Hubel DH, Wiesel TN. Receptive fields and functional architecture of monkey striate cortex. The Journal of Physiology. 1968; 195(1):215. [PubMed: 4966457]
24. Kantarci K, Jack CR Jr. Neuroimaging in Alzheimer disease: an evidence-based review. Neuroimaging Clinics of North America. 2003; 13(2):197. [PubMed: 13677801]
25. Kawachi T, Ishii K, Sakamoto S, Sasaki M, Mori T, Yamashita F, Matsuda H, Mori E. Comparison of the diagnostic performance of FDG-PET and VBM-MRI in very mild Alzheimer?s disease. European Journal of Nuclear Medicine and Molecular Imaging. 2006; 33(7):801–809. [PubMed: 16550383]
26. Knopman DS, DeKosky ST, Cummings JL, Chui H, Corey-Bloom J, Relkin N, Small GW, Miller B, Stevens JC. Practice parameter: diagnosis of dementia (an evidence-based review): report of the Quality Standards Subcommittee of the American Academy of Neurology. Neurology. 2001; 56(9):1143. [PubMed: 11342678]
27. Langbaum J, Chen K, Lee W, Reschke C, Bandy D, Fleisher AS, Alexander GE, Foster NL, et al. Categorical and correlational analyses of baseline fluorodeoxyglucose positron emission tomography images from the Alzheimer's Disease Neuroimaging Initiative (ADNI). Neuroimage. 2009; 45(4):1107–1116. [PubMed: 19349228]
28. Lovestone S. Searching for biomarkers in neurodegeneration. Nature Medicine. 2010; 16(12): 1371–1372.
29. Lowe DG. Object recognition from local scale-invariant features. 1999

30. Martinez LM, Alonso JM. Complex receptive fields in primary visual cortex. Neuroscientist. 2003; 9(5):317–31. [PubMed: 14580117]
31. Mosconi L, Sorbi S, de Leon MJ, Li Y, Nacmias B, Myoung PS, Tsui W, Ginestroni A, Bessi V, Fayyazz M, et al. Hypometabolism exceeds atrophy in presymptomatic early-onset familial Alzheimer's disease. Journal of Nuclear Medicine. 2006; 47(11):1778. [PubMed: 17079810]
32. Mueller SG, Weiner MW, Thal LJ, Petersen RC, Jack C, Jagust W, Trojanowski JQ, Toga AW, Beckett L. Alzheimer?s Disease Neuroimaging Initiative. Advances in Alzheimer?s and Parkinson?s Disease. 2008:183–189.
33. Mutch J, Lowe DG. Multiclass object recognition with sparse, localized features. 2006
34. Nordberg A, Rinne JO, Kadir A, Laanom B. The use of PET in Alzheimer disease. Nature Reviews Neurology. 2010; 6(2):78–87. [PubMed: 20139997]
35. Rachakonda V, Tian Hong PAN, Le WD. Biomarkers of neurodegenerative disorders: How good are they? Cell Research. 2004; 14(5):349–358.
36. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez JC, Muller M. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics. 2011; 12:77. [PubMed: 21414208]
37. Roe CM, Fagan AM, Williams MM, Ghoshal N, Aeschleman M, Grant EA, Marcus DS, Mintun MA, Holtzman DM, Morris JC. Improving CSF biomarker accuracy in predicting prevalent and incident Alzheimer disease. Neurology. 2011
38. Schwartz O, Pillow JW, Rust NC, Simoncelli EP. Spike-triggered neural characterization. J Vis. 2006; 6(4):484–507. [PubMed: 16889482]
39. Serre T, Wolf L, Poggio T. Object recognition with features inspired by visual cortex. 2005:994–1000.
40. Tartaglia MC, Rosen HJ, Miller BL. Neuroimaging in Dementia. Neurotherapeutics. 2011:1–11. [PubMed: 21274679]
41. Thompson PM, Hayashi KM, De Zubicaray G, Janke AL, Rose SE, Semple J, Herman D, Hong MS, Dittmer SS, Doddrell DM, et al. Dynamics of gray matter loss in Alzheimer's disease. Journal of Neuroscience. 2003; 23(3):994. [PubMed: 12574429]
42. Tzourio-Mazoyer N, Landeau B, Papathanassiou D, Crivello F, Etard O, Delcroix N, Mazoyer B, Joliot M. Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. Neuroimage. 2002; 15(1):273–289. [PubMed: 11771995]
43. Van Herk M. A fast algorithm for local minimum and maximum filters on rectangular and octagonal kernels. Pattern Recognition Letters. 1992; 13(7):517–521.

Feed-forward hierarchical model of the ventral visual stream applied to functional brain image classification

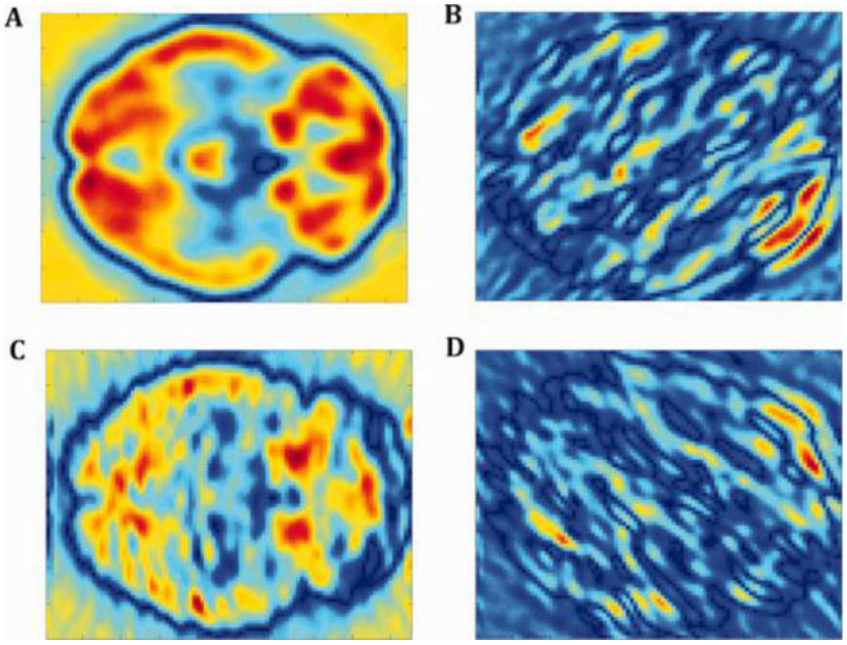

Figure I.

Examples of Gabor filtered slices. For each example, the filter size, σ, and λ remained constant at 5^3, 2.1, and 2.6 respectively while the orientation parameters θ and φ were varied. A) θ=0, φ=0; B) θ=π/4, φ=π/4; C) θ=π/2, φ=π/4; D) θ=3π/4, φ=π/4. The maximum filter responses are shown in red. As the orientation of the filters change (A–D), signals of similar orientations are selected by the filter.

Feed-forward hierarchical model of the ventral visual stream applied to functional brain image classification

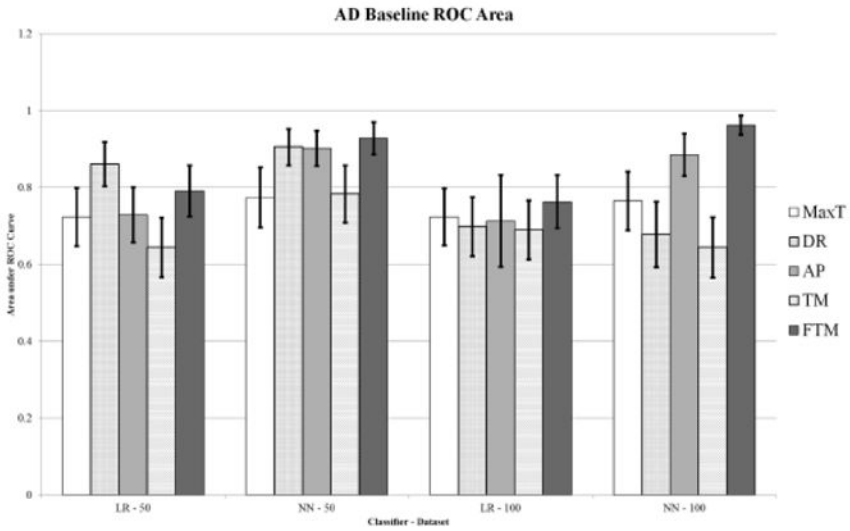

Figure IIa.

Area under the ROC curve for AD classification of the ADNI baseline data set for logistic regression (LR) and neural network (NN) classifiers for both 50 and 100 feature datasets (MaxT = maximum t-score, DR = PCA data reduction, AP = average patch, TM = template matching, FTM = Gabor filtering + template matching). The FTM method outperforms the others in 94% of the cases and is statistically better in 50% of the cases.

Feed-forward hierarchical model of the ventral visual stream applied to functional brain image classification

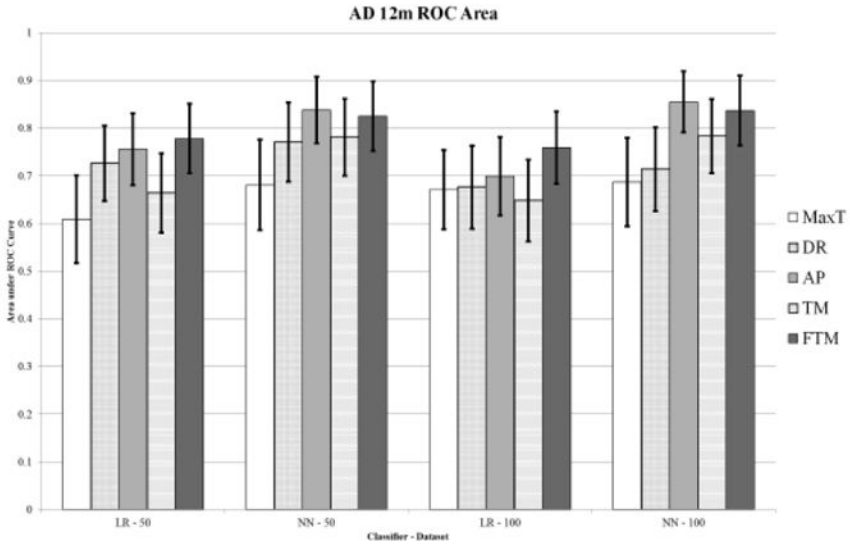

Figure IIb.

Area under the ROC curve for AD classification of the ADNI 12m data set for logistic regression (LR) and neural network (NN) classifiers for both 50 and 100 feature datasets (MaxT = maximum t-score, DR = PCA data reduction, AP = average patch, TM = template matching, FTM = Gabor filtering + template matching). The FTM method outperforms the others in 88% of the cases and is statistically better in 38% of the cases.

Feed-forward hierarchical model of the ventral visual stream applied to functional brain image classification

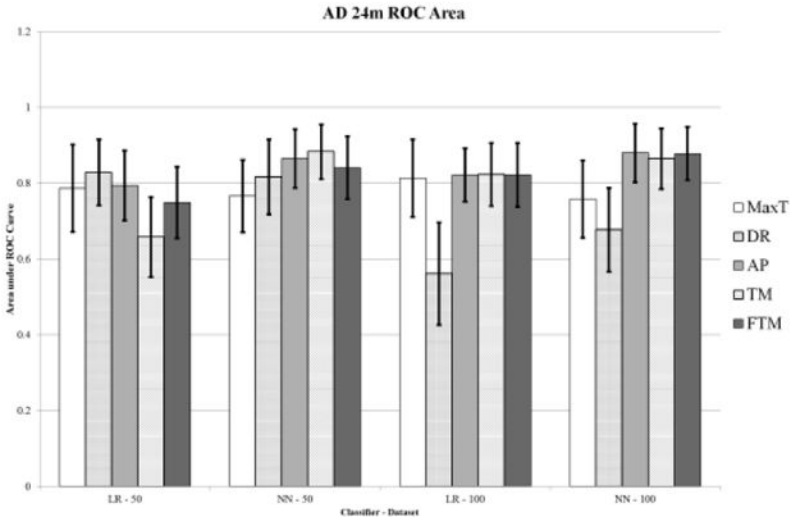

Figure IIc.

Area under the ROC curve for AD classification of the ADNI 24m data set for logistic regression (LR) and neural network (NN) classifiers for both 50 and 100 feature datasets (MaxT = maximum t-score, DR = PCA data reduction, AP = average patch, TM = template matching, FTM = Gabor filtering + template matching). The FTM method outperforms the others in 56% of the cases and is statistically better in 19% of the cases.

Feed-forward hierarchical model of the ventral visual stream applied to functional brain image classification

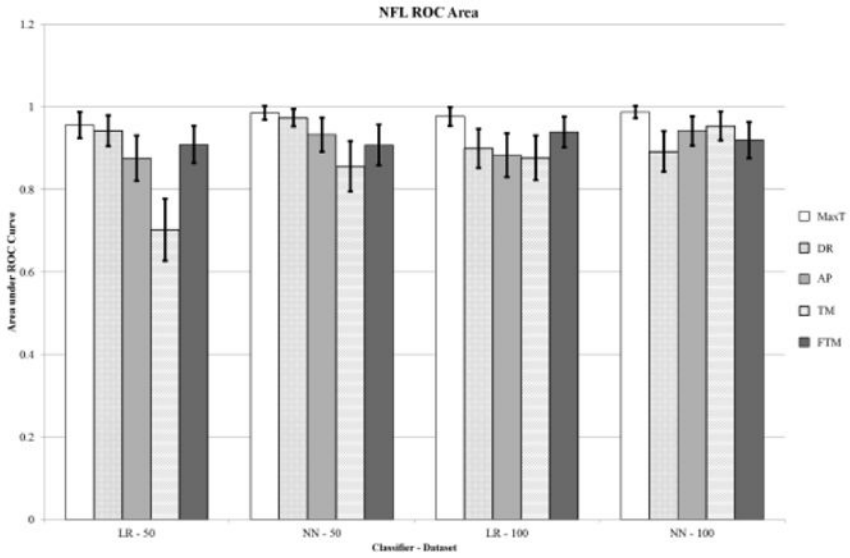

Figure III.

Area under the ROC curve for NFL classification for logistic regression (LR) and neural network (NN) classifiers for both 50 and 100 feature datasets (MaxT = maximum t-score, DR = PCA data reduction, AP = average patch, TM = template matching, FTM = Gabor filtering + template matching). The MaxT method outperformed the other methods, statistically better than the FTM method in all comparisons except in the LR-50 feature dataset. The FTM ROC-AUC was still very good, always greater than 0.900 and as high as 0.939 in the NN-100 feature dataset.

Table I

S1 Layer Gabor filter sizes and parameters by band (rows 1–3) where bands are used to group similar filter sizes.

Band	1	2	3	4	5	6	7	8
Filter	3, 5	7, 9	11, 13	15, 17	19, 21	23, 25	27, 29	31, 33
Sigma	1.4, 2.1	3.0, 3.9	4.6, 5.6	6.5, 7.5	8.5, 9.6	10.6, 11.7	12.9, 14.1	15.3, 16.5
Lambda	1.7, 2.6	3.6, 4.8	5.7, 6.8	8.0, 9.2	10.4, 11.8	13.1, 14.5	15.9, 17.4	18.9, 20.5
Max Gird	4^3	6^3	8^s	10^3	12^3	14^3	16^3	18^3
Patch	5^3, 9^3, 13^3, 17^3							

Row 4 shows the C1 layer grid size for maximums over Gabor filter scales. Row 5 shows the template patch sizes common to all bands.

Hum Brain Mapp. Author manuscript; available in PMC 2016 February 08.

Table IIa

Results from the AD ROC-AUC analysis of the ADNI baseline data.

Dataset (#feat)	Classifier	Method	ROC-AUC	95% Cont	z-Score (X_{AUC} − FTM_{AUC})	P (FTM_{AUC} − X_{AUC})
		FTM	0.791	0.857–0.725		
		TM	0.644	0.721–0.567	**−3259**	**0.001**
		AP	0.729	0.801–0.657	−1.556	0.120
		DR	0.861	0.919–0.803	1.854	0.064
		MaxT	0.692	0.768–0.616	**−2.187**	**0.029**
AD-Bas (50)	NN					
		FTM	0.928	0.970–0.886		
		TM	0.783	0.858–0.709	**−4321**	**1.55E−04**
		AP	0.902	0.951–0.854	−1.132	0.158
		DR	0.905	0.952–0.858	−0.833	0.405
		MaxT	0.777	0.855–0.698	**−3.661**	**2.51E−04**
AD-Bas (100)	LR					
		FTM	0.763	0.832–0.694		
		TM	0.689	0.766–0.612	−1.761	0.078
		AP	0.713	0.832–0.694	−1.320	0.187
		DR	0.698	0.775–0.620	−1.604	0.109
		MaxT	0.687	0.761–0.614	−1.574	0.115
AD-Bas (100)	NN					
		FTM	0.962	0.987–0.938		
		TM	0.644	0.722–0.567	**−8.623**	**2.20E−16**
		AP	0.885	0.940–0.831	**−3336**	**8.50E−04**
		DR	0.678	0.763–0.594	**−6.697**	**2.13E−11**
		MaxT	0.773	0.849–0.696	**−5.053**	**435E−07**

The table lists ROC-AUC measurements, 95% confidence intervals, Z-scores, and probabilities for comparisons of the FTM method with the other methods within each dataset and classifier combination. Negative z-scores indicate methods that are lower in ROC-AUC than the FTM method. Significant differences are highlighted in bold.

MaxT = maximum *t*-score, DR = PCA data reduction, AP = average patch, TM = template matching, FTM = Gabor filtering + template matching.

Table IIb

Results from the AD ROC-AUC analysis of the ADNI 12m data.

Dataset (#feat)	Classifier	Method	ROC-AUC	95% Conf	*z*-Score (X_{AUC} − FTM_{AUC})	*P* (FTM_{AUC} − X_{AUC})
AD-12m (50)	LR					
		FTM	0.778	0.851–0.705		
		TM	0.664	0.747–0.582	**−2.173**	**0.030**
		AP	0.756	0.831–0.682	−0.499	0.618
		DR	0.726	0.805–0.648	−1.060	0.289
		MaxT	0.609	0.701–0.518	**−2.830**	**0.005**
AD-12m (50)	NN					
		FTM	0.825	0.898–0.753		
		TM	0.781	0.862–0.701	−0.952	0.341
		AP	0.838	0.908–0.769	0.319	0.750
		DR	0.771	0.854–0.689	−1.292	0.196
		MaxT	0.681	0.776–0.585	**−2371**	**0.019**
AD-12m (100)	LR					
		FTM	0.759	0.835–0.683		
		TM	0.648	0.734–0.561	**−2.166**	**0.030**
		AP	0.699	0.781–0.618	−1.210	0.226
		DR	0.676	0.763–0.588	−1.546	0.122
		MaxT	0.671	0.754–0.588	−1.532	0.127
AD-12m (100)	NN	FTM	0.837	0.910–0.764		
		TM	0.783	0.861–0.706	−1.411	0.158
		AP	0.855	0.919–0.791	0.590	0.555
		DR	0.714	0.802–0.627	**−2234**	**0.022**
		MaxT	0.687	0.780–0.594	**−2.482**	**0.014**

The table lists ROC-AUC measurements, 95% confidence intervals, Z-scores, and probabilities for comparisons of the FTM method with the other methods within each dataset and classifier combination. Negative z-scores indicate methods that are lower in ROC-AUC than the FTM method. Significant differences are highlighted in bold.

MaxT = maximum t-score, DR = PCA data reduction, AP = average patch, TM = template matching, FTM = Gabor filtering + template matching).

Table IIc

Results from the AD ROC-AUC analysis of the ADNI 24m data.

Dataset (#feat)	Classifier	Method	ROC-AUC	95% Conf	*z*-Score ($X_{AUL.} - FTM_{AUC}$)	*P* ($FTM_{AUC} - X_{AUC}$)
AD-24m (50)	LR					
		FTM	0.749	0.843–0.655		
		TM	0.658	0.763–0.553	−1.437	0.151
		AP	0.794	0.886–0.702	0.794	0.427
		DR	0.828	0.915–0.740	1.371	0.171
		MaxT	0.787	0.902–0.673	0.502	0.616
AD-24m (50)	NN					
		FTM	0.841	0.924–0.758		
		TM	0.883	0.955–0.810	0.991	0.322
		AP	0.865	0.942–0.788	0.736	0.462
		DR	0.816	0.915–0.717	−0.459	0.646
		MaxT	0.766	0.861–0.670	−1.335	0.182
AD-24m (100)	LR					
		FTM	0.822	0.906–0.737		
		TM	0.823	0.906–0.740	0.026	0.979
		AP	0.822	0.892–0.716	−0.319	0.75
		DR	0.561	0.426–0.696	**−4.620**	**3.84E−06**
		MaxT	0.813	0.915–0.710	−0.143	0.886
AD-24m (100)	NN					
		FTM	0.878	0.948–0.806		
		TM	0.864	0.944–0.783	−0383	0.702
		AP	0.880	0.957–0.804	0.102	0.919
		DR	0.677	0.788–0.566	**−8214**	**2.20E−16**
		MaxT	0.758	0.860–0.656	**−2273**	**0.023**

The table lists ROC-AUC measurements, 95% confidence intervals, Z-scores, and probabilities for comparisons of the FTM method with the other methods within each dataset and classifier combination. Negative z-scores indicate methods that are lower in ROC-AUC than the FTM method. Significant differences are highlighted in bold.

MaxT = maximum t-score, DR = PCA data reduction, AP = average patch, TM = template matching, FTM = Gabor filtering + template matching.

Table III

Results from the NFL ROC-AUC analysis.

Dataset (#feat)	Classifier	Method	ROC-AUC	95% Conf	*z*-Score (X_{AUC} − FTM_{AUC})	*P* (FTM_{AUC} = X_{AUC})
NFL (50)	LR					
		FTM	0.909	0.954–0.864		
		TM	0.702	0.777–0.628	**−4.662**	**5.09E−06**
		AP	0.876	0.931–0.821	−0.907	0.365
		DR	0.942	0.979–0.906	1.137	0.255
		MaxT	0.956	0.988–0.924	**2.059**	**0.040**
NFL (50)	NN					
		FTM	0.908	0.957–0.860		
		TM	0.856	0.917–0.795	−1.306	0.193
		AP	0.933	0.974–0.893	0.776	0.438
		DR	0.974	0.995–0.954	**2.405**	**0.016**
		MaxT	0.986	1.000–0.969	**2.944**	**0.003**
NFL (100)	LR					
		FTM	0.939	0.976–0.902		
		TM	0.877	0.931–0.822	−1.856	0.065
		AP	0.883	0.936–0.830	−1.705	0.089
		DR	0.900	0.947–0.853	−1.315	0.188
		MaxT	0.977	1.000–0.954	1.753	0.080
NFL (100)	NN					
		FTM	0.920	0.964–0.876		
		TM	0.954	0.989–0.918	1.167	0.245
		AP	0.942	0.977–0.907	0.765	0.445
		DR	0.892	0.941–0.842	−0.866	0.386
		MaxT	0.988	1.000–0.973	**2.900**	**0.004**

The table lists ROC-AUC measurements, 95% confidence intervals, Z-scores, and probabilities for comparisons of the FTM method with the other methods within each dataset and classifier combination. Negative z-scores indicate methods that are lower in ROC-AUC than the FTM method. Significant differences are highlighted in bold.

MaxT = maximum t-score, DR = PCA data reduction, AP = average patch, TM = template matching, FTM = Gabor filtering + template matching.

Table IV

Results from the visual ratings of neuroanatomist J.H. Fallon from the ADNI baseline data.

Method	AD-TP	AD-FP	HC-TP	HC-FP
J.H.F	0.718	0.380	0.671	0.244
FTM	0.875	0.122	0.878	0.115
DR	0.622	0.389	0.611	0.378
MaxT	0.829	0.375	0.625	0.171

The table lists true positive (TP) and false positive (FP) values for the Alzheimer's disease (AD) and healthy control (HC) classes compared to the FTM, DR, and MaxT methods. The FTM method outperforms both the human rater and the other methods.

MaxT = maximum t-score, DR = PCA data reduction, FTM = Gabor filtering + template matching.