DISEASE CLASSIFICATION: A PROBABILISTIC APPROACH

Yogesh Rathi *, J. Malcolm, S. Bouix, R. McCarley, L. Seidman, J. Goldstein, C-F Westin, M. E. Shenton

Harvard Medical School, Boston

ABSTRACT

We describe a probabilistic technique for separating two populations whereby analysis is performed on affine-invariant representations of each patient. The method begins by converting each voxel from a high-dimensional diffusion weighted signal to a low-dimensional diffusion tensor representation. Three orthogonal measures that capture different aspects of the local tissue are derived from the tensor representation to form a feature vector. From these feature vectors, we form a probabilistic representation of each patient. This representation is affine invariant, thus obviating the need for registration of the images. We then use a Parzen window classifier to estimate the likelihood of a new patient belonging to either population. To demonstrate the technique, we apply it to the analysis of 22 first-episode schizophrenic patients and 20 normal control subjects. With leave-many-out cross validation, we find a detection rate of 90.91% (10% false positives).

Index Terms— Diffusion Tensor Imaging (DTI), Classification, schizophrenia, affine-invariant

1. INTRODUCTION

Nearly 1% of the population in the US is affected by schizophrenia. A growing body of evidence suggests that the early stages of schizophrenia (and many other brain disorders) are critical in forming and predicting the course and outcome of the disorder. The classification tools proposed in this work can serve as a first step towards early detection of schizophrenia, which may result in a better prognosis and functional outcome.

Both, post-mortem and neuroimaging studies have contributed significantly to what we know about the brain. Moreover, MRI studies of volumetric reduction in several brain regions in schizophrenia have been particularly informative with respect to confirming early speculations that the brain is disordered in schizophrenia. This work is largely the result of advances in neuroimaging that allowed for more careful measurement of regions of interest within the brain (see review in [1,2]). In particular, a shape based framework was recently developed by the authors in [3], which utilized volumetric differences as discriminatory features for distinguishing a population of schizophrenia from normal controls (NC). Another

recent work in this direction was proposed in [4], where the authors used the affine parameters obtained during registration of each subject to a given atlas as a discriminant feature. However, both works, used structural MRI data, as opposed to diffusion MRI being used in this study.

The advent of diffusion magnetic resonance imaging (dMRI) has provided the opportunity for non-invasive investigation of neural architecture of the brain. One of the simplest and widely used model that is used to analyze dMRI images is the diffusion tensor model. Diffusion tensors represent the diffusion of water molecules in three orthogonal directions with the principal direction aligned with the fiber orientation. Several scalar measures derived from this model are used to assess the strength of connectivity in neural fiber bundles.

While there has been an explosion in the number of studies reporting statistical differences in various regions of the brain (see [2] and the references therein), very few works have addressed the problem of classifying schizophrenic patients from healthy subjects. Apart from structural MRI data being used by [3, 4], there has been work done by [5, 6], where the authors use dimensionality reduction followed by linear discriminant analysis for classification of patients with schizophrenia (chronic). They use the fractional anisotropy and/or linear anisotropy images derived from single tensor estimation as discriminant features.

In all of the approaches listed above [3, 5], the authors first perform a registration of all the subjects to a common atlas space and subsequent analysis is done on this dataset. However, as has been shown by the authors in [7], transforming the images from subject space to atlas space can significantly change certain quantities under study (for example, volume). Thus, errors accrued during registration are carried forward in subsequent analysis, in particular in classification tasks. To avoid such errors and to reduce the computational complexity, we propose an affine-invariant classification scheme, which removes the need for registration of the images. We employ several anisotropy features of the white matter of the brain to discriminate first-episode (FE) schizophrenia patients from normal controls (NC) in a probabilistic fashion.

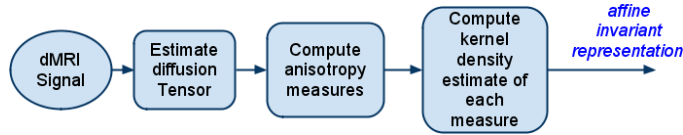*yogesh@bwh.harvard.edu

Fig. 1. Overall outline for computing an affine invariant representation of each subject

2. METHODS

The algorithm proceeds by first computing an affine-invariant probabilistic representation of each subject (see Figure 1). From the dMRI scan of a subject, diffusion tensors are first estimated. Three orthogonal anisotropy measures [8] that form the discriminatory features of our classifier are then computed at each voxel in the white matter region. A nonparametric density estimator is then used to convert the discriminatory features of each subject into an affine-invariant proababilistic representation. This representation is subsequently used by a Parzen window classifier to compute the probability of a previously unseen subject being FE or NC in a cross-validation scheme. Details on each of these steps is given in the next section.

2.1. Preliminaries

In diffusion weighted imaging, image contrast is related to the strength of water diffusion. At each image voxel, diffusion is measured along a set of distinct gradients, $\mathbf{u}_1, ..., \mathbf{u}_n \in \mathbb{S}^2$ (on the unit sphere), producing the corresponding signal, $\mathbf{s} = [s_1, ..., s_n]^T \in \mathbb{R}^n$. The diffusion tensor is related to the signal using the following relation [9, 10]:

$$s_i = s_0 \exp\left(-b\mathbf{u}_i^T D \mathbf{u}_i\right),$$

where s_0 is a baseline signal intensity, b is an acquisition-specific constant, and D is a tensor describing the diffusion pattern. D can be estimated using a weighted least-squares approach [11].

Several scalar measures derived from the single tensor model have been proposed in the literature [8, 12, 13]. In particular, we use a set of three orthogonal invariants studied in [8], namely the norm N, fractional anisotropy FA and mode M_d. These measures capture different (orthogonal) aspects of the shape of the tensor. Given, a diffusion tensor D, these measures can be computed as follows:

$$N = \| D \|, \quad FA = \frac{\sqrt{3} \| D - \frac{1}{3}tr(D)I \|}{\sqrt{2} \| D \|},$$

$$M_d = 3\sqrt{6} \left| \left(\frac{\tilde{D}}{\| \tilde{D} \|} \right) \right|, \quad \tilde{D} = D - \frac{1}{3}tr(D)I \tag{1}$$

where, $|.|$ denotes the determinant, $tr(.)$ is the trace and $\| . \|$ denotes the frobenius norm of a matrix. Thus, FA measures how the shape of the tensor deviates from that of a sphere. M_d indicates the mode of the tensor, *i.e.* $M_d = -1$ indicates planar anisotropy, $M_d = 0$ indicates an orthotropic tensor

and $M_d = 1$ indicates linear anisotropic tensor. Norm N measures the "size" of the diffusion tensor. From the above discussion, our feature vector is given by the 3-dimensional vector

$$f = [\, N, FA, M_d \,]^T. \tag{2}$$

2.2. Affine Invariant Representations

All previous methods [3, 5] that performed classification of schizophrenic patients, performed spatial normalization (registration) of the structural or FA images to a common atlas space. A recent study by [7] has shown that registration can introduce unwanted effects in the data set. In this work, we avoid the errors accrued during registration by using a probability density based classifier.

Probability density functions (pdf) are invariant to translation, rotation, scale and shear of an image, *i.e.* pdf's are invariant under linear transformation of the coordinates of an image. A nonparametric estimate of the pdf can be computed using the following expression [14]:

$$p(z) = \frac{1}{Mh} \sum_{x=1}^{M} G\left(\frac{z - I(x)}{h} \right), \quad z \in \{\text{Range of } I\} \tag{3}$$

where $I(x)$ is a scalar image at spatial location x, M is the number of data points, G is a Gaussian kernel and h denotes the bandwidth of the kernel.

The proposed feature set f lives in a 3-dimensional space. Computing the joint pdf in such a high-dimensional space is computationally intensive. Further, the features N, FA, M_d are mutually orthogonal and hence independent. As such, we compute a 1D pdf for each feature separately using (3). Note that, each of these features capture different aspects of the variation in "shape" of the diffusion tensor and hence are independent of the orientation. Thus, these measures themselves are invariant under affine transformations.

Several schizophrenia studies [2] have shown abnormalities in the white matter region of the brain. We thus choose this entire region (white matter) to compute the pdf. Specifically, a diffusion tensor is estimated at each voxel and FA is computed in the entire image volume. Regions of the brain that have $FA \geq 0.4$ are selected for further analysis. This roughly corresponds to the white matter region in the brain. All the other features (such as, M_d, N) are computed in this region.

Using (3), we compute the pdf for each of the three discriminatory measures and combine them into a matrix representation denoted by $\mathbf{p} = [p_n \; p_{fa} \; p_{md}]$. Thus, each patient

scan i can now been transformed into a probabilistic representation (matrix) $\mathbf{p}^i$ of dimension $n_b \times 3$, where n_b is the number of bins used in the pdf computation. In our subsequent discussions, we will use this representation in our classifier.

Figures 2a,2b,2c, show the pdf's for 22 first-episode (FE) schizophrenic patients (red) along with 20 age-matched normal controls (NC) (blue). A visual inspection shows differences between the two groups (blue and red) for each of these measures.

2.3. Parzen Window Classifier

The Parzen window classifier was first introduced by [15]. In this method, a Parzen window based density estimate is used to compute the probability that a new data point belongs to one of the groups in the training data set.

Let $\{\mathbf{p}_{fe}^i\}_{i=1}^{N_{fe}}$ and $\{\mathbf{p}_{nc}^i\}_{i=1}^{N_{nc}}$ be the set of N_{fe} FE and N_{nc} NC subjects in the training data set. Given a test data point $\hat{\mathbf{p}}$, the probability that it belongs to either group can be computed using the Parzen window density estimator as follows:

$$\mathcal{P}^{fe}(\hat{\mathbf{p}}) = \frac{1}{N_{fe}} \sum_{i=1}^{N_{fe}} K(\hat{\mathbf{p}}, \mathbf{p}_{fe}^i),$$

$$\mathcal{P}^{nc}(\hat{\mathbf{p}}) = \frac{1}{N_{nc}} \sum_{i=1}^{N_{nc}} K(\hat{\mathbf{p}}, \mathbf{p}_{nc}^i), \qquad (4)$$

where $K(.,.)$ is a Gaussian kernel given by

$$K(\mathbf{p}^i, \mathbf{p}^j) = \exp\left(\sum_{f=1}^{3} \frac{-\| p_f^i - p_f^j \|^2}{\sigma_f^2} \right), \qquad (5)$$

with $f = [N, FA, M_d]^T$ as described earlier, and i, j represent the indices for i^{th} and j^{th} subject.

For each of the two groups, we choose σ_f using the following relation:

$$\sigma_f = \frac{c_f}{N} \sum_{i=1}^{N} \min_{i \neq j} \| p_f^i - p_f^j \|, \quad j = 1, 2, ..N, \quad f = 1 \text{ to } 3,$$

where N is N_{fe} for the group of FE patients and $N = N_{nc}$ for NC subjects. The constant c_f is a scalar that is computed so that the training error is minimized. Typical values for c_f lie in the range $c_f \in [1.5, \ 2]$. This data driven approach of choosing σ_f is quite common in the literature and has been used in other works as well [16]. This choice of σ_f is guided by the following considerations: 1) σ_f varies appropriately with the scaling of each of the components of f, 2) It minimizes the training error of the classifier, 3) it respects the distribution of points within the clusters (whether the points are spread out or densely packed).

Thus, from the probabilities obtained in (4), we obtain the following simple classification rule:

$$\text{Classification of } \hat{\mathbf{p}} = \begin{cases} \mathcal{P}^{fe}(\hat{\mathbf{p}}) > \mathcal{P}^{nc}(\hat{\mathbf{p}}), & \hat{\mathbf{p}} \text{ is FE patient} \\ \mathcal{P}^{fe}(\hat{\mathbf{p}}) \leq \mathcal{P}^{nc}(\hat{\mathbf{p}}), & \hat{\mathbf{p}} \text{ is NC}. \end{cases}$$

3. RESULTS

3.1. Data acquisition protocol

Our dataset consisted of 22 FE patients (all males) with average age 20.89 ± 4.8 years and 20 NC (all males) with average age 22.3 ± 4.2 years. All the subjects were scanned as part of Dr. Martha Shenton's NIH grant (P50 MH 080272) on a 3-Tesla GE system using an echo planar imaging (EPI) diffusion weighted image sequence. A double echo option was used to reduce eddy-current related distortions. To reduce impact of EPI spatial distortion, an eight channel coil was used to perform parallel imaging using Array Spatial Sensitivity Encoding Techniques (GE) with a SENSE-factor (speed-up) of 2. Acquisitions have 51 gradient directions with b-value = 900 and eight baseline scans with b=0. The original GE sequence was modified to increase spatial resolution, and to further minimize image artifacts. The following scan parameters were used: TR 17000 ms, TE 78 ms, FOV 24 cm, 144x144 encoding steps, 1.7 mm slice thickness. All scans had 85 axial slices parallel to the AC-PC line covering the whole brain.

The raw diffusion weighted images were preprocessed using the Rician noise removal algorithm of [17] followed by eddy current and head motion correction algorithm [18] (part of the FSL package - `http://www.fmrib.ox.ac.uk/fsl/flirt/`).

3.2. Classification Results

Leave-many-out (LMO) is an unbiased technique for cross-validation of classification results [19, 20]. We used this technique to test our classifier. In this method, M subjects are removed at random from the dataset and the classifier is trained on the remaining samples. Testing is then done on these M subjects and classification scores are computed from several such runs.

In this work, we performed several experiments by training the classifier on different number of samples. Training was done separately by randomly selecting $Tp = \{$ 60%, 70%, 80%, 90%, 98% $\}$ of the total (here 42) samples respectively. For each Tp (say 60%), 1000 different training datasets were randomly generated and testing was performed on the remaining (say 40%) data. A particular test subject was classified as FE or NC based on a majority voting rule. In all of the scenarios, where the training set consisted of 60% to 98% of the total samples, **the classifier consistently gave a detection rate (true positives) of** 90.91% **with** 10% **false positives. The overall classification error was 9.52%**. Alternatively, the sensitivity was 90.91% and specificity was 90%. Figure 3 shows the frequency plot (the number of times a subject was classified as FE or NC on 1000 different training samples) for the case when 30% of the samples were used in testing. During this run, on an average, each subject was tested approximately 310 times on different training data sets. Similar results were obtained with other training data sets as well.
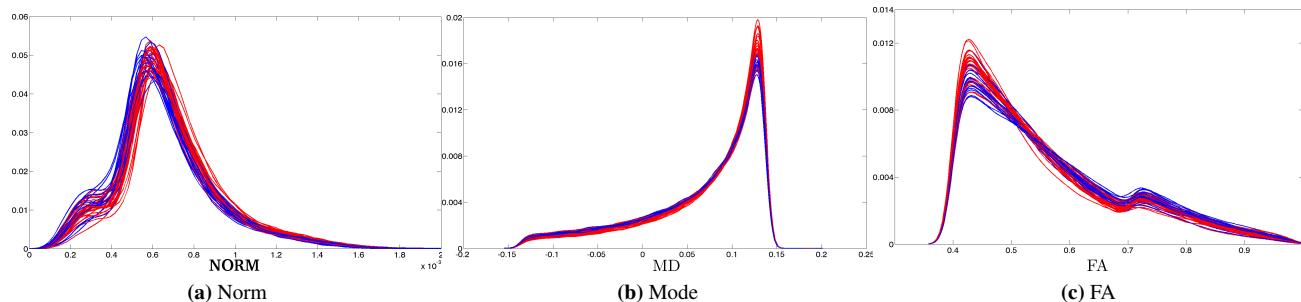
(a) Norm   **(b)** Mode   **(c)** FA

Fig. 2. Probability density functions of the three anisotropy measures for 22 FE patients (red) and 20 NC (blue).
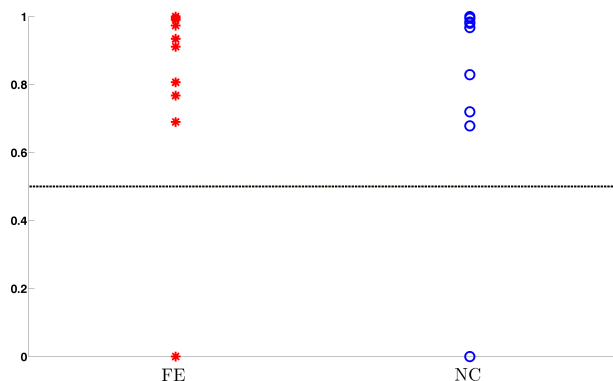

Fig. 3. Frequency plot showing the number of times each subject was classified as FE or NC with 70% of the data used in training. Frequency below 0.5 indicates misclassification by the classifier.

4. CONCLUSION

In this work, we proposed a probabilistic and affine-invariant representation of each subject obtained by extracting anisotropy features from diffusion tensor estimated at each voxel. This representation is used within a Parzen window classifier framework to classify each subject as being FE or NC. A leave-many-out cross-validation scheme was used to test the classifier on previously unseen data. On several different experimental settings, the classifier consistently gave a detection rate of 90.91% with 10% false positives.

5. REFERENCES

[1] M.E. Shenton, C.C. Dickey, M. Frumin, and R.W. McCarley, "A review of MRI findings in schizophrenia," *Schizophrenia Research*, vol. 49, no. 1-2, pp. 1–52, 2001.

[2] M. Kubicki, R. McCarley, C.-F. Westin, H.-J. Park, S. Maier, R. Kikinis, F.A. Jolesz, and M.E. Shenton, "A review of diffusion tensor imaging studies in schizophrenia," *J. of Psychiatric Research*, vol. 41, pp. 15–30, 2007.

[3] C. Davatzikos, D. Shen, R.C. Gur, X. Wu, D. Liu, Y. Fan, P. Hughett, B.I. Turetsky, and R.E. Gur, "Whole-brain morphometric study of schizophrenia revealing a spatially complex set of focal abnormalities," *Archives of general psychiatry*, vol. 62, no. 11, pp. 1218–1227, 2005.

[4] Kilian M. Pohl and Mert R. Sabuncu, "A unified framework for mr based disease classification," in *Information Processing in Medical Imaging*, Jerry L. Prince, Dzung L. Pham, and Kyle J. Myers, Eds. 2009, vol. 5636 of *Lecture Notes in Computer Science*, pp. 300–313, Springer.

[5] MWA Caan, KA Vermeer, LJ van Vliet, C. Majoie, BD Peters, GJ den Heeten, and FM Vos, "Shaving diffusion tensor images in discriminant analysis: A study into schizophrenia," *Medical Image Analysis*, vol. 10, no. 6, pp. 841–849, 2006.

[6] A. Caprihan, GD Pearlson, and VD Calhoun, "Application of principal component analysis to distinguish patients with schizophrenia from healthy controls based on fractional anisotropy measurements," *Neuroimage*, vol. 42, no. 2, pp. 675–682, 2008.

[7] J.S. Allen, J. Bruss, S. Mehta, T. Grabowski, C.K. Brown, and H. Damasio, "Effects of spatial transformation on regional brain volume estimates," *NeuroImage*, vol. 42, no. 2, pp. 535–547, 2008.

[8] G. Kindlmann, DB Ennis, RT Whitaker, and C.F. Westin, "Diffusion tensor analysis with invariant gradients and rotation tangents," *IEEE Transactions on Medical Imaging*, vol. 26, no. 11, pp. 1483–1499, 2007.

[9] PJ Basser, J. Mattiello, and D. LeBihan, "MR diffusion tensor spectroscopy and imaging," *Biophysical Journal*, vol. 66, no. 1, pp. 259–267, 1994.

[10] D. LeBihan, J.F. Mangin, C. Poupon, C.A. Clark, S. Pappata, N. Molko, and H. Chabriat, "Diffusion tensor imaging: concepts and applications," *Journal of Magnetic Resonance Imaging*, vol. 13, pp. 534–546, 2001.

[11] AC Aitken, "On least squares and linear combination of observations," in *Proc. R. Soc. Edinb*, 1934, vol. 55, pp. 42–48.

[12] E. Ozarslan, B.C. Vemuri, and T.H. Mareci, "Generalized scalar measures for diffusion MRI using trace, variance, and entropy," *Magnetic Resonance in Medicine*, vol. 53, no. 4, pp. 866–876, 2005.

[13] C.-F. Westin, S. E. Maier, H. Mamata, A. Nabavi, F. A. Jolesz, and R. Kikinis, "Processing and visualization of diffusion tensor MRI," *Medical Image Analysis*, vol. 6, no. 2, pp. 93–108, 2002.

[14] E. Parzen, "On estimation of a probability density function and mode," *The annals of mathematical statistics*, pp. 1065–1076, 1962.

[15] AK Jain and MD Ramaswami, "Classifier design with Parzen windows," *Pattern Recognition and artificial intelligence*, pp. 211–228, 1988.

[16] D. Cremers, T. Kohlberger, and C. Schnrr, "Nonlinear shape statistics in mumford-shah based segmentation," in *7th ECCV '02*, 2002, vol. 2351, pp. 93–108.

[17] S. Aja-Fernandez, M. Niethammer, M. Kubicki, M. E. Shenton, and C. F. Westin, "Restoration of DWI data using a rician LMMSE estimator," *IEEE Trans Med Imaging*, vol. 27, pp. 1389–1403, 2008.

[18] 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.

[19] V.N. Vapnik, *The nature of statistical learning theory*, springer, 2000.

[20] G.C. Cawley and N.L.C. Talbot, "Efficient leave-one-out cross-validation of kernel Fisher discriminant classifiers," *Pattern Recognition*, vol. 36, no. 11, pp. 2585–2592, 2003.