ANALYSIS OF PHOW REPRESENTATIONS FOR ALZHEIMER DISEASE CLASSIFICATION ON BRAIN STRUCTURAL MRI

Ricardo Mendoza-León [1,2], Fabio A. González [3], Pablo Arbeláez [4], John Puentes [2], and Marcela Hernández Hoyos [1]

[1]Systems and Computing Engineering Department, School of Engineering, Universidad de los Andes, Bogotá, Colombia.

[2]Lab-STICC UMR CNRS 6285 Équipe DECIDE, Département Image et Traitement de l'Information, Institut Mines-Télécom, Télécom Bretagne, Brest, France.

[3]MindLab Research Group, Universidad Nacional de Colombia, Bogotá, Colombia.

[4] Biomedical Engineering Department, School of Engineering, Universidad de los Andes, Bogotá, Colombia.

ABSTRACT

Alzheimer's Disease (AD) is a neurodegenerative pathology characterized by progressive atrophy of brain and impairment of memory and cognitive functions. Physicians frequently use structural brain imaging to identify abnormal patterns in brain structure that may indicate probable AD. Thus, shape information is central for brain imaging analysis and AD diagnosis. This paper examines how three variants of Pyramid Histograms Of visual Words (PHOW) descriptions, a data-driven approach, handle the complex task of AD classification. 87 pathological cases and 87 controls from OASIS dataset were used to study the impact of shape and surface information. Best performance was 89.3%, a current mark for AD classification, and an increase (27.1%) in contrast to a naive approach. Additionally, controls were better classified than demented subjects (94.5% and 84.0%, respectively), while young, early-onset AD subjects, and elderly controls were the most difficult. Finally, dictionary word analysis revealed discriminative surface features. Also, local patterns induced by global word distribution appear to be more significant for classification than word location.

Index Terms— Alzheimer's classification, PHOW, support vector machine, bag of words, structural MRI.

1. INTRODUCTION

Today, 46.8 million people worldwide live with dementia [1]. The most common type of dementia is Alzheimer's Disease (AD), a neurodegenerative pathology, characterized by progressive atrophy of brain structure and impairment of cognitive functions, especially memory. Furthermore, the incidence of dementia is expected to increase to 131 million by 2050, making AD one of the most critical worldwide health challenges. Additionally, given the dramatic cognitive impairment in advanced AD subjects, early identification of AD along with appropriate treatment is very important, since it may slow down disease progression, reduce the economic impact of care, and improve the quality of life of patients and families.

Several brain imaging techniques are currently being used to support AD identification. Particularly, structural magnetic resonance imaging (MRI) has become an increasingly important diagnostic tool, because AD induces structural changes in brain matter (atrophy) visible in these images.

In this paper, we study the importance of shape information, which is coded in many complex surface patterns over the brain matter, for AD diagnosis.

Diverse proposals have been made to classify AD disease based on MRI brain images with promising results [2], relying on different data-driven modeling strategies such as: voxel-based morphometry [3], intensity-based Bag-of-Words (BoW) models [4], local binary patterns [5], support vector machine (SVM) saliency maps [6], and SIFT-based descriptors [7]. Also, strategies that extensively rely on domain knowledge biomarkers, including: atlas-based parcelations [8], white matter, gray matter segmentations [2, 8], and descriptions of localized structures (e.g. frontoparietal lobular cortical thickness, hippocampal shape) [2, 3, 8]. However, it is difficult to estimate the importance of global shape and surface information in those results, since different features are commonly fused.

In this work three alternatives to classify AD are evaluated, based on PHOW, which is an extended BoW model that uses Pyramid Histogram of Oriented Gradients [9] to describe images. The first one, PHOW-2D, considers 2D patches extracted from individual orientation-quantized image slices. Similarly, PHOW-2.5D, uses 2D patches, but extracted over a single orientation-quantized mosaic of ordered slices, implicitly introducing slice order information, hence the 2.5D name. The last one, is a full 3D orientation quantization approach, called PHOW-3D. This adaptable,

978-1-4799-2349-6/16/$31.00 ©2016 IEEE

data-driven, global modeling approach has the advantage of capturing shape, by quantization of directional edge responses. Moreover, in contrast to domain knowledge driven approaches, PHOW, being general, can be seamlessly applied in other classification tasks, without the need of extensive biomarker evaluation and selection. An example is the hippocampal structure, profusely studied in AD vs. control classification, but suboptimal when discriminating AD vs. control vs. frontotemporal dementia instances [10]. This distinctive feature has high value for differential diagnosis.

For all cases, the same classification approach with linear SVMs was applied. An Intersection Kernel chi^2 mapping procedure called IK-SMV [11], was used to map histogram features to a suitable feature space. This SVM approach is both simple and notoriously less computationally expensive than a nonlinear classifier. We also included as a baseline a naive 256-bin full-volume intensity histogram representation, since it neglects spatial data completely, yet can capture intensity and volumetric information, and has been previously used in MRI-based AD classification tasks [12].

Our dataset comprised 174 volumes, selected from OASIS dataset [13]. These volumes were averaged-motion and gain-field corrected, atlas-registered, and skull-stripped beforehand by the OASIS project standardized protocol.

2. METHODS

The PHOW method comprises four main steps: a) build a dictionary, b) extract image features, c) train an SVM classifier, and, finally, d) perform predictions.

To build a patch-based dictionary, volume information is first prepared: individual slices in PHOW-2D; an ordered mosaic of slices in PHOW-2.5D case; and anatomical ordered slices in a tri-dimensional intensity matrix in PHOW-3D. Bi-dimensional (i.e. 2D and 2.5D cases) or tri-dimensional (i.e. 3D case) convolutional Sobel filters are evaluated on every image setup obtaining a gradient response, which is more intense on edge (surface) locations. Afterwards, the gradient's dot-product response to directional unit-length quantization vectors is evaluated: 8 vectors for PHOW-2D and PHOW-2.5D, and 26 for PHOW-3D, spanning 45° angle steps along the θ polar axis and θ, ϕ spherical axes, respectively. The gradient response quantization value corresponds to the vector index giving the larger absolute response. However, an additional index (0) is given if the highest response is below an empirical threshold ($\gamma = 0.01$). Fig. 1, illustrates these steps for PHOW-3D.

Afterwards, a collection of regular sized bi-dimensional or tri-dimensional patches are randomly sampled on every train volume. Then, using a visual pyramid, a collection of increasingly finer spatial grids or "levels", collection's patches are further divided in 2D or 3D cells. The final patch description vector is obtained by concatenating histograms [14] of 9-bins for PHOW-2D and PHOW-2.5D, and 27-bins for PHOW-3D, computed separately at each cell, using the voxel's quantized directionality. Once all patches from a collection of volumes (a train set) have been sampled and

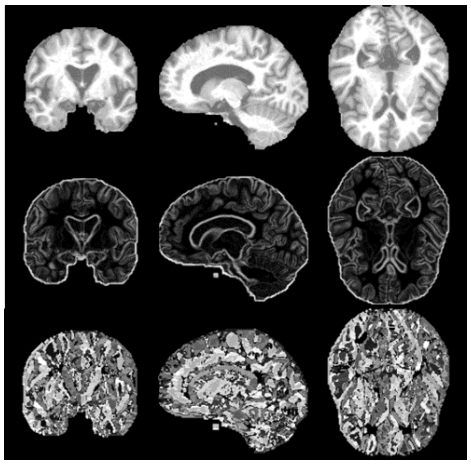

Fig. 1. Results of a directional quantization process in PHOW-3D. From left to right, the three columns represent coronal, sagittal, and axial planes views. From top to bottom, the three rows show, intensity images, absolute response of the Sobel kernels, and the gray-scale view of the quantized directional indexes, respectively.

described, a dictionary model is created, using a Euclidean K-Means clustering algorithm. This dictionary comprises the resulting centroids or "words" after clustering. In the feature extraction step, given a dictionary, an input image is represented by, sampling patches uniformly from the quantized image. Then, for each extracted patch, the closest word is found (a word hit) according to the Euclidean metric. Aggregated word hits are used to compute word frequency histograms for each image, and used to train a classifier.

The OASIS dataset sample comprised 87 control and 87 AD subjects, with varying ages and disease progression according to the corresponding Clinical Dementia Rating (CDR): CDR = 0 (87 healthy subjects), CDR = 0.5 (60 early-onset AD subjects), and CDR ≥ 1 (27 advanced AD subjects). In order to define equally distributed train and test sets, we divided the sample cases according the CDR and further individually subdivided into six age ranges, with similar number of subjects. Then, around 23% of the subjects were randomly selected on every CDR-range group (20 AD and 20 controls) for the test set, leaving the remaining subjects for the train set (67 AD and 67 controls).

For evaluation purposes, given a configuration of pyramid levels, patch size and number of words, five different dictionary representations (dissimilar clustering arrange-ments) of train and test sets were used to estimate linear SVM classifiers, and the resulting accuracies were averaged. In the case of PHOW-2D and PHOW-3D we explored one and two level pyramids: a 1x1 grid for the first level and 2x2 grid for the second level in PHOW-2D, and a 1x1x1 grid for the first level and 2x2x2 grid for the second level in PHOW-3D; 100, 200, 400, 800, 1600, and 3200 words dictionaries; and patch sizes of 8 and 16 voxels spanning 2 dimensions for PHOW-2D and three dimensions for PHOW-3D; totaling 24 parameter configurations and 120 models. Besides, in the case of PHOW-2.5D, the explored parameters were: input image mosaics of 4 (2x2), 16 (4x4), 64 (8x8) and 169 (13x13)

coronal slices taken from the volume center towards the anterior and posterior sides; 100, 200, 400, 800 and 3200 words dictionaries; two and three-level pyramids (2x2 grid for the first level, 4x4 grid for the second level, and a 8x8 grid for the third level); totaling 48 parameter configurations and 240 models, all with patch size 64x64.

3. RESULTS

The best averaged result obtained (Table 1) was 89.2% in a PHOW-3D model with one level pyramid (1x1x1 grid), patch size 8x8x8 and a dictionary of 800 words, a 27.1% difference with respect to the histogram representation (67.5% accuracy). On the other hand, the best PHOW-2D average accuracy was 87.1% using a one-level pyramid (1x1 grid), patch size 16x16 with a 200-word dictionary, and the best PHOW-2.5D result was 86.3% in two different configurations using 64-slice mosaics as inputs, a two-level (2x2, 4x4) pyramid, and dictionaries with 800 and 400 words. Furthermore, the control's classification rate was superior (averaged true negatives) than the AD classification rate.

The average accuracy showed varied behavior with respect to patch representation dimensionality (the number of patch cells times the number of quantization directions) and dictionary size parameters (Fig. 2 and Fig. 3, respectively), indicating a larger sensibility for PHOW-3D, followed by PHOW-2D and PHOW-2.5D.

Finally, for the best model, the average of all the absolute-gradient-response patch hits for every word in the dictionary, was evaluated in order to analyze how words captured shape features (Fig. 4). This analysis revealed that shape patterns were included in the dictionary, and also showed visual differences between computed averages on AD and control volumes in the test group.

4. DISCUSSION

PHOW models obtained positive average performance, closer to 90% for PHOW-3D, and above 80% for the others, being reasonable marks for AD classification standards. For example, A. Rueda, et al. [4], report balanced accuracies above 80% and 90% in different sample groups, and the best ranked proposal in CADDementia challenge [8] was able to reach a 97% score in control's true positive fraction, which in our case was 94.5% (PHOW-3D's true negative average, albeit not in identical datasets). Additionally, the gap in performance observed between histogram and PHOW representations could be partially attributed to the lack of spatial data in the first one. This highlights the importance of including shape representation in AD classification models.

The occurrence of observable shape patterns in the word averaged gradient patch hits, displays valuable capabilities for modeling surface-based shapes using PHOW. Likewise, for these average patches, visual differences in contrast and shape sharpness between AD and controls in the test volumes were found. This may be indicative of proper discriminant capabilities of the words i.e. increased patch-average contrast

	TP	TN	FP	FN	Accuracy
PHOW-3D	84.0%	94.5%	5.5%	16.0%	89.2%
PHOW-2D	84.1%	90.0%	10.0%	15.9%	87.1%
PHOW-2.5D	81.5%	91.0%	9.0%	19.5%	86.3%
Histogram	65.0%	70.0%	30.0%	35.0%	67.5%

Table 1. Average confusion matrix values and average accuracy for the best model parameters on each variant.

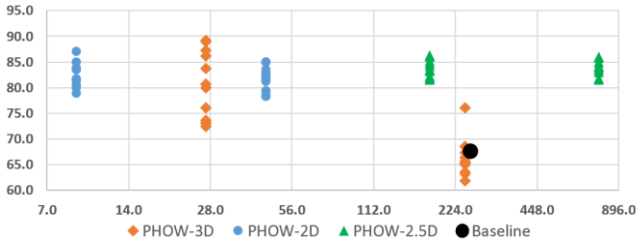

Fig. 2. Patch dimensionality (horizontal axis) vs. average accuracy for all parameter configurations used in the three PHOW variants.

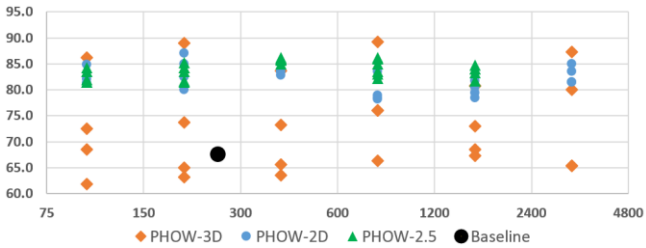

Fig. 3. Dictionary size (horizontal axis) vs. average accuracy for all parameter configurations used in the three PHOW variants.

and sharpness, indicate reduced shape variability, which could point to less noise in the patch-word quantization process. This behavior was observed in many word-averaged patches of test controls (Fig. 4), which is consistent with the results. Only one control and three AD cases were not labeled correctly by the best model in any PHOW variant.

Globally compared, performance differences between the three examined PHOW approaches appear to be smaller than expected. On the other hand, the inferior accuracy of PHOW-2.5D with respect to PHOW-2D was unanticipated. However, best results in both 3D and 2D variants used only one level grids. Such detail may indicate that what is essential in the representation of the pathology is global word distribution (local patterns) rather than word location inside the volumes, namely, mosaics configuration seem to induce 3D location implicitly. Moreover, differences in accuracy sensibility to patch and dictionary size, showed an inverse relation for classification performance order. PHOW-3D was the best method despite its larger variability. This behavior may be related, in part, with the steep increase of patch dimension and sample size in PHOW-3D. As the number of pyramid levels and grid division's increase, the model could steer to over fitting. Remarkably, PHOW-2.5D was only affected marginally by a larger patch dimensionality, in contrast to the behavior shown by PHOW-3D and PHOW-2D.

Another observed compelling aspect is that both patch dimension and dictionary size tuning seem to have the same importance, even though the final description dimensionality

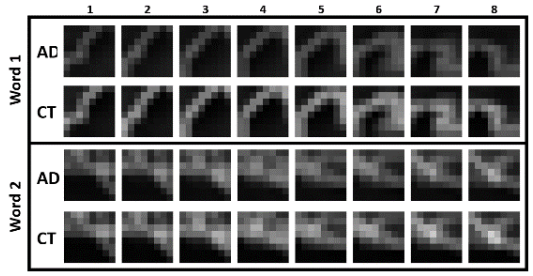

Fig. 4. Illustration of averaged 3D absolute-gradient 8x8x8 patches for two different words of best PHOW-3D model. Patch sequence (columns 1 to 8), follows the coronal axis. Averages were independently obtained over all AD or control (CT) volumes in the test set.

depends on dictionary size alone. Moreover, it can also be noticed that dictionary size and sensibility of the best PHOW-2D model were very similar to those reported by A. Rueda, et al. [4] for intensity based BoW. In addition, an inspection of classification results, revealed that the most complex subjects to classify correspond to healthy elderly (age above 85 years) and young subjects with early-onset dementia (CDR = 0.5 and age below 70 years), in agreement with previous reports [2]. In the first case, brain atrophy due to normal aging process seems to overlap with atrophy patterns of demented patients (CDR $\geq$ 1). Similarly, atrophy in subjects with early-onset AD are very subtle, thus easily mistaken as healthy young subjects. This unveils challenges with inter and intra-subject variability in AD classification.

Finally, even if our study framework may be restricted, we consider it useful for a wider context. For instance, results were influenced in non-obvious ways by the explored image configurations, emphasizing the importance of volume setup as a method's parameter. Likewise, our experimental approach, which provides an extensive view on the impact of dictionary hyper-parameters, give valuable insight on the potential of data-driven approaches in future complex classification task involving multiple diseases.

5. CONCLUSION

An exploration of parametric aspects, and a performance - sensibility analysis of three PHOW variants for global data-driven AD classification was presented. Accuracies were considered competitive, and differences between PHOW variants were narrow. It was found that controls were better classified, being elderly controls and young early onset AD subjects the most difficult, presumably due to overlapping atrophic patterns associated with intra-and inter-subject variability. Local patterns, besides word location, appeared essential for pathology description, and patch-averaged contrast and sharpness, indicated reduced noise during quantization and revealed discriminative surface features. Finally, the outcome highlights the potential value of shape descriptions for neurodegenerative disease classification.

As future work, we propose to further elucidate the properties of dictionary based volumetric descriptions with larger datasets, more pathological classes, increased shape representation strategies, and additional volume setups.

6. ACKNOWLEDGMENTS

The authors wish to thank Fundación CEIBA and Alcaldía Mayor de Bogotá, for the financial support of Ricardo Mendoza's PhD studies through the scholarship program "*Becas Rodolfo Llinás*", Amazon Inc., for providing valuable computing resources for this work through an "AWS in Education Research" grant, and CNRS - Lab-STICC Lab for providing financial aid to disseminate this contribution.

7. REFERENCES

[1] M. Prince, A. Wimo, and M. Guerchet, "World Alzheimer Report 2015," WHO, 2015.

[2] R. Cuingnet, E. Gerardin, et al., "Automatic classification of patients with Alzheimer's disease from structural MRI: a comparison of ten methods using the ADNI database," *Neuroimage*, vol. 56, no. 2, pp. 766–781, 2011.

[3] E. Bron, M. Smits, et al., "Feature selection based on SVM significance maps for classification of dementia," *Machine Learning in Medical Imaging*, Springer, pp. 272–279, 2014.

[4] A. Rueda, J. Arevalo, et al., "Bag of features for automatic classification of Alzheimer's disease in Magnetic Resonance Images," in *Progress in Patt. Rec., Image Analysis, Computer Vision, and Applications*, Springer, vol. 7441, pp. 559–566, 2012.

[5] A. R. Lopes Simoes, "Towards earlier detection of Alzheimer's disease using magnetic resonance images," PhD thesis, University of Twente, Netherlands, 2013.

[6] A. Rueda, F. González, et al., "Extracting salient brain patterns for imaging-based classification of neurodegenerative diseases,", *IEEE Trans on Med. Im.*, vol. 33, no. 6, pp. 1262–1274, 2014.

[7] P. Mondal, J. Mukhopadhyay, et al., "3D-SIFT feature based brain atlas generation: An application to early diagnosis of Alzheimer's disease," *Int. Conf. on Medical Imaging, m-Health and Emerging Communication Systems (MedCom)*, pp. 342–347, 2014.

[8] E. Bron, M. Smits, et al., "Standardized evaluation of algorithms for computer-aided diagnosis of dementia based on structural MRI: The CADDementia challenge," *NeuroImage*, vol. 111, pp. 562–579, 2015.

[9] A. Bosch, A. Zisserman, and X. Munoz, "Representing shape with a spatial pyramid kernel," *in Proc. of the 6th ACM int. conf. on image and video retrieval*, pp. 401–408, 2007.

[10] P. R. Raamana, H. Rosen, et al., "Three-Class Differential Diagnosis among Alzheimer Disease, Frontotemporal Dementia, and Controls," *Front. Neurol*., vol. 5, May 2014.

[11] S. Maji, A. Berg, and J. Malik, "Classification using intersection kernel support vector machines is efficient," in *2008 Proc. CVPR*, pp. 1–8, 2008.

[12] A. Ortiz, J. M. Górriz, et al., "LVQ-SVM based CAD tool applied to structural MRI for the diagnosis of the Alzheimer's disease," *Patt. Rec. Letters*, vol. 34, no. 14, pp. 1725–1733, Oct. 2013.

[13] D. Marcus, T. Wang, et al., "Open Access Se-ries of Imaging Studies (OASIS): cross-sectional MRI data in young, middle aged, non-demented, and demented older adults," *J. of cognitive neuroscience*, vol. 19, no. 9, pp. 1498–1507, 2007.

[14] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories," in *2006 Proc. CVPR*, vol. 2, pp. 2169–2178. 2006.