Inferring Disease Status by non-Parametric Probabilistic Embedding

Nematollah Kayhan Batmanghelich[1,2], Ardavan Saeedi[1],
Raul San Jose Estepar[2], Michael Cho[2], William M. Wells III[1,2]

[1]Computer Science and Artificial Intelligence Lab, MIT, Cambridge, USA,
{kayhan, ardavans}@mit.edu,
[2]Harvard Medical School, Brigham and Women's Hospital, Boston, USA
{rjosest@bwh, remhc@channing, sw@bwh}.harvard.edu

Abstract. Computing similarity between all pairs of patients in a dataset enables us to group the subjects into disease subtypes and infer their disease status. However, robust and efficient computation of pairwise similarity is a challenging task for large-scale medical image datasets. We specifically target diseases where multiple subtypes of pathology present simultaneously, rendering the definition of the similarity a difficult task. To define pairwise patient similarity, we characterize each subject by a probability distribution that generates its local image descriptors. We adopt a notion of affinity between probability distributions which lends itself to similarity between subjects. Instead of approximating the distributions by a parametric family, we propose to compute the affinity measure indirectly using an approximate nearest neighbor estimator. Computing pairwise similarities enables us to embed the entire patient population into a lower dimensional manifold, mapping each subject from high-dimensional image space to an informative low-dimensional representation. We validate our method on a large-scale lung CT scan study and demonstrate the state-of-the-art prediction on an important physiologic measure of airflow (the forced expiratory volume in one second, FEV1) in addition to a 5-category clinical rating (so-called GOLD score).

1 Introduction

As the size of an image dataset grows, the chance of observing more phenotypically similar patients increases. This premise makes analysis of large-scale image datasets attractive: subject similarities can reveal subtypes or the underlying biology of disease. In addition to the computational challenges of large datasets, defining robust image similarity measures in the presence of significant anatomical variation is a difficult task. Our approach targets heterogeneous diseases where the pathology in each patient can be thought of as a superposition of different processes, or subtypes of a disease. We propose a method that is computationally efficient and statistically robust. Our motivation comes from a study of Chronic Obstructive Pulmonary Disease (COPD), but the resulting model is applicable to a wide range of heterogeneous disorders.

A common method to compute similarities is based on image registration. Gerber *et al.* [5] applied pairwise registrations and defined similarity based on geodesic distance on the Riemannian manifold of diffeomorphic transformations. Hamm *et al.* [6] proposed a similar method except they restricted their analysis to a smaller subset of transformations and incorporated the residual of the registration into the similarity measure. Both methods rely on pairwise registration which is computationally demanding in large-scale settings and less applicable in the presence of large variations in anatomy. Unlike brain abnormalities in Alzheimers disease, the lung abnormalities in COPD are scattered and less localized [10]. This renders the definition of similarity between two images more challenging.

One approach to this challenge is to model image content as a *set* of local features. More specifically in the context of lung disease, Sorensen *et al.* [16] use histogram and texture features of local patches to create a binary ab/normal classification and suggest aggregation of the posterior probabilities to a subject-level score. Similarly Toews *et al.* [17] propose to represent images as collections of scale-invariant features and construct an approximate nearest neighbor graph of local features. To infer the subject-level score, they sum the log-likelihood function of the class associated with observed image features. In both cases, the presence of the patch-level labels [16] or subject-level labels [17] is required to infer the patient score. It is not clear how those methods can be applied in an unsupervised fashion.

We propose a general method that aggregates similarities from local-level image descriptors to infer subject-level similarities. The local descriptors are viewed as samples from subject-specific probability distributions; therefore the similarity between subjects is naturally reduced to a notion of similarity between probability distributions which should be estimated from their observed samples. Although a parametric approach can be used to infer the distributions for each subject [1], estimating those parameters can be computationally expensive if only pairwise similarities are of interest. Also, a misspecified parametric family biases the similarity estimation. We adopt a non-parametric approach proposed by Wang *et al.* [19] where the computation of similarity only depends on distances of each local feature from its k-nearest neighbors and does not require kernel density estimators (KDE). Using fast methods to approximate a nearest neighbor graph [12] enables us to achieve computational efficiency comparable to that of Toews *et al.* [17]. Another advantage is that no patch- or subject-level labels are required hence the method can be applied in an unsupervised fashion (*e.g.,* for sub-typing).

We illustrate an application of the method on a large-scale study of COPD. Our method outperforms the state-of-the-art approach in predicting clinical values related to COPD. We show how this method is used to embed the patient population in a lower dimensional space and its effectiveness in capturing disease structure in the embedding space.

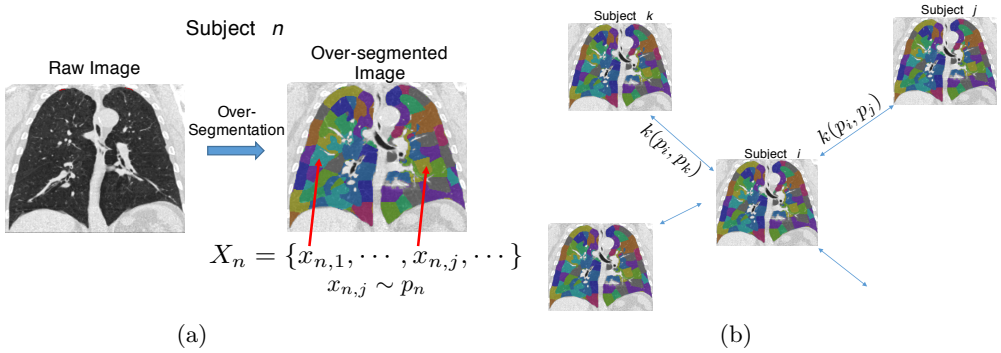


Fig. 1. (a) Feature extraction procedure for each subject. We extract local image descriptors (*e.g.,* $x_{n,j}$) from each super-pixel. X_n denotes the set of all local features from subject n. We model each subject with its corresponding probability density (*e.g.,* p_n). (b) Similarity graph between subjects. $k(p_i, p_j)$ denotes the similarity strength (affinity) between subjects i and j.

2 Method

In this section, we first describe the notation and the general setting. Then, we explain the algorithm to compute the pairwise patient similarities. Finally, we will explain how we use the similarity measurements to embed the patient population into a lower dimensional representation which is used to predict clinical values.

General Setting: Let each of $X_1, \cdots, X_N$ denote the *set* of local image features extracted from images of subjects $1, \cdots, N$ in the dataset. More specifically, we use an over-segmentation approach [7] to subdivide areas of a lung into groups of homogeneous super-pixels while preserving the boundaries of objects in the image. $X_n = \{x_{n,1}, \cdots, x_{n,m_n}\}$ is a set of image signatures extracted from m_n super-pixels where $x_{n,i} \in \mathbb{R}^d$ are local image descriptors extracted from region i of subject n. We will explore different options for the local descriptors in the experiment section of the paper. Following so-called "bag-of-words" representation [15], we model X_n as sample points from an unknown subject-specific distribution, $X_n \sim p_n$ (*i.e.,* $x_{n,i} \sim p_n$). We define similarity between subject i and subject j by defining a similarity measure between the corresponding distributions $k(p_i, p_j)$. We aim to estimate this quantity without estimating the underlying distribution. The general scheme is shown in Fig.1.

Distance between Distributions: To define similarity between images of two subjects given their observed bags of local descriptors $X \sim p$, $X' \sim q$, we need to define similarity between their corresponding distributions. We first define the distance between distributions and convert it to a similarity measure. We use

Kullback Leibler (KL) as the distance between distributions:

$$\mathrm{KL}(p\|q) = \int_{\mathbb{R}^d} \log \frac{p(x)}{q(x)} p(x) dx. \tag{1}$$

There is no closed-form for KL even for a mixture of two density distributions. We adopt a non-parametric approach proposed by Wang *et al.* [19] that does not require an explicit density estimation and estimates KL directly using a k-nearest neighbor graph.

Given sets of observations X, X' from the two probability distributions p, q, $X = \{x_i | x_i \sim p; i = 1, \cdots, N\}$ and $X' = \{x'_i | x'_i \sim q; i = 1, \cdots, M\}$, a $k-$nearest neighbor estimator of a point z only depends on the distance from z to the elements of X and X' [9]:

$$\hat{p}_k(z) = \frac{k/N}{\mathrm{vol}(z, \rho_k(z))} = \frac{k}{Nc\rho_k^d(z)}, \quad \hat{q}_k(z) = \frac{k/M}{\mathrm{vol}(z, \nu_k(z))} = \frac{k}{Mc\nu_k^d(z)}, \tag{2}$$

where $\mathrm{vol}(x, R)$ is the volume of a ball of radius R centered at z, $\rho_k(z)$ and $\nu_k(z)$ are the distance from the k'th nearest neighbor of z in the sets X and X' respectively, and c stands for the volume of a d-dimensional unit ball.

An unbiased estimator for the $\mathrm{KL}(p\|q)$ from the corresponding set of observed local descriptors, X and X' is the following:

$$\hat{\mathrm{KL}}_{N,M}(p\|q) = \frac{d}{N} \sum_{n=1}^{N} \log \frac{\nu_k(x_n)}{\rho_k(x_n)} + \log \frac{M}{N-1}. \tag{3}$$

Notice that the method directly estimates KL without estimating p and q and it only depends on the $k-$nearest neighbor distances (*i.e.,* $\rho_k(\cdot)$, $\nu_k(\cdot)$). The approximate $k-$nearest neighbor graph is constructed efficiently using [12]. Wang *et al.* [19] proved the estimator is asymptotically unbiased: $\lim_{N,M\to\infty} \mathbb{E}\left[\hat{\mathrm{KL}}_{N,M}(p\|q)\right] \to \mathrm{KL}(p\|q)$.

Subject-level Score Vector and Prediction: Let matrix L denote exponentiated symmetric KL distance; *i.e.,* $L_{ij} = \exp\left(-\mathrm{KL}_{\mathrm{sym}}(X_i, X_j)/\sigma^2\right)$ where $\mathrm{KL}_{\mathrm{sym}}(X_i, X_j) = \hat{\mathrm{KL}}(X_i\|X_j) + \hat{\mathrm{KL}}(X_j\|X_i)$. $\hat{\mathrm{KL}}(X_i\|X_j)$ is estimated using (3). We form the similarity kernel by projecting L on the positive definite cone as suggested by Chen *et al.* [4]. As suggested by Chang *et al.* [3], we set σ to the median value of the $\mathrm{KL}_{\mathrm{sym}}$ in the dataset in all of our experiments.

Computing the similarity matrix enables us to employ an embedding method and project each subject to a lower dimensional space by unfolding the manifold space of the subjects. To do that, we apply the Cholesky decomposition on the similarity matrix and feed the resulting factorization to a Linear Embedding (LLE) [20] algorithm and derive a lower dimensional subject-specific score vector. The resulting vector will be used for prediction of clinical measurements and visualization.

3 Experiments

In this section, we apply our method to a large-scale study of a COPD. We validate our method by predicting clinical measurements related to COPD and characterizing the disease continuum. The goal of this experiment is to compare the proposed method with classical baselines and investigate its robustness with respect to different choices of local image descriptors.

We apply our method on various local image descriptors and compare our performance with a global baseline feature and a classical representation method. As a baseline feature, we use two clinically important CT measurements of lung density, `INSP950` and `EXP950`. `INSP950`, the percentage of voxels $<$ -950HU, is a quantitative measure of emphysema. `EXP950`, the percentage of voxels $<$ -950HU after exhalation, reflects the degree of gas trapping [14, 11]. We also compare our approach with a classical representation method, Bag-of-Words (`BoW`), where images are represented by a histogram of words; words are clustered features from super-pixels. We used k-means clustering for `BoW`.

Data Preparation and Experimental Setting: We apply the method to CT images of lungs on 7292 subjects from the COPDGene study [13]. After automatic segmentation of the lung, we employ an over-segmentation approach [7] to subdivide areas of a lung into groups of spatially homogeneous super-pixels. We extract the following local features:

Histogram: Local histogram have been shown to be effective in characterizing emphysema [1, 2, 16]. We follow two procedures to extract histogram features. In the first, we extract a 32-bin histogram from each super-pixel (ref. as `Hist32`); 32 is roughly the third root of the average number of pixels in the super-pixel as suggested [16]. In the second procedure, we divide the histogram into 400 bins, followed by a PCA to reduce the dimensions to 30 (ref. as `HistPCA`) as suggested [1].

Texture: Texture features are shown to be important in characterizing lung tissue [16, 18]. Sorensen *et al.* [16] suggested using rotational invariant texture features. We adopt a rotation invariant histogram of gradient descriptors as proposed by Liu *et al.* [8]. Their method considers a gradient histogram as a continuous angular signal represented by the spherical harmonics (ref. as `sHOG`). We also extract `Harilick` features from the Gray-Level Co-occurrence Matrix (GLCM) following the pipeline [18] where the histogram information is already incorporated.

Evaluation: After computing the similarity matrix and the embedding vector scores (see Section 2), the resulting vectors are used as features in the following experiments. We use the Random Forest method to predict the `GOLD` score and linear Ridge regression (with the regularization weight set to 1) to estimate the continuous respiratory score (`FEV1`). Since neither of these clinical scores are derived from images, this experiment independently validates how well the embedding coordinates computed from the image similarity measure characterize

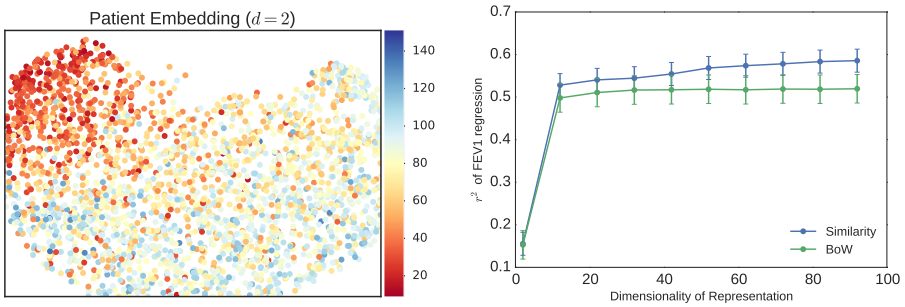
Table 1. Mean and bootstrap 95% confidence interval width (in parentheses) of the prediction performance for `GOLD` score and `FEV1`. The best results are shown in bold. The six first rows are the baseline methods: global feature and the traditional Bag-of-Words representation respectively.

		`FEV1`		`GOLD`
	Image Feature	r^2	MSE	% Accuracy
	`Baseline`	0.50 (0.03)	0.018 (0.001)	42.8 (1.5)
BoW	`Hist32`	0.51 (0.03)	0.018 (0.001)	47.2 (1.7)
BoW	`HistPCA`	0.51 (0.04)	0.018 (0.001)	47.3 (1.5)
BoW	`Hist32+sHOG`	0.56 (0.03)	0.016 (0.001)	47.2 (1.7)
BoW	`HistPCA+sHOG`	0.51 (0.04)	0.018 (0.001)	**47.2 (1.3)**
BoW	`Harilick`	0.33 (0.03)	0.025 (0.002)	39.6 (1.6)
Ours	`Hist32`	0.57 (0.03)	0.016 (0.001)	45.7 (1.5)
Ours	`HistPCA`	0.57 (0.03)	0.015 (0.001)	47.1 (1.7)
Ours	`Hist32+sHOG`	**0.59 (0.03)**	**0.015 (0.001)**	47.0 (1.8)
Ours	`HistPCA+sHOG`	0.57 (0.03)	0.015 (0.001)	**47.3 (1.7)**
Ours	`Harilick`	0.56 (0.03)	0.016 (0.001)	45.4 (1.6)

the underlying disease process. We report r^2 and the Mean Squared Error (MSE) of the prediction of `FEV1` and accuracy for `GOLD` score. We train on 99%, test on 1%, and repeat this process 50 times.

The results are reported in Table 1. All similarity-based predictions outperform the traditional threshold-based approach (*i.e.,* `Baseline`) irrespective of the local descriptors. To be comparable, we set the number of clusters in `BoW` to the dimensionality of our embedding method ($d = 100$). Our similarity-based representation outperforms `BoW` in r^2 and MSE and ties on accuracy. We computed p-values of the performance differences using a paired t-test. Our method is significantly better than the clinical image feature with $-\log p\text{-value} \gg 5$. The $-\log p$-values of the difference between the best performances of `BoW` and our method for r^2, MSE, and accuracy are 3.4, 3.2, and 0.01 respectively. The significant performance difference between `BoW` and our method for `Harilick` descriptor demonstrates robustness of the method with respect to choice of texture feature.

Fig.2 reports the effect of dimensionality of the representation on the prediction performance. Fig.2a shows the projection of patients on a 2D embedding space. A dot represents a patient and its color denotes `FEV1`. Even 2D embedding captures the structure of the disease; subjects on the bottom right are healthier than subjects on top left of the embedding space. Fig.2b reports the r^2 for `FEV1` with respect to dimensionality of the representation (*i.e.,* cluster size for `BoW` and embedding dim.) for `Hist32+sHOG` features. Both methods stabilize quickly in terms of performance and our method outperforms `BoW`.



(a) (b)

Fig. 2. (a) Embedding patients on a 2D space. A dot represents a patient and its color denotes `FEV1` (severity of COPD). Hotter colors indicate more severe disease. (b) Prediction performance (r^2) of `FEV1` with respect to the dimensionality of the embedding.

4 Conclusion

In this paper, we proposed to embed subject images into a manifold using an efficient pairwise similarity between probability distributions. We adopted a non-parametric approach requiring very few assumptions about the probability distributions that scales well as shown in our large-scale study. The entire process of computing similarities and the embedding takes less than few hours for all subjects (Python implementation). The experimental results showed that even projection on a two dimensional space can capture the continuum of the disease. This was evaluated quantitatively by predicting two clinical scores, none of which are derived from images, thus validating the benefits of the similarity-based method in characterizing the underlying disease process. Our approach can be used in longitudinal analysis to study disease exacerbation since we can associate coordinates in the embedding space to the clinical phenotype. Although we focus on COPD, our approach can be widely used in other scenarios particularly for heterogeneous diseases and when the bag-of-words model applies.

References

1. Batmanghelich, N.K., Saeedi, A., Cho, M., Estepar, R.S.J., Golland, P.: Generative method to discover genetically driven image biomarkers. International Conference on Information Processing and Medical Imaging 17(1), 30–42 (2015)
2. Castaldi, P.J., San José Estépar, R., Mendoza, C.S., Hersh, C.P., Laird, N., Crapo, J.D., Lynch, D.A., Silverman, E.K., Washko, G.R.: Distinct quantitative computed tomography emphysema patterns are associated with physiology and function in smokers. American journal of respiratory and critical care medicine 188(9), 1083–1090 (2013)
3. Chang, B., Kruger, U., Kustra, R., Zhang, J.: Canonical correlation analysis based on hilbert-schmidt independence criterion and centered kernel target alignment.

In: Proceedings of The 30th International Conference on Machine Learning. pp. 316–324 (2013)
4. Chen, Y., Garcia, E.K., Gupta, M.R., Rahimi, A., Cazzanti, L.: Similarity-based classification: Concepts and algorithms. The Journal of Machine Learning Research 10, 747–776 (2009)
5. Gerber, S., Tasdizen, T., Joshi, S., Whitaker, R.: On the manifold structure of the space of brain images. Med Image Comput Comput Assist Interv 12(Pt 1), 305–312 (2009)
6. Hamm, J., Ye, D.H., Verma, R., Davatzikos, C.: Gram: A framework for geodesic registration on anatomical manifolds. Med Image Anal 14(5), 633–642 (10 2010)
7. Holzer, M., Donner, R.: Over-segmentation of 3d medical image volumes based on monogenic cues. Cvww (JANUARY 2014), 35–42 (2014)
8. Liu, K., Skibbe, H., Schmidt, T., Blein, T., Palme, K., Brox, T., Ronneberger, O.: Rotation-invariant hog descriptors using fourier analysis in polar and spherical coordinates. International Journal of Computer Vision 106(3), 342–364 (2014)
9. Loftsgaarden, D.O., Quesenberry, C.P., et al.: A nonparametric estimate of a multivariate density function. The Annals of Mathematical Statistics 36(3), 1049–1051 (1965)
10. Lynch, D.A.: Progress in imaging copd, 2004-2014. Journal of the COPD Foundation Chronic Obstructive Pulmonary Diseases 1(2), 155–165 (2014)
11. Lynch, D.a., Al-Qaisi, M.a.: Quantitative computed tomography in chronic obstructive pulmonary disease. Journal of thoracic imaging 28(5), 284–90 (2013)
12. Muja, M., Lowe, D.G.: Scalable nearest neighbour algorithms for high dimensional data. IEEE Transactions on Pattern Analysis and Machine Intelligence 36(11), 2227–2240 (2014)
13. Regan, E.A., Hokanson, J.E., Murphy, J.R., Make, B., Lynch, D.A., Beaty, T.H., Curran-Everett, D., Silverman, E.K., Crapo, J.D.: Genetic epidemiology of copd (copdgene) study design. COPD: Journal of Chronic Obstructive Pulmonary Disease 7(1), 32–43 (2011)
14. Schroeder, J.D., McKenzie, A.S., Zach, J.A., Wilson, C.G., Curran-Everett, D., Stinson, D.S., Newell, J.D., Lynch, D.A.: Relationships between airflow obstruction and quantitative ct measurements of emphysema, air trapping, and airways in subjects with and without chronic obstructive pulmonary disease. American Journal of Roentgenology 201(3) (2013)
15. Sivic, J., Zisserman, A.: Efficient visual search of videos cast as text retrieval. Pattern Analysis and Machine Intelligence, IEEE Transactions on 31(4), 591–606 (2009)
16. Sorensen, L., Nielsen, M., Lo, P., Ashraf, H., Pedersen, J.H., De Bruijne, M.: Texture-based analysis of copd: A data-driven approach. IEEE Transactions on Medical Imaging 31(1), 70–78 (2012)
17. Toews, M., Wachinger, C., Estepar, R.S.J., Wells, W.M.: A feature-based approach to big data analysis of medical images. Information Processing in Medical Imaging (IPMI) 24, 339–50 (2015)
18. Vogl, W.D., Prosch, H., Muller-Mang, C., Schmidt-Erfurth, U., Langs, G.: Longitudinal alignment of disease progression in fibrosing interstitial lung disease. In: Lecture Notes in Computer Science. vol. 8674 LNCS, pp. 97–104 (2014)
19. Wang, Q., Kulkarni, S.R., Verdú, S.: Divergence estimation for multidimensional densities via-nearest-neighbor distances. Information Theory, IEEE Transactions on 55(5), 2392–2405 (2009)
20. Zhang, Z., Wang, J.: Mlle: Modified locally linear embedding using multiple weights. Advances in Neural Information Processing Systems pp. 1593–1600 (2006)