



Published in final edited form as:

*Inf Process Med Imaging*. 2015 ; 24: 339–350.

## A Feature-based Approach to Big Data Analysis of Medical Images

Matthew Toews<sup>1</sup>, Christian Wachinger<sup>2</sup>, Raul San Jose Estepar<sup>1</sup>, and William M. Wells III<sup>1</sup>

<sup>1</sup> Brigham and Women's Hospital, Harvard Medical School

<sup>2</sup> Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology  
wachinger@csail.mit.edu

### Abstract

This paper proposes an inference method well-suited to large sets of medical images. The method is based upon a framework where distinctive 3D scale-invariant features are indexed efficiently to identify approximate nearest-neighbor (NN) feature matches in  $O(\log N)$  computational complexity in the number of images  $N$ . It thus scales well to large data sets, in contrast to methods based on pair-wise image registration or feature matching requiring  $O(N)$  complexity. Our theoretical contribution is a density estimator based on a generative model that generalizes kernel density estimation and K-nearest neighbor (KNN) methods. The estimator can be used for on-the-fly queries, without requiring explicit parametric models or an off-line training phase. The method is validated on a large multi-site data set of 95,000,000 features extracted from 19,000 lung CT scans. Subject-level classification identifies all images of the same subjects across the entire data set despite deformation due to breathing state, including unintentional duplicate scans. State-of-the-art performance is achieved in predicting chronic pulmonary obstructive disorder (COPD) severity across the 5-category GOLD clinical rating, with an accuracy of 89% if both exact and one-off predictions are considered correct.

### 1 Introduction

Systems for storing and transmitting digital data are increasing rapidly in size and bandwidth capacity. Data collection projects such as COPDGene (19,000 lung CT scans of 10,000 subjects) [1] offer unprecedented opportunities to learn from large medical image sets, for example to discover subtle aspects of anatomy or pathology only observable in subsets of the population. For this, image processing algorithms must scale with the quantities of available data.

Consider an algorithm designed to discover and characterize unknown clinical phenotypes or disease subclasses from a large set of  $N$  images. Two major challenges that must be addressed are computational complexity and robust statistical inference. The fundamental operation required is often image-to-image similarity evaluation, which incurs a prohibitive computational complexity cost of  $O(N)$  when performed via traditional pair-wise methods

such as registration [2, 3] or feature matching [4, 5]. For example, computing the pairwise affinity matrix between 416 low resolution brain scans via efficient deformable registration requires about one week on a 50GHz cluster computer [2]. Furthermore, robustly estimating variables of interest requires coping with myriad confounds, including arbitrarily misaligned data, missing data due to resection or variable scan cropping (e.g. in order to reduce ionizing radiation exposure), inter-subject anatomical variability including abnormality, intra-subject variability due to growth and deformation (e.g. lung breathing state), inter-scanner variability in multi-site data, to name a few.

The technical contribution of this paper is a computational framework that provides a solution to both of these challenges. The framework is closely linked to large-scale data search methods where data are stored and indexed via nearest neighbor (NN) queries [6, 10]. Images are represented as collections of distinctive 3D scale-invariant features [7, 8], information-rich observations of salient image content. Robust, local image-to-image similarity is computed efficiently in  $O(\log N)$  computational complexity via approximate nearest neighbor search.

The theoretical contribution is a novel estimator for class-conditional densities in a Naive Bayes classification formulation, representing a hybrid of kernel density [11] and KNN [12] methods. A mixture density is estimated for each input feature of a new image as the weighted sum of 1) a variable bandwidth kernel density computed from a set of nearest neighbors and 2) a background distribution over unrelated features. This estimator achieves state-of-the-art performance in automatically predicting the 5-class GOLD severity label from a large multi-site data set, improving upon previous results based on lung-specific image processing pipelines and single-site data [13, 28, 29]. Furthermore, subject-level classification is used to identify all instances duplicate subjects in a large set of 19,000 subjects, despite deformation due to breathing state.

## 2 Related Work

Our work derives from two main bodies of research, local invariant image feature methods and density estimation. Here, a local feature refers to a salient image patch or region identified via an interest operator, e.g. extrema of the difference-of-Gaussian operator in the popular scale-invariant feature transform (SIFT) [7]. Local features effectively serve as a reduced set of information-rich image content that enable highly efficient image processing algorithms, for example feature correspondence in  $O(\log N)$  computational complexity via fast approximate NN search methods [9, 10]. Early feature detection methods identified salient image locations [14], scale-space theory [15] lead to the development of so-called scale-invariant [7] and affine-invariant [16] feature detection methods capable of repeatedly detecting the same image pattern despite global similarity and affine image transforms, in addition to intensity variations.

Once detected, salient image regions are cropped and spatially normalized to patches of size  $D$  voxels, then encoded as compact, highly distinctive descriptors for efficient indexing, e.g. the gradient orientation histogram (GoH) representation [7]. Note that there is a tradeoff between the patch dimension  $D$  and the number of unique samples available to populate the

space of image patches  $\mathbb{R}^D$ . At one extreme, voxel-size patches lead to a densely sampled but relatively uninformative  $R^1$  space, while at the other extreme, entire images provide an information-rich but severely under-sampled  $R^N$  space. Intermediately-sized observations have been shown to be most effective for tasks such as classification [17]. Note also that many patches are rarely observed in natural medical images, and the typical set of patches is concentrated within a subspace or manifold  $\mathcal{M} \in \mathbb{R}^D$ . Furthermore, local saliency operators further restrict the manifold to a subset of highly informative patches  $\mathcal{M} \bullet \in \mathcal{M}$ , as common, uninformative image patterns are not detected, e.g. non-localizable boundary structures or regions of homogenous image intensity.

Salient image content can thus be modeled as a set of local features, e.g. within a spatial configuration or as an unordered bag-of-features representation when inter-feature spatial relationships are difficult to model. Probabilistic models for inference typically require estimating densities from feature data. Non-parametric density estimators such as KDE or KNN estimators are particularly useful as an explicit model of the joint distribution is not required. They can be computed on-the-fly without requiring computationally expensive training, e.g. via instance-based or lazy-learning methods [18, 13, 19]. KDE seeks to quantify density from kernel functions centered around training samples [20, 11], whereas KNN estimators seek to quantify the density at a point from a set of K nearest neighbor samples [12, 18, 13, 19]. An interesting property of KNN estimators is that when used in classification, their prediction error is asymptotically upper bounded by no more than twice the optimal Bayes error rate as the number of data grow [12]. This property is particularly relevant given the increasing size of medical image data sets.

In the context of medical image analysis, scale-invariant features have been used to align and classify 3D medical image data [8, 21, 22], however they have not yet been adapted to large-scale indexing and inference. Although our work here focuses on inference, the feature-based correspondence framework is generally related to medical image analysis methods using nearest neighbors or proximity graphs across image data, including subject-level recognition [5], manifold learning [3, 2, 23], in particular methods based on local image characteristics [24, 25] and multi-atlas labeling [26].

The experimental portion of this paper investigates chronic pulmonary obstructive disorder (COPD) in a large set of multi-site lung CT images. A primary focus of COPD imaging research has been to characterize and classify disease phenotypes. Song et al. investigate various 2D feature descriptors for classifying lung tissues including local binary patterns and gradient orientation histograms [27]. Several authors propose subject-level COPD prediction as an avenue of exploratory research. Sorensen et al. [13] use texture patches in a binary classification COPD = (0,1) scenario on single-site data to achieve an area-under-the-curve (AUC) classification of 0.71. Mets et al. [28] use densitometric measures computed from single-site data of 1100 male subjects, to achieve an AUC value of 0.83 for binary COPD classification. Gu et al. [29] use automatic lung segmentation and densitometric measures to classify single-site data according to the GOLD range, achieving an exact classification rate of 0.37, or 0.83 if classification into neighboring categories is considered correct. A major challenge is classifying multi-site data acquired across different sites and scanners. On a large multi-site data set, our method achieves exact and one-off classification rates of 0.48

and 0.89, respectively, which to our knowledge are the highest rates reported for GOLD classification.

### 3 Method

#### 3.1 Estimating Class Probabilities

Let  $f_i \in \mathbb{R}^D$  be a D-dimensional vector encoding the appearance of a scale-normalized image patch, e.g. a scale-invariant feature descriptor, and let  $F = \{f_i\}$  be a set of such features extracted in an image. Let  $C$  be a clinical variable of interest, e.g. a discrete measure of disease severity, defined over a set of  $M$  values  $[1, \dots, m]$ . Finally let  $C_i$  represent the value of  $C$  associated with feature  $f_i$ .

We seek the posterior probability  $p(C|F)$  of clinical variable  $C$  conditioned on feature data  $F$  extracted in a query image, which can be expressed as

$$p(C|F) \propto p(C) p(F|C) = p(C) \prod_i p(f_i|C). \quad (1)$$

In Equation (1), the first equality follows from Bayes' rule, and the second from the so-called Naive Bayes assumption of conditional feature independence. This strong assumption is often made for computational convenience when modeling the true joint distribution over all features  $F$  is intractable. Nevertheless, it often leads to robust, effective modeling even in contexts where conditional independence does not strictly hold. Conditional independence is reasonable in the case of local image observations  $f_i$ , as patches separated in scale and space do not typically exhibit direct correlations. On the RHS of Equation (1),  $p(C)$  is a prior distribution over the clinical variable of interest  $C$ , and  $p(f_i|C)$  is the likelihood function of  $C$  associated with observed image feature  $f_i$ .

We use a robust variant of kernel density estimation for calculating the class conditional likelihood densities:

$$p(f_i|C) \propto \left[ \sum_{j: C_j=C} \frac{N}{N_C} \exp\left(-\frac{d^2(f_i, f_j)}{\alpha^2 + 1}\right) \right] + \beta \frac{N_C}{N}. \quad (2)$$

Here  $N_C$  is the number of features of class  $C$  in the training data, and  $N = \sum_C N_C$  is the total feature count.  $d(f_i, f_j)$  is the distance between  $f_i$  and neighboring descriptor  $f_j$ , here the Euclidean distance between descriptors.  $\alpha$  is an adaptive kernel bandwidth parameter that is empirically set to  $d_{NN_i}$  for each input feature  $f_i$ , i.e. the distance between  $f_i$  and the nearest neighboring descriptor  $f_j$  in a data base of training data:

$$d_{NN_i} = \min_j \{d(f_i, f_j)\}. \quad (3)$$

Finally,  $\beta$  is a weighting parameter empirically set to  $\beta = 1$  in experiments for best performance. Note that the overall scale of the likelihood is unimportant, as normalization can be performed after the product in Equation (1) is computed.

In practice, Equation 2 is computed for each  $f_i$  from a set of  $KNN_i = \{(f_j, C_j)\}$  of  $K$  feature/label pairs  $(f_j, C_j)$  identified via an efficient approximate KNN search over a set of training feature data. Because the adaptive exponential kernel falls off quickly, it is not crucial to determine an optimal  $K$  as in some KNN methods [18], but rather to set  $K$  large enough include all features contributing to the kernel sum. Intuitively, the two terms in Equation 2 are designed as a mixture model that is aimed at increasing the robustness of estimates when some of the features are “uninformative”. The first term is a density estimator that accounts for informative features in the data. It is a variant that combines aspects of kernel density estimation and  $KNN$  density estimation, using a kernel where the bandwidth is scaled by the distance to the first nearest neighbor as in Breiman et al. [20]. The second term  $\beta \frac{N_C}{N}$  provides a default estimate for the case of uninformative features, curiously this class-specific value results in noticeably superior classification performance than a value that is uniform across classes.

### 3.2 Computational Framework

To scale to large data sets of medical images, our inference method focuses on rapidly indexing a large set of image features. A variety of local feature detectors exist, we adopt a 3D generalization of the SIFT algorithm [8], where the location and scale of distinctive image patches are detected as extrema of a difference-of-Gaussian operator. Once detected, patches are reoriented, rescaled to a fixed size ( $11^3$  voxels) and transformed into a GoH representation over 8 spatial bins and 8 orientation bins, resulting in a 64-element feature descriptor. Finally, rank-ordering[30] transforms descriptor elements into an ordinal representation, where elements take on their rank in an array sorted according to GoH value. Once extracted, descriptors can be stored in tree data structures for efficient NN indexing. Again,  $d(f_i, f_j)$  can be defined according to a variety of measures such as geodesic distance, here we adopt the Euclidean distance between descriptor elements. Exact NN search is difficult for high dimensional descriptors, however approximate KNN methods can be used to identify NNs with high probability in  $O(\log N)$  search time, for example via randomized K-D search tree [9].

## 4 Experiments

Experiments focus on analyzing Chronic Obstructive Pulmonary Disorder (COPD), an important health problem and a leading cause of death. We test our method on lung CT images from the COPDGene data set [1] acquired for the purpose of characterizing COPD phenotypes and associated genetic links. The COPDGene dataset consists of 19,000 lung CT images of 10,300 subjects with clinical and demographic labels, where expiration and inspiration images are acquired for most subjects. Data are acquired at 21 clinical centers with CT scanners from a variety of different vendors, making the dataset a diverse multi-site test bed for practical big data algorithms. The clinical COPD measurement of interest is the GOLD score, which quantifies COPD severity on a scale of 0-4 based on spirometry measurements.

The only data preprocessing step in our pipeline is 3D scale-invariant feature extraction from images using the implementation described in [8]. Note that feature extraction is robust

to variations in image geometry and intensity, and domain-specific pre-processing steps such as lung segmentation are unnecessary. Feature extraction is a one-time processing step, requiring on the order of 20 seconds for an image of size  $256^3$  voxels (0.6mm isotropic resolution), after which features can be efficiently indexed. All feature extraction was performed on a commodity cluster computing system over the course of several hours. Each lung CT image results in approximately 5000 features, and the set of 19,000 images results in a total of 95,000,000 features. While the original image data occupies 3.8TB when gzip-compressed, the feature data requires only 8.6GB, representing a data reduction of 440X. Feature extraction can thus be viewed as a form of lossy compression, where the goal is to retain as much salient information as possible while significantly reducing the memory footprint. Given the relatively small size and usefulness of feature data, it may be useful to include it as part of a standard markup for efficiently indexing future image formats, e.g. DICOM extensions.

#### 4.1 Computer-assisted COPD Prediction

A primary goal of the COPDGene project is to identify disease phenotypes in order to better understand and characterize COPD. To this end, we investigate computer-assisted prediction of COPD, based on the five GOLD categories ranging from 0 to 4 in increasing order of severity. For clarity we experiment with a balanced set of data with 523 images per GOLD category, for a total of 2615 images and 13,000,000 features. We use only expiratory images in order to evaluate algorithm performance in isolation from confounds such as shape change during breathing. Maximum a-posterior (MAP) estimation is used to predict the most probable GOLD score  $C_{MAP}$  for each new feature set  $F$  in a leave-one-out manner:

$$C_{MAP} = \underset{C}{\operatorname{argmax}} \{p(C|F)\}. \quad (4)$$

Note that for experiments, the prior  $p(C)$  from Equation (1) is taken to be uniform. The leave-one-out methodology is implemented efficiently using the K-D search tree method of Muja and Lowe [9] to compute NN correspondences. In this method, features are indexed in a set of independent search trees, whose data splits are chosen randomly amongst the subset of feature descriptor elements exhibiting the highest variability. As a figure of merit we consider both the accuracy of exact prediction and one-off prediction, i.e. where  $C_{MAP}$  predicts a GOLD score one off from the true label. Prediction is also tested on various training set sizes, in order to investigate the effect on prediction accuracy. Graphs of prediction results are shown in Figure 1, and Table 1 lists the confusion matrix for prediction on 2615 training subjects.

In our method, feature-wise densities  $p(f_i|C)$  quantify the informativeness of individual features  $f_i$  with respect to labels  $C_i$ , e.g. as in feature-based morphometry [21]. These densities may be useful in investigating and characterizing COPD phenotypes, this is a topic of ongoing investigation in our group. Figure 2 shows the 20 features with the highest  $p(f_i|C)$  for images from either extreme of the GOLD severity rating.

Parameter  $K$  is typically important in KNN estimation methods, however our method does not vary significantly with changes in  $K$  above a certain value due to the drop-off of the



adaptive exponential kernel. Figures 3 a) and b) illustrate inter-feature distances and their weighted kernel values for several typical features across  $K=100$  neighbors. The result of prediction tends to stabilize for  $K=100$ . We attempted KNN density estimation for various values of  $K$  via standard counting as in [13], however classification performance was relatively poor and varied noticeably with values of  $K$ . We found the best means of improving performance was to adopt the kernel weighting scheme in Equation (2).

Although the training sets used here for prediction are balanced in terms of the number of subjects, individual images produce different numbers of features. Figure 4 a) shows the feature counts  $N_C$  for GOLD categories. Figures 4 b) and c) show how changes in either  $\alpha$  or  $\beta$  result in prediction that is noticeably skewed towards or away from GOLD categories (e.g. here GOLD 1, 3 or 4) with higher feature counts.

## 4.2 Subject-level Indexing

Large, multi-site image data sets can quickly become difficult to manage. Image labeling errors may be introduced in DICOM headers [31], and images of subjects may be inadvertently duplicated, removed or modified, compromising the data integrity and usefulness. We propose subject-level indexing to identify all instances of the same subject within a data set, in a manner similar to recent work in brain imaging [5], in order to inspect and verify data integrity. Subject ID is used as the clinical label of interest  $C$ , and inference seeks to identify highly probable labels  $p(C|F)$  given feature data  $F$  from a test subject. Note that inference must be robust to large deformations due to inhale and exhale state, our method accomplishes this purely from local feature appearance information, parameters of feature geometry are not used (i.e. image location, scale, orientation).

Subject-level indexing effectively computes the image-to-image affinity matrix between all 19,000 image feature sets. The processing time is  $\approx 6$  hours on a laptop (MacBook Pro) using a single core, with a breakdown of  $\approx 1$  hr for data read-in and  $\approx 5$  hrs for KNN feature correspondence ( $\approx 1$  second per subject). This is effectively equivalent to  $\approx 180,000,000$  pair-wise image registrations (assuming a symmetric registration technique). Note that here, correspondences are established across significant lung shape variation due to breathing state, as both expiration and inspiration are used. To our knowledge, all images are correctly grouped according to subject labels, using an empirically determined threshold on the posterior probability. A set of 65 images are flagged as abnormal, either duplicate images (unusually high posterior probability indicating identical feature sets) or different images of the same subject (high posterior probability). A partial list of 20 unintentional duplicate subject scans was compiled from genetic information, all were successfully identified as same subjects via subject-level indexing. The remaining abnormalities are currently being investigated.

## 5 Discussion

We presented a general method for analyzing large sets of medical images, based on nearest neighbor scale-invariant feature correspondences and kernel density estimation. The method scales well to large medical image sets in two respects. First, efficient approximate NN search techniques can be used to achieve correspondence in  $O(\log N)$  computational

complexity, as to opposed  $O(N)$  pairwise image or feature matching algorithms which quickly become intractable for large data sets. Second, probabilistic inference can be performed on-the-fly from feature data, without parametric models or potentially expensive training procedures. A hybrid KDE-KNN kernel density estimator with an adaptive bandwidth parameter is used to robustly estimate likelihood factors from nearest neighbor features.

Our method is demonstrated on 19,000 lung CT images of 10,300 subjects from the multi-site COPDGene data set. Subject-level indexing demonstrates that images of the same subject can be robustly identified across deformation due to breathing, and erroneous instances of subject duplication can be flagged. State-of-the-art results are obtained for multi-site multi-class prediction of clinical GOLD scores, improving on methods involving special purpose lung segmentation and densitometric measures. The prediction result is important, because it suggests the existence of disease-related anatomical patterns that could help to better understand COPD. It may be that the 3D SIFT representation is particularly well-tuned to anatomical structure of lung parenchyma related to COPD. Future work will focus on analysis of disease-informative features, identifying disease phenotypes. One avenue will be to incorporate feature geometry (e.g. location, scale) within modeling. Finally, our method is general and could be used to organize large sets of general medical image data, e.g. brain or full-body scans. Software described in this paper will be provided to the public for research use.

We believe there is a good deal of potential for studying large sets of medical images via the local feature framework, e.g. 3D SIFT features or other suitable data-driven extractors. While they may be a coarse approximation to the original image, local features often contain enough salient information to robustly and efficiently perform tasks such as registration or classification. This is particularly true where the quantity of data an algorithm is capable of exploiting begins to compensate for the coarseness of its representation, i.e. via efficient search methods. The general framework can be used to efficiently generate proximity graphs between large sets of medical image data, and may thus be useful in the context of other computational approaches such as manifold learning [32, 2].

## Acknowledgements

This research was supported by NIH grants P41EB015902, P41EB015898 5K25HL104085, 5R01HL116931 and 5R01HL116473.

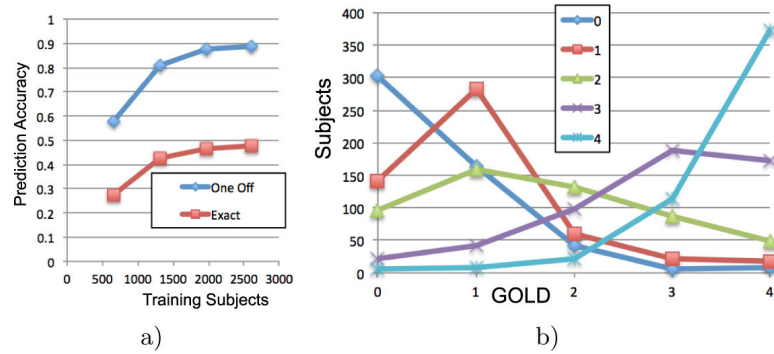
## References

1. Regan EA, Hokanson JE, Murphy JR, Make B, Lynch DA, Beaty TH, Curran-Everett D, Silverman EK, Crapo JD. Genetic epidemiology of COPD (COPDGene) study design. COPD: Journal of Chronic Obstructive Pulmonary Disease. 2011; 7(1):32–43. [PubMed: 20214461]
2. Gerber S, Tasdizen T, Thomas Fletcher P, Joshi S, Whitaker R. Manifold modeling for brain population analysis. Medical image analysis. 2010; 14(5):643–653. [PubMed: 20579930]
3. Hamm J, Ye DH, Verma R, Davatzikos C. Gram: A framework for geodesic registration on anatomical manifolds. Medical image analysis. 2010; 14(5):633–642. [PubMed: 20580597]
4. Toews M, Zöllei L, Wells WM III. Invariant feature-based alignment of volu-metric multi-modal images. IPMI. 2013:25–36.

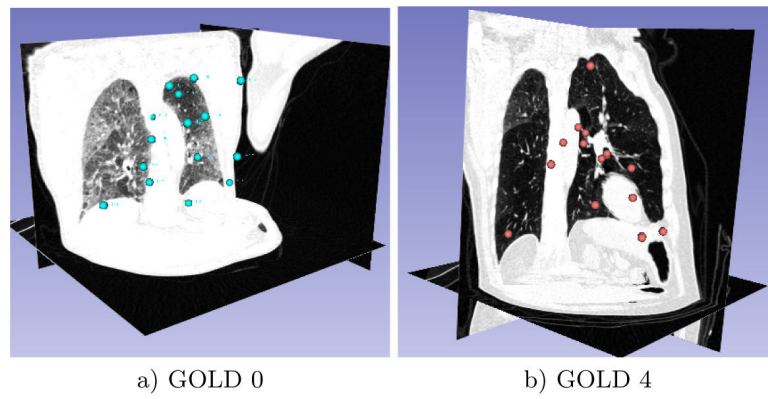


5. Wachinger, C.; Golland, P.; Reuter, M. MICCAI. 2014. Brainprint: Identifying subjects by their brain.; p. 41–48.
6. Brin S. Near neighbor search in large metric spaces. VLDB. 1995:574–584.
7. Lowe DG. Distinctive image features from scale-invariant keypoints. IJCV. 2004; 60(2):91–110.
8. Toews M, Wells WM III. Efficient and robust model-to-image alignment using 3d scale-invariant features. Medical Image Analysis. 2013; 17(3):271–282. [PubMed: 23265799]
9. Muja M, Lowe D. Scalable nearest neighbour algorithms for high dimensional data. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2014; 36(11):2227–2240. [PubMed: 26353063]
10. Kleinberg JM. Two algorithms for nearest-neighbor search in high dimensions. ACM symposium on Theory of computing. 1997:599–608.
11. Terrell GR, Scott DW. Variable kernel density estimation. The Annals of Statistics. 1992:1236–1265.
12. Cover T, Hart P. Nearest neighbor pattern classification. Information Theory, IEEE Transactions on. 1967; 13(1):21–27.
13. Sorensen L, Nielsen M, Lo P, Ashraf H, Pedersen JH, De Bruijne M. Texture-based analysis of COPD: a data-driven approach. Medical Imaging, IEEE Transactions on. 2012; 31(1):70–78.
14. Moravec HP. Visual mapping by a robot rover. Proc. of the 6th International Joint Conference on Artificial Intelligence. 1979:598–600.
15. Lindeberg T. Feature detection with automatic scale selection. IJCV. 1998; 30(2):79–116.
16. Mikolajczyk K, Schmid C. Scale and affine invariant interest point detectors. IJCV. 2004; 60(1): 63–86.
17. Ullman S, Vidal-Naquet M, Sali E. Visual features of intermediate complexity and their use in classification. Nature Neuroscience. 2002; 5(7):682–687. [PubMed: 12055634]
18. Toussaint GT. Proximity graphs for nearest neighbor decision rules: recent progress. Interface. 2002; 34
19. Yu A, Grauman K. Predicting useful neighborhoods for lazy local learning. Advances in Neural Information Processing Systems. 2014:1916–1924.
20. Breiman L, Meisel W, Purcell E. Variable kernel estimates of multivariate densities. Technometrics. 1977; 19(2):135–144.
21. Toews M, Wells WM III, Collins DL, Arbel T. Feature-based morphometry: Discovering group-related anatomical patterns. NeuroImage. 2010; 49(3):2318–2327. [PubMed: 19853047]
22. Gill G, Toews M, Beichel RR. Robust initialization of active shape models for lung segmentation in ct scans: A feature-based atlas approach. International journal of biomedical imaging. 2014:2014
23. Singh N, Thomas Fletcher P, Samuel Preston J, King RD, Marron J, Weiner MW, Joshi S. Quantifying anatomical shape variations in neurological disorders. Medical image analysis. 2014; 18(3):616–633. [PubMed: 24667299]
24. Aljabar P, Wolz R, Srinivasan L, Counsell SJ, Rutherford MA, Edwards AD, Hajnal JV, Rueckert D. A combined manifold learning analysis of shape and appearance to characterize neonatal brain development. Medical Imaging, IEEE Transactions on. 2011; 30(12):2072–2086.
25. Ye, DH.; Hamm, J.; Kwon, D.; Davatzikos, C.; Pohl, KM. MICCAI. Springer; 2012. Regional manifold learning for deformable registration of brain mr images.; p. 131–138.
26. Aljabar P, Heckemann RA, Hammers A, Hajnal JV, Rueckert D. Multi-atlas based segmentation of brain images: atlas selection and its effect on accuracy. Neuroimage. 2009; 46(3):726–738. [PubMed: 19245840]
27. Song Y, Cai W, Zhou Y, Feng DD. Feature-based image patch approximation for lung tissue classification. IEEE TMI. 2013; 32(4):797–808.
28. Mets OM, Buckens CF, Zanen P, Isgum I, van Ginneken B, Prokop M, Gietema HA, Lammers JWJ, Vliegenthart R, Oudkerk M, et al. Identification of chronic obstructive pulmonary disease in lung cancer screening computed tomographic scans. JAMA. 2011; 306(16):1775–1781. [PubMed: 22028353]

29. Gu S, Leader J, Zheng B, Chen Q, Sciurba F, Kminski N, Gur D, Pu J. Direct assessment of lung function in COPD using CT densitometric measures. *Physiological measurement*. 2014; 35(5):833. [PubMed: 24710855]
30. Toews M, Wells WM III. SIFT-RANK: Ordinal description for invariant feature correspondence. *Computer Vision and Pattern Recognition (CVPR), 2009 IEEE Conference on, IEEE*. 2009:172–177.
31. Guld, MO.; Kohnen, M.; Keysers, D.; Schubert, H.; Wein, B.; Bredno, J.; Lehmann, TM. *Int. Symposium on Medical Imaging*. Vol. 4685. SPIE; 2002. Quality of DICOM header information for image categorization.; p. 280-287.
32. Torki, M.; Elgammal, A. *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on. IEEE; 2010. Putting local features on a manifold.; p. 1743-1750.*

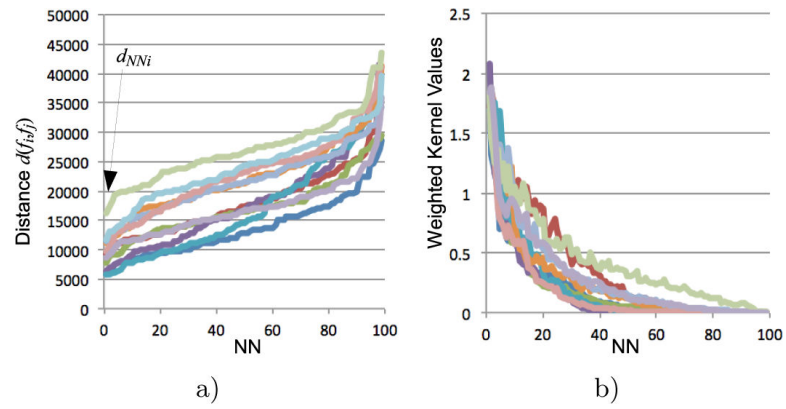
**Fig. 1.**

a) GOLD prediction accuracy (one-off and exact) as a function of the number of training subjects. b) Curves for predicted vs. actual GOLD values for all 2615 training subjects. State-of-the art classification is achieved, with an accuracy of 48% for exact prediction and 89% for one-off prediction. Note the gradual transitions across predicted GOLD labels in b), as expected in the case of a COPD severity continuum.

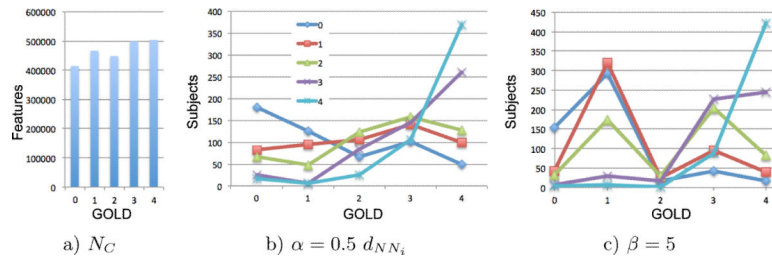


**Fig. 2.**

Visualizing the 20 disease-informative features with the highest  $p(f_i|C)$  for a) GOLD 0 and b) GOLD 4. Informative features are typically scattered throughout the lungs and range in size from from 2-5mm. Note feature scale is not displayed.

**Fig. 3.**

a) NN feature distance  $d(f_i, f_j)$  for 10 typical features  $f_i$  and  $K=100$  neighbors  $f_j$ , sorted by increasing distance right-to-left. b) Weighted kernel values of Equation (2) for the same features and neighbors, note kernel values become negligible by  $K=100$ .

**Fig. 4.**

Graph a) illustrates feature counts  $N_C$  over GOLD categories. Graphs b) and c) illustrate skewed prediction from unoptimal kernel parameters in Equation (2), b)  $\alpha = 0.5 d_{NN_i}$  and c) from  $\beta = 5$ .



**Table 1**

Confusion matrix for COPD GOLD category prediction, 523 subjects per category using K=100. Bold values indicate exact prediction.

GOLD	Predicted GOLD				
	0	1	2	3	4
0	<b>303</b>	164	43	5	8
1	141	<b>283</b>	60	21	18
2	95	160	<b>132</b>	87	49
3	21	43	98	<b>188</b>	173
4	5	9	21	114	<b>374</b>