Dissimilarity representations in lung parenchyma classification

Lauge Sørensen[a] and Marleen de Bruijne[a,b]

[a]Department of Computer Science, University of Copenhagen, Copenhagen, Denmark;
[b]Biomedical Imaging Group Rotterdam, Departments of Radiology & Medical Informatics, Erasmus MC, Rotterdam, The Netherlands

ABSTRACT

A good problem representation is important for a pattern recognition system to be successful. The traditional approach to statistical pattern recognition is feature representation. More specifically, objects are represented by a number of features in a feature vector space, and classifiers are built in this representation. This is also the general trend in lung parenchyma classification in computed tomography (CT) images, where the features often are measures on feature histograms. Instead, we propose to build normal density based classifiers in dissimilarity representations for lung parenchyma classification. This allows for the classifiers to work on dissimilarities between objects, which might be a more natural way of representing lung parenchyma. In this context, dissimilarity is defined between CT regions of interest (ROI)s. ROIs are represented by their CT attenuation histogram and ROI dissimilarity is defined as a histogram dissimilarity measure between the attenuation histograms. In this setting, the full histograms are utilized according to the chosen histogram dissimilarity measure.

We apply this idea to classification of different emphysema patterns as well as normal, healthy tissue. Two dissimilarity representation approaches as well as different histogram dissimilarity measures are considered. The approaches are evaluated on a set of 168 CT ROIs using normal density based classifiers all showing good performance. Compared to using histogram dissimilarity directly as distance in a k nearest neighbor classifier, which achieves a classification accuracy of 92.9%, the best dissimilarity representation based classifier is significantly better with a classification accuracy of 97.0% ($p = 0.046$).

Keywords: classifier design, lung, COPD, emphysema, dissimilarity representation, earth movers distance

1. INTRODUCTION

The traditional approach to statistical pattern recognition is feature representation. More specifically, objects are represented by a number of features in a feature vector space, and classifiers are built in this representation.[1] This is also the general trend in lung parenchyma classification.[2–6] Duin *et al.* motivated the idea of basing classification directly on distances between objects, thereby completely avoiding features.[7] Instead of focusing on finding good features for describing objects, the focus is moved to finding good dissimilarity measures for comparing objects. Dissimilarity representations may be preferable to the traditional feature representation approach, e.g., when there is not enough expert knowledge available to define proper features or when data is high dimensional.[8]

Working in a dissimilarity representation of objects, a k nearest neighbor (kNN) classifier,[9] which is applied directly on distances between objects, is a natural and simple choice. However, there exist techniques that make it possible to use other classifiers such as normal density based classifiers on dissimilarity data.[8] The general idea is to represent data by a distance matrix containing pair-wise dissimilarities between objects, also called dissimilarity representation. From this representation, a feature space is derived in which traditional pattern recognition techniques then can be applied. Embedding of a Euclidean dissimilarity representation into a Euclidean space via classical scaling is one way of doing this.[10] A second approach is to treat the dissimilarity representation as a new data set with the rows being observations and the columns being dimensions in a dissimilarity space. Each

Further author information:
Lauge Sørensen: E-mail: lauges@diku.dk, Telephone: (+45) 35 32 07 39

dimension in this space measures the dissimilarity to a particular training prototype, and the set of prototypes is called the representation set.[8] A third approach that will not be considered further in this paper, is embedding in a pseudo-Euclidean space in the case of a non-Euclidean dissimilarity representation.[10,11]

Compared to a density based classifier built in a dissimilarity space, kNN has high computational complexity and large storage requirements. In kNN, distances to all training set objects need to be computed when classifying novel patterns, and therefore the entire training set needs to be stored. In a dissimilarity space, a few objects can be selected from the training set as prototypes in the representation set, keeping the dimensionality low and only requiring storage of the representation set and the trained classifier. A kNN classifier makes the classification decision based only on a local neighborhood, i.e., the k closest prototypes, which makes it sensitive to noise. Density based classifiers in a dissimilarity representation are more global, in the sense that parameters of Gaussian functions are estimated off-line using all available dissimilarity training data while still working in a low dimensional dissimilarity space or embedding, which has a natural smoothing effect. Also, the classification is based on a weighted combination of the dissimilarities between the novel pattern and the prototypes. These weights are estimated using the entire training set and thus "essential" prototypes are given more weight in the classification decision. A density based classifier is therefore expected to achieve better generalization when dealing with a small and noisy data set, especially in cases of normal distributed classes.

Previously, we investigated the use of feature histograms for lung disease pattern classification in computed tomography (CT) using a histogram dissimilarity measure directly as distance in a kNN classifier, which showed promising results.[12] In the literature, measures of histograms, such as moments of filter response histograms and measures on co-occurrence matrices, are often used as features in a feature space when classifying lung disease patterns in CT.[2–6] Using only the first few moments of a histogram might discard valuable information. Instead, using the full histogram for classification may improve classification accuracy.[13] This paper investigates the possible benefit of building classifiers in a histogram dissimilarity representation compared to using histogram dissimilarity directly as distance in a kNN classifier. In light of the previous discussions, we see several possible benefits of using a density based classifier trained in a histogram dissimilarity representation for lung parenchyma classification. To our knowledge, this has not been investigated before.

Pekalska *et al.* have applied dissimilarity representations in numerous standard data sets, including handwritten digits, polygons, road signs, and chromosome band profiles.[8,14] Dissimilarity representations have also been used in various other pattern recognition applications. Trosset *et al.* used dissimilarity representations for discriminating patients with Alzheimer's disease from normal elderly subjects in magnetic resonance images. The dissimilarities were based on hippocampal dissimilarity obtained from image registration deformations.[15] In this work, we represent images by histograms and construct dissimilarity representations based on histogram dissimilarities, which is an approach also taken by other authors. Bruno *et al.* used a dissimilarity representation based on symmetrized Kullback-Leibler divergence between RGB histograms for image retrieval.[16] Paclik *et al.* investigated the use of dissimilarity representations in hyperspectral data classification using various histogram dissimilarity measures.[17]

The specific application of this paper is classification of emphysema subtype and normal tissue in regions of interest (ROI), based on the CT attenuation histogram. Emphysema is a major component of chronic obstructive pulmonary disease (COPD) and is characterized by gradual loss of lung tissue. COPD is a growing health problem worldwide. In the United States alone, it is the fourth leading cause of morbidity and mortality, and it is estimated to become the fifth most burdening disease worldwide by 2020.[18] Methods for reliable classification of emphysema in lungs are therefore of interest, since they may form the basis for computer-aided diagnosis. CT imaging is gaining more and more attention as a diagnostic tool for COPD, and it is a sensitive method for diagnosing emphysema, assessing its severity, and determining its subtype. Both visual and quantitative CT assessment are closely correlated with the pathological extent of emphysema.[19] Emphysema is usually classified into three subtypes, or patterns, in CT,[20] and the two of the three subtypes we focus on in this paper are the following: centrilobular emphysema (CLE), defined as multiple small low-attenuation areas; and paraseptal emphysema (PSE), defined as multiple low-attenuation areas in a single layer along the pleura often surrounded by interlobular septa visible as thin white walls.

2. METHODS

This section describes the methodology that we use. Section 2.1 briefly describes how the attenuation histograms are computed from the ROIs. Section 2.2 describes three different histogram dissimilarity measures used for comparing histograms. Section 2.3 describes two dissimilarity representation approaches: the dissimilarity space approach and an embedding approach based on classical scaling. Both are based on a distance matrix that in turn is based on a histogram dissimilarity measure. Finally, Section 2.4 describes two classifiers, a linear discriminant and a quadratic discriminant classifier, that both will be trained and tested in the dissimilarity representations.

2.1 Histogram estimation

We represent each ROI by its attenuation histogram. The histogram is estimated using non-linear binning by choosing the histogram bins such that the total distribution of the attenuation values in the training set is approximately uniform.[13] All histograms are normalized to sum to one.

2.2 Histogram dissimilarity measures

Three histogram dissimilarity measures L are considered: one based on histogram intersection (HI),[21] earth movers distance (EMD),[22] and the L_2-norm. HI is given by

$$HI(H,K) = \sum_{i=1}^{N_b} \min(H_i, K_i)$$

where $H \in \mathbb{R}^{N_b}$ and $K \in \mathbb{R}^{N_b}$ are histograms each with N_b bins. $HI(\cdot,\cdot)$ is a similarity measure, and a dissimilarity measure based on this can be obtained by

$$L_{HI}(H,K) = 1 - HI(H,K). \tag{1}$$

All histograms considered in this work sum to one, thus $L_{HI}(\cdot,\cdot) \in [0,1]$. EMD is given by

$$L_{EMD}(H,K) = \sum_{i=1}^{N_b}\sum_{j=1}^{N_b} C_{ij}F_{ij} \tag{2}$$

where $C \in \mathbb{R}^{N_b \times N_b}$ is a ground distance matrix and $F \in \mathbb{R}^{N_b \times N_b}$ is a flow matrix. The flow matrix contains the optimal flows obtained by solving the transportation problem of moving the mass of H such that it matches the mass of K. The L_2-norm is given by

$$L_2(H,K) = \sqrt{\sum_{i=1}^{N_b}(H_i - K_i)^2}. \tag{3}$$

2.3 Dissimilarity representations

Computing all pairwise dissimilarities L between the objects from the set $\mathcal{A} = \{a_1, \ldots, a_n\}$ and the set $\mathcal{B} = \{b_1, \ldots, b_m\}$ we obtain the $n \times m$ dissimilarity, or distance, matrix[8,14]

$$D_L(\mathcal{A},\mathcal{B}) = \begin{pmatrix} L(a_1,b_1) & \ldots & L(a_1,b_m) \\ \vdots & \ddots & \vdots \\ L(a_n,b_1) & \ldots & L(a_n,b_m) \end{pmatrix}. \tag{4}$$

Using (4) with either (1), (2), or (3) as histogram dissimilarity, we obtain three different distance matrix representations of the data $D_{L_{HI}}(\mathcal{A},\mathcal{B})$, $D_{L_{EMD}}(\mathcal{A},\mathcal{B})$, and $D_{L_2}(\mathcal{A},\mathcal{B})$.

2.3.1 Dissimilarity space

One way to utilize the distance matrix (4) is by extracting a representation set of prototypes $\mathcal{R}$. Given a training set $\mathcal{T}$, this approach selects a set of objects $\mathcal{R} \subseteq \mathcal{T}$ from $\mathcal{T}$. All objects in $\mathcal{T}$ are represented in a dissimilarity space, where the i'th dimension corresponds to the dissimilarity with prototype $\mathcal{R}_i \in \mathcal{R}$, i.e., we compute $D_L(\mathcal{T}, \mathcal{R})$.[8] Selecting a representation set is conceptually similar to selecting a limited number of prototypes for the kNN classifier. However, where the prototypes define the kNN classifier independently of the remaining training set, $\mathcal{R}$ defines a dissimilarity space in which the entire training set is represented and used to train a classifier. The final classifier is therefore expected to be less sensitive to the specific choice of prototypes.

There are different ways of choosing the representation set, e.g., random selection or feature selection methods, in this context searching for prototypes. For simplicity, we will only consider random prototype selection in this work. Random selection has previously been found to give reasonable results.[10]

2.3.2 Embedding

Instead of selecting prototypes, another approach is to embed $D_L(\mathcal{T}, \mathcal{T})$ in a vector space and reduce the dimensionality of this space. Standard inner product based techniques can be applied in this space.

A $D_L(\mathcal{T}, \mathcal{T})$ based on an Euclidean dissimilarity measure L can be perfectly embedded in an Euclidean space by classical scaling, which is a distance preserving linear mapping.[10] It is based on the positive definite Gram matrix

$$G = -\frac{1}{2} J (D_L \odot D_L) J$$

where $\odot$ denotes entry-wise matrix multiplication and the centering matrix $J = I - \frac{1}{n}\mathbf{1}\mathbf{1}^T$ where n is the number of training set objects and $\mathbf{1} = [1, \ldots, 1]^T \in \mathbb{R}^n$. G is factorized using an eigendecomposition

$$G = Q \Lambda Q^T$$

where Λ is a diagonal matrix containing eigenvalues ordered by descending magnitude and Q is a matrix containing the corresponding eigenvectors. For $k \leq n$ non-zero eigenvalues, a k-dimensional Euclidean embedding is then obtained by

$$E = Q_k \Lambda_k^{\frac{1}{2}} \tag{5}$$

where $Q_k \in \mathbb{R}^{n \times k}$ contains the first k leading eigenvectors and $\Lambda_k \in \mathbb{R}^{k \times k}$ contains the square roots of the corresponding eigenvalues.

When $D_L(\mathcal{T}, \mathcal{T})$ is based on a non-Euclidean dissimilarity measure, B is not positive definite and therefore has negative eigenvalues. In this case, an Euclidean embedding cannot be obtained using (5) since the computations rely on square roots of the eigenvalues. This problem can be addressed by considering only positive eigenvalues and corresponding eigenvectors in (5).[10]

Two of the histogram dissimilarity measures used in this work, (1) and (2), are non-Euclidean and one, (3), is Euclidean. When using Euclidean distance, i.e., (3), classical scaling recovers the original $n \times N_b$ data matrix from the $n \times n$ distance matrix up to location, reflection, and rotation.

2.4 Classifiers

Two classifiers are evaluated in the different dissimilarity representations: a linear discriminant classifier (LDC) and a quadratic discriminant classifier (QDC).[1,9] These classifiers have previously shown to perform well in dissimilarity spaces.[14] Both are density based classifiers using multivariate Gaussian functions to represent classes $\omega_i = \{\mu_i, \Sigma_i\}$

$$G_i(\mathbf{x}; \mu_i, \Sigma_i) = \frac{1}{(2\pi)^{N/2} |\Sigma_i|^{1/2}} \exp\left(-\frac{1}{2} (\mathbf{x} - \mu_i)^T \Sigma_i^{-1} (\mathbf{x} - \mu_i) \right)$$

where N is the dimensionality of the input space and $\mathbf{x} \in \mathbb{R}^N$ is a position in the input space. In LDC, equal class covariance matrices Σ are assumed resulting in the following linear discriminant function

$$g_i(\mathbf{x}) = \mathbf{x}^T \Sigma^{-1} \mu_i - \frac{1}{2} \mu_i^T \Sigma^{-1} \mu_i + \log P(\omega_i) \tag{6}$$

where Σ and the class sample means μ_i are estimated in the dissimilarity representation obtained from $D_L(\mathcal{T}, \mathcal{T})$ and $P(\omega_i)$ is the class prior. In QDC, each class covariance matrix Σ_i is estimated separately resulting in the following quadratic discriminant function

$$g_i(\mathbf{x}) = -\frac{1}{2} \log |\Sigma_i| - \frac{1}{2} (\mathbf{x} - \mu_i)^T \Sigma_i^{-1} (\mathbf{x} - \mu_i) + \log P(\omega_i). \tag{7}$$

The density based classifiers assigns class ω_i to observation $\mathbf{x}$ according to the maximum discriminant function

$$\hat{g}(\mathbf{x}) = \arg\max_i g_i(\mathbf{x}). \tag{8}$$

3. EXPERIMENTS AND RESULTS

The data used for the experiments originates from a set of thin-slice CT images of the thorax. CT was performed using GE equipment (LightSpeed QX/i; GE Medical Systems, Milwaukee, WI, USA) with four detector rows, using the following parameters: In-plane resolution 0.78×0.78 mm, 1.25 mm slice thickness, tube voltage 140 kV, and tube current 200 milliampere (mA). The slices were reconstructed using a high spatial resolution (bone) algorithm. A population of 25 patients, 8 healthy non-smokers, 4 smokers without COPD, and 13 smokers diagnosed with moderate or severe COPD according to lung function tests[18] were scanned in the upper, middle, and lower lung, resulting in a total of 75 CT slices.

Visual assessment of the leading pattern, either NT, CLE, or PSE, in each of the 75 slices was done individually by an experienced chest radiologist and a CT experienced pulmonologist. 168 non-overlapping ROIs of size 31×31 pixels were annotated in the slices, representing the three classes: NT (59 observations), CLE (50 observations), and PSE (59 observations). The NT ROIs were annotated in the non-smokers and the CLE and PSE ROIs were annotated in the smokers, within the area(s) of the leading pattern.

Figure 1 shows an ROI from each of the three classes, together with the CT slices in which they were annotated, and Figure 2 shows the attenuation histograms of all 168 ROIs estimated using the non-linear binning principle described in Section 2.1.

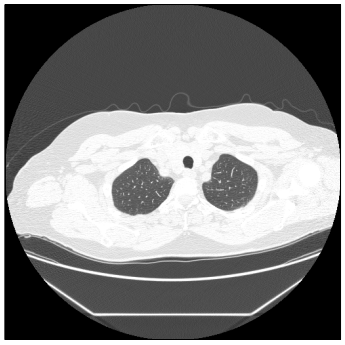
3.1 Visualizing dissimilarity spaces

Three prototypes are selected at random, one from each class, and the resulting pair-wise dissimilarity spaces are inspected by plotting the dissimilarities between all observations and one prototype versus the dissimilarities between all observations and second prototype. The results can be seen in Figure 3. The class separation is already quite good using only two prototypes and it can be expected to be even better when using more than two prototypes. In some cases, there is a tendency to degenerate behavior of the resulting spaces, e.g., in Figure 3(i) where the PSE samples almost reside on a line in the two-dimensional dissimilarity space.

3.2 Visualizing embeddings

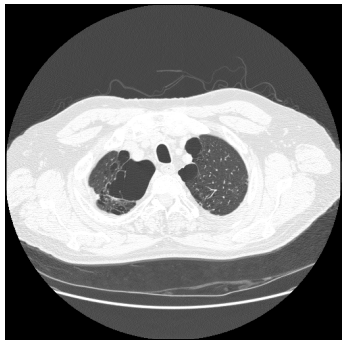
Figure 4 shows the eigenvalues derived in the embedding process for $D_{L_{HI}}$, $D_{L_{EMD}}$, and D_{L_2} on our data. As seen in Figure 4(a) and 4(b), the non-Euclidean property of L_{HI} and L_{EMD} is revealed by the presence of negative eigenvalues. The number of eigenvalues that are significantly different from zero is small in all three cases, showing that the intrinsic dimensionality of the three dissimilarity representations of the data is rather low.

Figure 5 shows two-dimensional embeddings obtained by using the two eigenvectors with the largest positive eigenvalues. The class separation is generally good in all three representations.

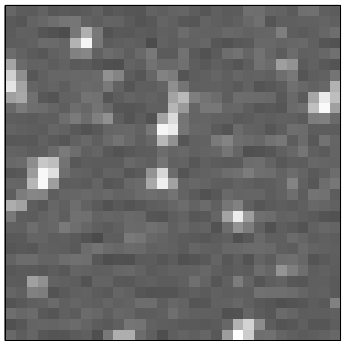


(a) CT slice with leading NT pattern.

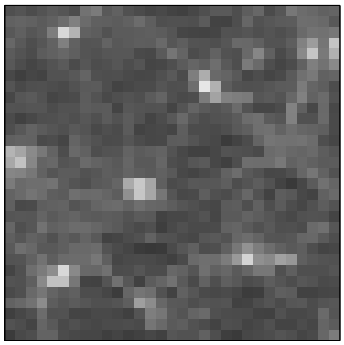
(b) CT slice with leading CLE pattern.



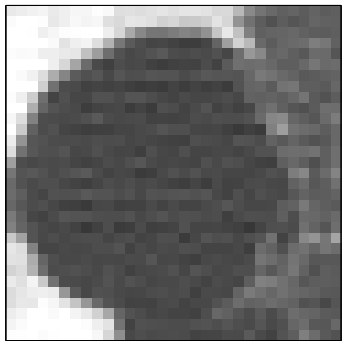
(c) CT slice with leading PSE pattern.



(d) NT ROI.

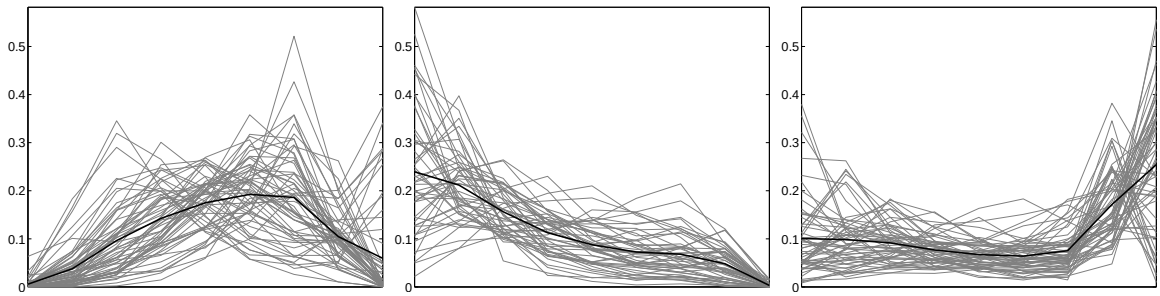


(e) CLE ROI.



(f) PSE ROI.

Figure 1. Examples slices and ROIs annotated in the same slices. The ROI in 1(d) is from 1(a) etc.

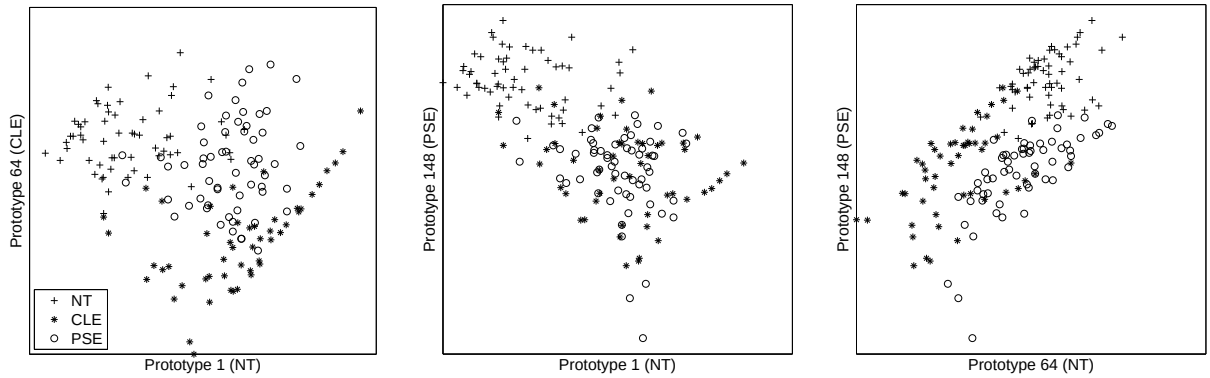


(a) NT.

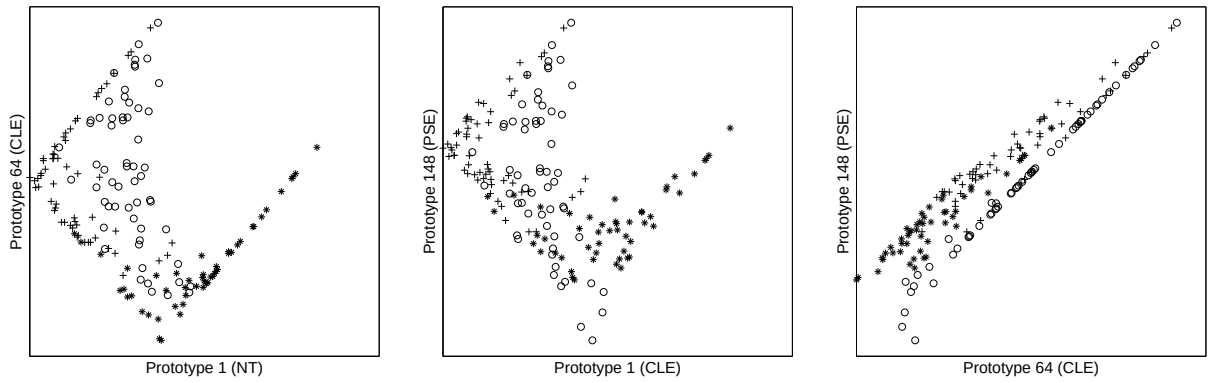
(b) CLE.

(c) PSE.

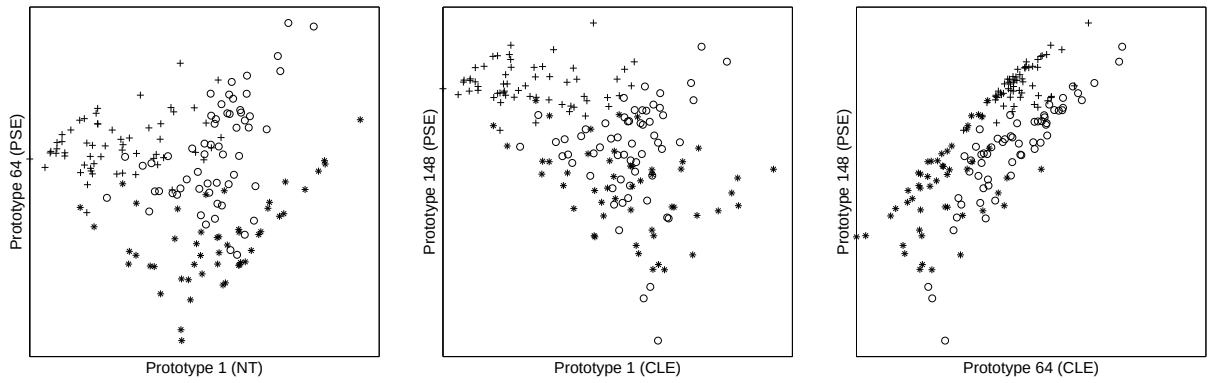
Figure 2. Attenuation histograms estimated from the data. Individual histograms are shown in gray and the mean histogram is shown in black.



(a) $D_{L_{HI}}$. (b) $D_{L_{HI}}$. (c) $D_{L_{HI}}$.

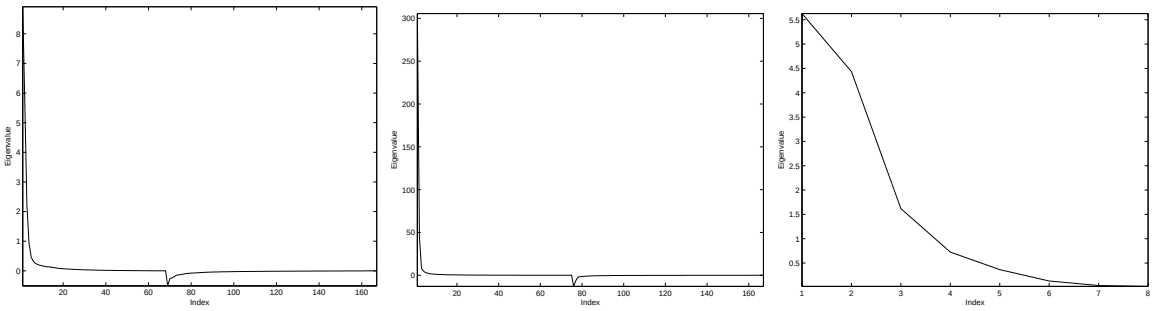


(d) $D_{L_{EMD}}$. (e) $D_{L_{EMD}}$. (f) $D_{L_{EMD}}$.



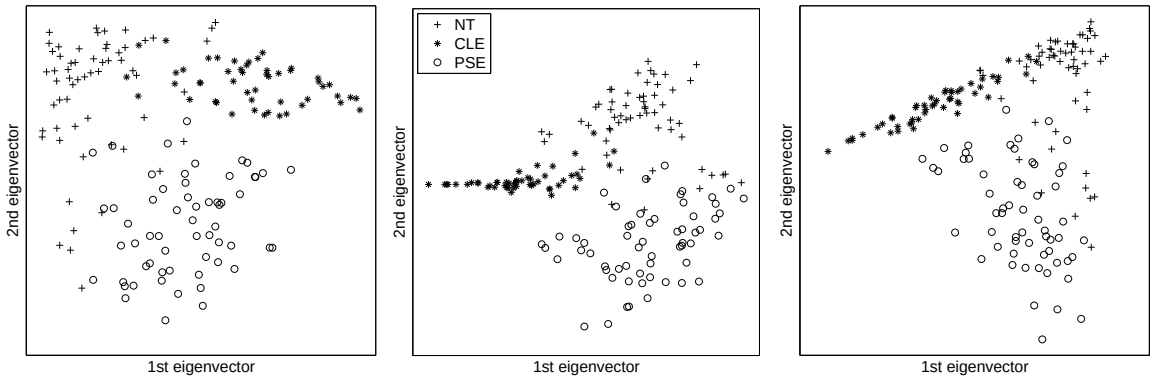
(g) D_{L_2}. (h) D_{L_2}. (i) D_{L_2}.

Figure 3. Examples of dissimilarity spaces obtained using representation sets with two random prototypes.



(a) $D_{L_{HI}}$. (b) $D_{L_{EMD}}$. (c) D_{L_2}.

Figure 4. Eigenvalues derived in the embedding process sorted by absolute value. In 4(a) and 4(b) the eigenvalues are divided in a positive and a negative part.



(a) $D_{L_{HI}}$. (b) $D_{L_{EMD}}$. (c) D_{L_2}.

Figure 5. Two-dimensional embedding of D_L using the two eigenvectors with the largest positive eigenvalue.

3.3 Classifier stability

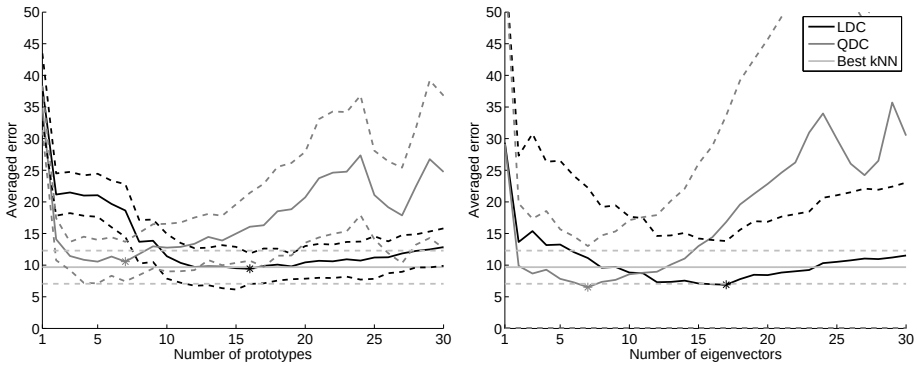
We use feature curves for inspecting the stability of the dissimilarity representation based classifiers as a function of the number of dimensions in the representation. That is, as a function of the number of prototypes in $\mathcal{R}$ and number of retained eigenvectors in E. The feature curves are computed based on thirty repeated random 50%/50% data splits. In these splits, balanced class distributions are ensured by placing half the ROIs representing one class in the training set and the other half in the test set. In each split, the dimension range $N = [1, 2, \ldots, 30]$ is used in turn by selecting N random prototypes in the dissimilarity space approach and N positive eigenvectors in the embedding approach, in both cases from the training set.

Figure 6 shows the resulting prototype curves. QDC is more sensitive to the number of dimensions compared to LDC. This phenomenon can be explained by the increasing number of parameters in QDC, which requires more training samples for reliable estimation.

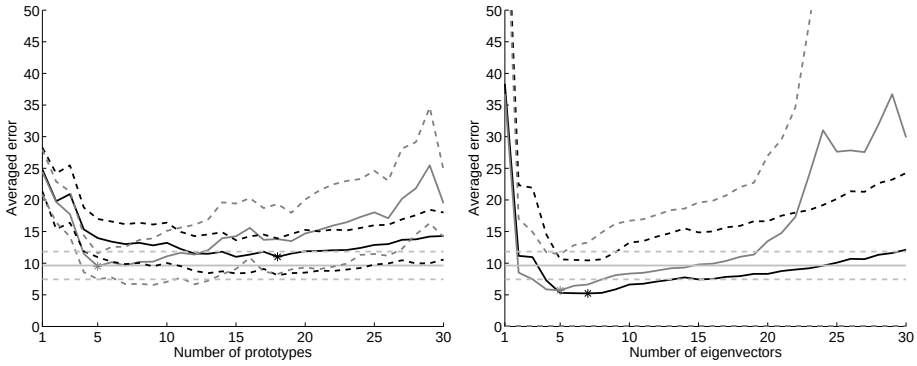
3.4 Classifier accuracy

The classification accuracy is evaluated using leave-one-out error estimation on the 168 ROIs, and the following classifier setups are evaluated:

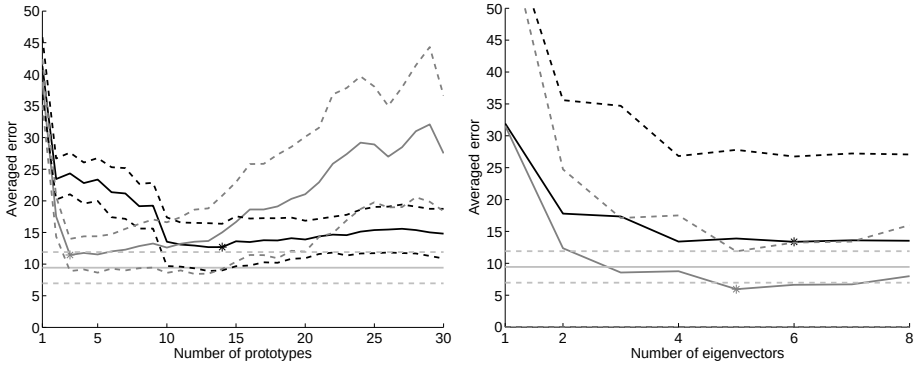
- kNN using histogram dissimilarity measure L as distance. $k = [1, 2, \ldots, 5]$, $L = \{L_{HI}, L_{EMD}, L_2\}$.
- Classifier C in a dissimilarity space defined by random representation set selection from distance matrix D_L. $C = \{\text{LDC}, \text{QDC}\}$, $D_L = \{D_{L_{HI}}, D_{L_{EMD}}, D_{L_2}\}$.
- Classifier C in an embedding of a distance matrix D_L. $C = \{\text{LDC}, \text{QDC}\}$, $D_L = \{D_{L_{HI}}, D_{L_{EMD}}, D_{L_2}\}$.



(a) Dissimilarity space using $D_{L_{HI}}$. (b) Embedding of $D_{L_{HI}}$.



(c) Dissimilarity space using $D_{L_{EMD}}$. (d) Embedding of $D_{L_{EMD}}$.



(e) Dissimilarity space using D_{L_2}. (f) Embedding of D_{L_2}.

Figure 6. Feature curves of dissimilarity representation based LDC and QDC. Standard deviations are shown as dashed lines. The asterisks mark the minimum of each curve. The performance of the best kNN classifier, for $k = [1, \ldots, 5]$, using the training set as prototypes and the histogram dissimilarity in question as distance is also shown for reference as a horizontal line.

Classifier		L_{HI}	L_{EMD}	L_2
kNN using L as distance	1NN	91.7	92.9	92.3
	2NN	91.1	91.1	91.7
	3NN	92.9	91.1	92.3
	4NN	92.3	90.5	91.7
	5NN	91.1	89.3	91.1
Dissimilarity space	LDC	88.6 ($\pm$1.0)	88.3 ($\pm$1.7)	87.6 ($\pm$1.3)
	QDC	93.1 ($\pm$1.1)	90.1 ($\pm$2.0)	93.3 ($\pm$1.2)
Embedding	LDC	91.1	97.0	86.3
	QDC	94.1	95.2	95.2

Table 1. Results of the leave-one-out evaluation. The reported performance of the dissimilarity space experiments is an average of ten repeated leave-one-out experiments where the representation set is drawn at random each time. The same random representation set is used for all tested configurations. The standard deviations of these experiments are shown in parenthesis.

The number of bins in the non-linear attenuation histogram is chosen as $N_b = \lfloor \sqrt[3]{N_p} \rfloor$, where N_p is the number of pixels in the ROI. In calculating L_{EMD}, the ground distance matrix, C in (2), is constructed such that the distance between two neighboring bins the attenuation histograms is one. More generally, the ground distance between bin i and bin j is $C_{ij} = |i - j|$. Further, we use the EMD implementation by Rubner.[23] The LDC and QDC class priors, $P(\omega_i)$ in (6) and (7), are estimated from data. The dimensionality of the dissimilarity spaces in all classifier setups is, somewhat arbitrarily, fixed to seven. All dissimilarity representation based classifiers perform reasonably well at this dimensionality according to the feature curves in Figure 6. The experiments are carried out in Matlab using the PRTools toolbox.[24]

In general, all the classifiers perform well, see Table 1, with classification accuracies in the range 88.3%−97.0%. Using the dissimilarity space approach with randomly chosen prototypes generally performs worse than using kNN with histogram dissimilarity as distance directly. However, the embedding approach shows very promising results, especially when L_{EMD} is used as histogram dissimilarity. The best estimated classification accuracy of 97.0% is achieved using LDC in the approximate embedding of $D_{L_{EMD}}$, and this is significantly better than the best kNN with histogram dissimilarity as distance according to a McNemar's test[25] ($p = 0.046$).

4. DISCUSSION AND CONCLUSIONS

The best dissimilarity representation based classifier achieves a classification accuracy of 97.0%, and this is significantly better ($p = 0.046$) than the best kNN classifier with histogram dissimilarity as distance, which achieved an accuracy of 92.9%. Generally, the embedding based classifiers perform slightly better than both the kNN and the dissimilarity space classifiers. Further, dissimilarity space based QDC, using only seven prototypes, performed similar to kNN. These results suggest that building classifiers in a dissimilarity representation, especially by embedding, is beneficial in the demonstrated application. The improved accuracy can be due to several factors. Firstly, a density based classifier built in a dissimilarity representation is more global, making use of all available training data in the classification decision, as opposed to a kNN classifier, which classifies only based on the k nearest prototypes. Second, in the embedding, the classes seem to be approximately normal distributed, see Figure 5, which fits well with normal density based classifiers like LCD and QDC.

Accuracies previously reported in the literature on lung parenchyma classification in CT including at least one type of emphysema, and using measures of feature histograms as features in a feature space, are generally lower and lie in the range $76\% - 93,5\%$.[2–6] These results are not directly comparable due to differences in the data, the choice of classes, etc. Nevertheless, the high accuracies of our experiments indicate that using the full feature histogram is beneficial and that a dissimilarity representation on histogram dissimilarities is a good way of utilizing the full feature histogram information.

In this work, we evaluated the dissimilarity space approach by drawing random prototypes for simplicity. However, prototype selection could be used instead, as in,[14] which could improve the performance of the representation set approach. Another possibility would be to draw the prototypes at random on class-level such that an equal amount of prototypes from each class are present in the representation set.

QDC, and to some degree also LDC, showed unstable behavior in high dimensional dissimilarity spaces and embeddings, as seen in the feature curves in Figure 6. This problem could be addressed by regularizing the estimated covariance matrices, allowing a larger number of dimensions to be used.[9] This could possibly improve the classification accuracy.

A natural next step would be to try dissimilarity representations based on other feature histograms than the attenuation histogram. For example, feature histograms describing local structure like local binary patterns[12] or other types of features previously used in lung parenchyma classification.[2–6] Combining the attenuation histogram and feature histograms describing local structure in a dissimilarity representation might improve performance.

In conclusion, we explore the use of normal density based classifiers built in a dissimilarity representation for lung parenchyma classification. Two different dissimilarity representation approaches are considered; embedding by classical scaling and the dissimilarity space approach, and dissimilarity representations based on different histogram dissimilarity measures are tried out. Two classifiers, LDC and QDC, are evaluated in the dissimilarity representations, and the best dissimilarity representation based classifier performed significantly better than using histogram dissimilarity directly as distance in a kNN classifier. A histogram dissimilarity representation allows for utilizing full feature histograms in classification, and through this representation, normal density based classifiers can be trained on histogram dissimilarity data. Further, sophisticated histogram dissimilarity measures, like the earth movers distance, fit naturally into this framework.

ACKNOWLEDGMENTS

This work is partly funded by the Danish Council for Strategic Research, under the Programme Commission for Nanoscience and Technology, Biotechnology and IT (NABIIT), by the Netherlands Organisation for Scientific Research (NWO), and by AstraZeneca, Lund, Sweden.

We would like to thank Saher B. Shaker (Hvidovre University Hospital, Denmark) and Asger Dirksen (Gentofte University Hospital, Denmark) for providing the data used in this work.

REFERENCES

[1] Duda, R. O., Hart, P. E., and Stork, D. G., [*Pattern Classification (2nd Edition)*], Wiley-Interscience (November 2000).

[2] Uppaluri, R., Hoffman, E. A., Sonka, M., Hartley, P. G., Hunninghake, G. W., and McLennan, G., "Computer recognition of regional lung disease patterns.," *Am J Respir Crit Care Med* **160**, 648–654 (Aug 1999).

[3] Chabat, F., Yang, G.-Z., and Hansell, D. M., "Obstructive lung diseases: texture classification for differentiation at CT," *Radiology* **228**, 871–877 (Sep 2003).

[4] Sluimer, I. C., van Waes, P. F., Viergever, M. A., and van Ginneken, B., "Computer-aided diagnosis in high resolution CT of the lungs," *Med Phys* **30**, 3081–3090 (Dec 2003).

[5] Xu, Y., Sonka, M., McLennan, G., Guo, J., and Hoffman, E. A., "MDCT-based 3-D texture classification of emphysema and early smoking related lung pathologies.," *IEEE Trans Med Imaging* **25**, 464–475 (Apr 2006).

[6] Park, Y. S., Seo, J. B., Kim, N., Chae, E. J., Oh, Y. M., Lee, S. D., Lee, Y., and Kang, S.-H., "Texture-based quantification of pulmonary emphysema on high-resolution computed tomography: Comparison with density-based quantification and correlation with pulmonary function test," *Investigative Radiology* **43**, 395–402 (June 2008).

[7] Duin, R., de Ridder, D., and Tax, D., "Featureless pattern classification," *Kybernetika* **34**(4), 399–404 (1998).

[8] Pekalska, E. and Duin, R. P. W., "Dissimilarity representations allow for building good classifiers," *Pattern Recognition Letters* **23**(8), 943–956 (2002).

[9] Jain, A., Duin, R., and Mao, J., "Statistical pattern recognition: a review," *IEEE Trans Pattern Anal Mach Intell* **22**, 4–37 (Jan. 2000).

[10] Pekalska, E., Paclík, P., and Duin, R. P. W., "A generalized kernel approach to dissimilarity-based classification," *Journal of Machine Learning Research* **2**, 175–211 (2001).

[11] Goldfarb, L., "A unified approach to pattern recognition," *Pattern Recognition* **17**(5), 575–582 (1984).

[12] Sørensen, L., Shaker, S. B., and de Bruijne, M., "Texture classification in lung CT using local binary patterns," *MICCAI* **11**(Pt. 1), 934–941 (2008).

[13] Ojala, T., Pietikäinen, M., and Harwood, D., "A comparative study of texture measures with classification based on featured distributions," *Pattern Recognition* **29**(1), 51–59 (1996).

[14] Pekalska, E., Duin, R. P. W., and Paclík, P., "Prototype selection for dissimilarity-based classifiers," *Pattern Recognition* **39**(2), 189–208 (2006).

[15] Trosset, M. W., Priebe, C. E., Park, Y., and Miller, M. I., "Semisupervised learning from dissimilarity data," *Comput. Stat. Data Anal.* **52**(10), 4643–4657 (2008).

[16] Bruno, E., Moënne-Loccoz, N., and Marchand-Maillet, S., "Asymmetric learning and dissimilarity spaces for content-based retrieval," in [*CIVR*], 330–339 (2006).

[17] Paclik, P. and Duin, R. P. W., "Dissimilarity-based classification of spectra: computational issues," *Real-Time Imaging* **9**(4), 237–244 (2003).

[18] Rabe, K. F., Hurd, S., Anzueto, A., Barnes, P. J., Buist, S. A., Calverley, P., Fukuchi, Y., Jenkins, C., Rodriguez-Roisin, R., van Weel, C., and Zielinski, J., "Global strategy for the diagnosis, management, and prevention of chronic obstructive pulmonary disease: GOLD executive summary.," *Am J Respir Crit Care Med* **176**, 532–555 (Sep 2007).

[19] Shaker, S. B., von Wachenfeldt, K. A., Larsson, S., Mile, I., Persdotter, S., Dahlbäck, M., Broberg, P., Stoel, B., Bach, K. S., Hestad, M., Fehniger, T. E., , and Dirksen, A., "Identification of patients with chronic obstructive pulmonary disease (COPD) by measurement of plasma biomarkers," *The Clinical Respiratory Journal* **2 (1)**, 17–25 (2008).

[20] Webb, W. R., Müller, N., and Naidich, D., [*High-Resolution CT of the Lung, Third Edition*], Lippincott Williams & Wilkins (2001).

[21] Swain, M. J. and Ballard, D. H., "Color indexing," *Int. J. Comput. Vision* **7**(1), 11–32 (1991).

[22] Rubner, Y., Tomasi, C., and Guibas, L. J., "The earth mover's distance as a metric for image retrieval," *Int. J. Comput. Vision* **40**(2), 99–121 (2000).

[23] Rubner, Y., "Code for the earth movers distance (EMD)." http://ai.stanford.edu/ rubner/emd/default.htm (May 1998).

[24] Duin, R., Juszczak, P., de Ridder, D., Paclík, P., Pekalska, E., and Tax, D., "PR-tools, a matlab toolbox for pattern recognition." http://www.prtools.org/ (April 2008). Version 4.1.3.

[25] Dietterich, T. G., "Approximate statistical test for comparing supervised classification learning algorithms," *Neural Computation* **10**(7), 1895–1923 (1998).