Locality-constrained Subcluster Representation Ensemble for Lung Image Classification

Yang Song[a,*], Weidong Cai[a], Heng Huang[b], Yun Zhou[c], Yue Wang[d], David Dagan Feng[a]

[a]*Biomedical and Multimedia Information Technology (BMIT) Research Group, School of IT, University of Sydney, NSW 2006, Australia*
[b]*Department of Computer Science and Engineering, University of Texas, Arlington, TX 76019, USA*
[c]*Russell H. Morgan Department of Radiology and Radiological Science, Johns Hopkins University School of Medicine, Baltimore, MD 21287, USA*
[d]*Bradley Department of Electrical and Computer Engineering, Virginia Polytechnic Institute and State University, Arlington, VA 22203, USA*

Abstract

In this paper, we propose a new Locality-constrained Subcluster Representation Ensemble (LSRE) model, to classify high-resolution computed tomography (HRCT) images of interstitial lung diseases (ILDs). Medical images normally exhibit large intra-class variation and inter-class ambiguity in the feature space. Modelling of feature space separation between different classes is thus problematic and this affects the classification performance. Our LSRE model tackles this issue in an ensemble classification construct. The image set is first partitioned into subclusters based on spectral clustering with approximation-based affinity matrix. Basis representations of the test image are then generated with sparse approximation from the subclusters. These basis representations are finally fused with approximation- and distribution-based weights to classify the test image. Our experimental results on a large HRCT database show good performance improvement over existing popular classifiers.

Keywords:
Medical image classification, Locality-constrained linear coding, Sparse representation, Clustering, Ensemble classification

*Corresponding author.
Email address: yson1723@uni.sydney.edu.au (Yang Song)

1. Introduction

The interstitial lung disease (ILD) refers to a group of more than 150 diseases affecting the lung parenchyma (Webb et al., 2008). Prolonged ILD may result in pulmonary fibrosis and affect breathing. To diagnose ILD with radiology, HRCT imaging is currently the preferred technique, which provides about 10 times more resolution than the conventional CT of the chest. This enables more detailed analysis of the pulmonary parenchymal abnormalities for a more confident diagnosis. Manual interpretation of HRCT imaging is however time-consuming and prone to inter-observer variability. In particular for ILD, the radiological patterns include various consolidation, linear or reticular opacities, small nodules, cystic airspaces, ground-glass opacities, and thickened interlobular septa (Ryu et al., 2002). The large variety of disease types and complexity of tissue patterns imply that manual analysis of ILD is challenging even for experienced radiologists (Webb et al., 2008). Computerized approaches, on the other hand, are normally considered capable of discovering image details that are difficult to perceive by human (Tourassi, 1999; Li et al., 2001), and are effective against inter-observer variability by providing a standardized solution.

Our aim of this study is to automatically classify HRCT image patches of five tissue classes: *normal*, *emphysema*, *ground glass*, *fibrosis*, and *micronodules* (examples shown in Figure 1). The latter four classes are prevalent characteristics of ILD and detecting them is important to identify the ILD types. The main challenge in accurate classification of ILD tissue patterns is the intra-class variation and inter-class ambiguity. Images of the same class can exhibit different visual patterns, while images of different classes can display similar visual features. The feature space will be complicated with scattering regions within the same class and overlapping areas between different classes, even with domain-customized feature design (Depeursinge et al., 2012b; Song et al., 2013). Such issues inevitably cause difficulties to the classifier. Our focus of this study is thus to design a new classifier to tackle the intra-class variation and inter-class ambiguity.

1.1. Related Work

1.1.1. Ensemble Classification

Classification based on an ensemble of classifiers has been quite popular in medical image analysis. The basic principle of ensemble classification is

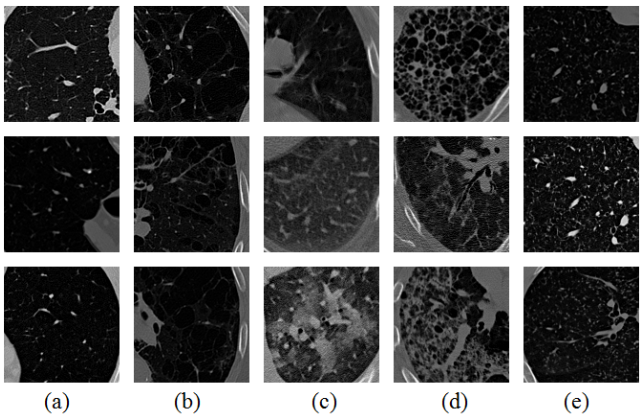

Figure 1: Example HRCT images (segments of axial slices) of the five ILD types: (a) normal, (b) emphysema, (c) ground glass, (d) fibrosis, and (e) micronodules. Note that each segment contains multiple image patches (details in Section 3).

that by integrating multiple base classifiers, better classification performance would be obtained than using the individual base classifiers (Rokach, 2010). Most of the existing ensemble classification methods can be categorized as the bagging (Lee et al., 2013; Chatelain et al., 2013; Criminisi et al., 2013; Yaqub et al., 2014; Allen et al., 2014; Zhao et al., 2014; Parrado-Hernandez et al., 2014) or boosting (Jacobs et al., 2011; Gorelick et al., 2013; Huang et al., 2014; Song et al., 2014c) models. A major difference between bagging and boosting is that the base classifiers in bagging are normally trained independently while in boosting the successive base classifiers are influenced by the prior ones. In both models, a single learning algorithm is used as the base classifier, including the support vector machine (SVM) (Parrado-Hernandez et al., 2014), decision or regression tree (Gorelick et al., 2013; Lee et al., 2013; Chatelain et al., 2013; Criminisi et al., 2013; Yaqub et al., 2014; Allen et al., 2014; Zhao et al., 2014), logistic regression (Jacobs et al., 2011), and sparse representation (Huang et al., 2014; Song et al., 2014c). Boosting algorithms can also be the building blocks in a tree structured classifier (Tu, 2005; Lu et al., 2012b; Feulner et al., 2013). The choice of base classifier is motivated by the particular imaging application, for which the base classifiers need to provide diverse but good classification performance to enable more accurate classification by the ensemble. The outputs from base classifiers are then fused via some variants of weighted averaging. The weighting schemes

are however usually predefined based on simple assumptions or the standard boosting algorithms, and might not adapt well to the specific data to be classified.

Generation of training subsets for the base classifiers is typically performed randomly or following a certain distribution. Random selection of feature subspace is also conducted with the random forest (Breiman, 2001) models (Lee et al., 2013; Chatelain et al., 2013; Criminisi et al., 2013; Yaqub et al., 2014; Allen et al., 2014; Zhao et al., 2014) to introduce further diversity in the base classifiers. The concept of representing multi-modal classes with multiple pseudo-classes and classifying them with separation hyperplanes between the pseudo-classes is proposed (Yu et al., 2010). Such a concept can be useful to classification problems with large intra-class variation. However, without a fusion component, the effectiveness of classification would be highly dependent on the individual separation hyperplanes. Recently an emerging trend of method is to sub-categorize / partition the training set of each class by feature clustering (Escalera et al., 2008; Dong et al., 2013; Song et al., 2014b). A base classifier is then trained for each subset, and the results are fused by weighted decoding (Escalera et al., 2008), kernel regression (Dong et al., 2013), or large margin aggregation (Song et al., 2014b). These sub-categorization methods partition the data by exploiting the characteristics of the feature space. Each subset clustered would have lower feature variation compared to the entire data set, hence helping to incorporate diversity and improve classification performance of the base classifiers. There are however few research in this area, and the design choices of the clustering algorithms, base classifiers, and fusion techniques, still remain under studied.

We can also consider the k-nearest neighbor (kNN) classifier as an ensemble model. With kNN, each data sample serves as a base classifier and majority voting from these base classifiers leads to the fused classification output. kNN is non-parametric and can naturally handle a large number of classes. Its effectiveness is however affected by the data distribution in the feature space. To improve the classification performance, discriminative learning has been incorporated. For example, in the SVM-KNN method (Zhang et al., 2006), SVM-based classification is applied in the localized feature space determined by the nearest neighbors of the test data. SVM can also be applied first to identify data samples that are close to the decision boundary, based on which cluster centers are derived and kNN is used to classify the data (Liu et al., 2011a). Learning-based distance metric is also a popular trend, such as the large margin nearest neighbor (LMNN) method

(Weinberger and Saul, 2009) that learns a Mahalanobis distance metric so that data from the same class would be more similar than those from different classes. Such a distance metric, however, is monolithic and could still be sensitive to the feature space complexity.

1.1.2. Sparse Representation

Sparse representation can be considered similar to kNN that a test image is classified based on similar images in the reference set. The major difference from kNN is that sparse representation identifies similar images by sparse approximation of the test image from linear combination of reference images, while kNN uses direct distance computation. Sparse representation does not require a parametric model to characterize the feature space separation (e.g. SVM) and can thus be particularly effective to handle the feature space complexity. Classification using sparse representation has recently been applied in many medical imaging applications (Liu et al., 2011b; Weiss et al., 2013; Xu et al., 2013; Tong et al., 2013; Song et al., 2013; Srinivas et al., 2014; Wang et al., 2014; Song et al., 2014a,c,b; Huang et al., 2014). Sparse representation has also been applied in multi-atlas models and the approximation coefficient is used as the weight vector to fuse the multiple atlases (Zhang et al., 2012; Song et al., 2012; Liao et al., 2013; Song et al., 2014a). The weights derived in this way would be adaptive to the test data and normally provide more desirable results than using predefined weighting schemes.

The effectiveness of sparse representation is largely affected by the quality of the reference data (Wright et al., 2010). In particular, if the reference set has large intra-class variation and inter-class ambiguity, a diverse set of reference images could be selected during the sparse approximation and this could result in better approximation from the wrong class than the correct class. One way to address this issue is to adapt the reference set to the test image. For example, the reference dictionary can be rescaled to increase the difference between the test image and the wrong class (Song et al., 2013). The locality-constrained linear coding (LLC) (Wang et al., 2010) has become widely popular in general computer vision. The essential idea is to encourage higher weights to be assigned to reference images that are more similar to the test image. The locality information in the feature space is thus exploited and the sparse approximation can be efficiently solved analytically. LLC has been directly applied for medical imaging and good performance has been demonstrated (Zhang et al., 2013; Xing and Yang, 2013; Wu et al., 2014). Another way is to partition the reference set into subsets and use sparse pre-

sentation as the base classifiers in a boosting (Huang et al., 2014; Song et al., 2014c) or sub-categorization (Song et al., 2014b) model. The subsets are expected to have lower intra-class variation and inter-class ambiguity, hence the sparse approximation at the subset-level could be more representative of the test image. The sub-categorization method (Song et al., 2014b) is expected to provide higher classification performance than the boosting-based methods (Huang et al., 2014; Song et al., 2014c), with its clustering-based reference partition and learning-based large margin fusion of base classifiers. However, the large margin fusion is relatively complex and could be sensitive to the selection of training set.

1.2. Our Contribution

In this work, we propose a *locality-constrained subcluster representation ensemble* (LSRE) model to classify HRCT image patches of five ILD tissue patterns. Our model comprises three stages. First, the images are partitioned hierarchically into subclusters based on spectral clustering with an approximation-based affinity matrix. Next, basis representations of the test image are obtained by sparse approximation with each subcluster as the reference dictionary. Finally, the basis representations are fused based on approximation- and distribution-based weights to classify the test image. Locality constraints are incorporated into the approximation objective for each of the three stages. Theoretically, the subclusters would capture the localized regions in the feature space, hence explicitly capturing the intra-class variation and inter-class ambiguity, and leading to diversity between the subclusters. Then with the basis representations as the base classifiers and data-adaptive design of the fusion weights, we expect that the final classification output would be more accurate than using sparse representation of the entire reference set.

Our LSRE model is closely related to ensemble classification, particularly the boosting techniques with sparse representation as the base classifiers (Huang et al., 2014; Song et al., 2014c), and the sub-categorization methods with clustering-based reference partition (Escalera et al., 2008; Dong et al., 2013; Song et al., 2014b). Different from these techniques, we designed a hierarchical spectral clustering algorithm with an approximation-based affinity matrix to partition the images. We also designed a data-adaptive weighting scheme to fuse the base classifiers with approximation- and distribution-based computations. In addition, we incorporated the locality constraints based on the LLC construct (Wang et al., 2010) to model local similarities between

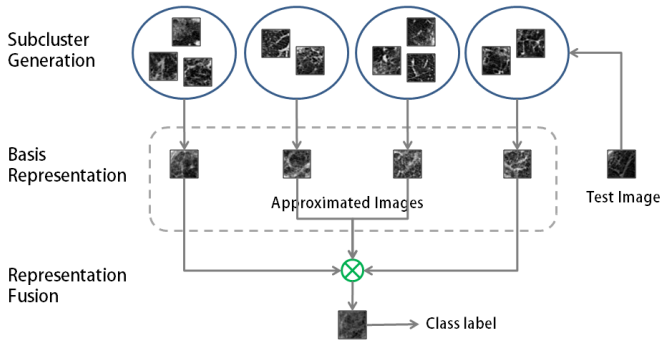

Figure 2: Overview of our LSRE model.

images and improve the computational efficiency. Our LSRE model is designed based on little domain knowledge about the ILD images, it can thus be generally applicable to other medical imaging applications.

The rest of the paper is organized as follows. A detailed description of our LSRE model is given in Section 2. The dataset and implementation details are described in Section 3. Our evaluation results and discussion are presented in Section 4. Finally Section 5 concludes the paper.

2. Methods

Given a set of N images (i.e. HRCT image patches), each with an H-dimensional feature vector $x_i \in \mathbb{R}^H$ and $X = \{x_i : i = 1, ..., N\}$, our aim is to determine the class label of each image in a multi-class classification setting. Figure 2 shows an overview of our LSRE model. First, the image set X is clustered hierarchically into subclusters, using spectral clustering with an affinity matrix derived in an LLC-based construct. A subcluster represents a localized region in the feature space, and could contain a mixture of images from different classes due to large inter-class ambiguity. Second, for a test image x, each subcluster (excluding x and any other images belonging to the same subject as x) is used as the reference dictionary to compute a basis representation of x based on LLC. A basis representation captures the approximation of x from a subcluster, and can be considered as a base classifier. Lastly, the basis representations are fused in an LLC-based model to obtain the weights of subclusters in approximating the test image x. Combined with the distribution-based weights estimating the reliabilities of basis

representations, the base classifiers at the subcluster-level are then fused to produce the class label of x.

2.1. Subcluster Generation

The first stage of our LSRE method is to divide the image set X into a union of K subclusters $\{S_k : k = 1, ..., K\}$ with minimum within-subcluster feature variation. Our design considerations for this stage are as follows. First, we expect that the images assigned to one subcluster are highly similar but they need not belong to the same class. While similar images should ideally be of the same class, this is usually not the case with large inter-class ambiguity. We thus do not require a subcluster to represent a single class and our LSRE method is designed to accommodate this heterogeneity in subclusters. Second, we expect to generate many subclusters with K determined at runtime. Due to large intra-class variation, the feature space of one class would be scattered into multiple local clusters. We would like to capture these localized regions (i.e. subclusters) so that the collection of subclusters would better describe the global feature space. This is different from the usual clustering approaches with one class represented by one cluster.

To this end, we design a hierarchical spectral clustering-based method with an affinity matrix derived using an LLC-based model. Assume an affinity matrix $A \in \mathbb{R}^{N \times N}$ is defined, with the (i, j)th element A_{ij} indicating the similarity between images x_i and x_j. With spectral clustering, X is divided into clusters by applying the normalized cuts algorithm on A (Shi and Malik, 2000). The problem is thus how to define the matrix A. An illustration of our subcluster generation method is shown in Figure 3.

2.1.1. Approximation-based Affinity Matrix

The affinity matrix A is essential in determining the clustering performance. Typically A is computed using the Gaussian similarity function based on a fully connected graph: $A_{ij} = \exp\{-\|x_i - x_j\|^2/(2\sigma^2)\}$, where σ sets the width of the neighborhood (von Luxburg, 2007). More advanced techniques based on sparse coding (SSC) (Elhamifar and Vidal, 2009), low-rank representation (LRR) (Liu et al., 2010), and least squares regression (LSR) (Lu et al., 2012a) have recently been reported in the area of subspace clustering. These techniques have a common hypothesis that images would be well approximated by the other images of the same cluster. They are different in the actual approximation algorithms used, and SSC produces a sparse matrix A while LRR and LSR output dense matrices. In our problem, we prefer A to

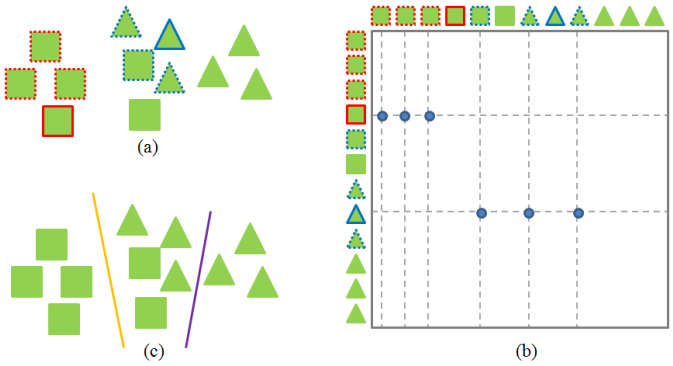

Figure 3: Illustration of subcluster generation. The example contains two classes, and the images belonging to these two classes are represented by squares and triangles, respectively. (a) gives a simplified view of the feature space of these images. An image of class 1 (rectangle with red solid outline) is nearest to three images of the same class (rectangle with red dashed outline); similarly an image of class 2 and its three nearest neighbors are shown. (b) visualizes the structure of the coefficient matrix Z. Each row vector contains the approximation coefficient with three non-zero elements, with which the image with solid outline is approximated by the three corresponding images. The coefficient matrix then converts to the affinity matrix A used in spectral clustering. (c) shows the 2-level hierarchical clustering output. At the first level, two clusters are created with the yellow line denoting the separation. At the second level, two more clusters are created indicated by the purple line. Altogether, three subclusters are generated during this process.

be sparse since we expect to generate many subclusters each containing only highly similar images, and SSC suits our aim in this aspect. However, SSC is time consuming due to the optimization routine of sparse coding. On the other hand, LLC as a sparse approximation algorithm is highly efficient with an analytical solution. We thus design a new approximation-based algorithm based on the LLC model to derive the affinity matrix A.

Specifically, we define the following approximation objective:

$$\begin{aligned} &\min_{\{z_i\}} \sum_{i=1}^{N} \|x_i - X z_i\|^2 + \lambda \|d_i \odot z_i\|^2 \\ &s.t.\ \ \mathbf{1}^T z_i = 1, \ \ \|z_i\|_0 \leq C_1, \ \ z_i(i) = 0, \ \ \forall i \end{aligned} \tag{1}$$

The first term penalizes the difference between the target image x_i and its approximation Xz_i. $X \in \mathbb{R}^{H \times N}$ denotes the concatenation of feature vectors of all images. The coefficient vector $z_i \in \mathbb{R}^N$ contains the weights of images in X in approximating x_i, and the total weight is 1. C_1-sparsity is expected

for z_i, i.e. $\|z_i\|_0 \leq C_1$, and C_1 is a constant. The target image x_i should have no contribution towards its own approximation and hence the corresponding weight $z_i(i) = 0$. The second term encourages to assign lower weights to images with larger distances from x_i. The symbol $\odot$ denotes the element-wise multiplication, and $d_i \in \mathbb{R}^N$ contains the pairwise Euclidean distances between the images in X and x_i. The parameter λ controls the balance between the two terms.

By minimizing this objective function, the non-zero elements in z_i indicate the degrees of similarity between the corresponding images and x_i. The second term is the essential idea of locality constraints in LLC, and z_i would contain only a small number of significant values, hence sparsity is encoded. This term also implies that the objective function can be efficiently solved with an approximative approach by replacing X with the top similar images. Similarly, we design the following approach to solve Eq. (1) efficiently.

We first use kNN with Euclidean distance to obtain the top C_1 similar images in X for the target image x_i. These C_1 images would not contain x_i itself to satisfy the constraint $z_i(i) = 0$. Denote the matrix concatenating the C_1 image feature vectors as $\tilde{X}_i \in \mathbb{R}^{H \times C_1}$. Eq. (1) is then reformulated as:

$$\begin{aligned} &\min_{\{\tilde{z}_i\}} \sum_{i=1}^{N} \|x_i - \tilde{X}_i \tilde{z}_i\|^2 + \lambda \|\tilde{d}_i \odot \tilde{z}_i\|^2 \\ &s.t. \;\; \mathbf{1}^T \tilde{z}_i = 1, \;\; \forall i \end{aligned} \tag{2}$$

where $\tilde{d}_i \in \mathbb{R}^{C_1}$ contains the pairwise Euclidean distances between $\tilde{X}_i$ and x_i, and $\tilde{z}_i \in \mathbb{R}^{C_1}$ represents the coefficient vector with a reduced dimension C_1 hence effectively satisfying $\|\tilde{z}_i\|_0 \leq C_1$. Next, by introducing the Lagrange multiplier α, we define the Lagrange function $\mathcal{L}(\tilde{z}_i, \alpha)$ as:

$$\|x_i - \tilde{X}_i \tilde{z}_i\|^2 + \lambda \|\tilde{d}_i \odot \tilde{z}_i\|^2 + \alpha(\mathbf{1}^T \tilde{z}_i - 1) \tag{3}$$

which is equivalent to:

$$\tilde{z}_i^T G \tilde{z}_i + \lambda \tilde{z}_i^T D \tilde{z}_i + \alpha(\mathbf{1}^T \tilde{z}_i - 1) \tag{4}$$

where $G = (\tilde{X}_i - x_i \mathbf{1}^T)^T (\tilde{X}_i - x_i \mathbf{1}^T)$, and $D = diag([\tilde{d}_1^2, ..., \tilde{d}_{C_1}^2])$. Then, let $\partial \mathcal{L}(\tilde{z}_i, \alpha) / \partial \tilde{z}_i = 0$, we have:

$$2(G + \lambda D)\tilde{z}_i + \alpha \mathbf{1} = 0 \tag{5}$$

The vector $\tilde{z}_i$ is derived analytically by:

$$\begin{aligned} \tilde{z}_i^* &= (G + \lambda D) \setminus \mathbf{1} \\ \tilde{z}_i &= \tilde{z}_i^* / (\mathbf{1}^T \tilde{z}_i^*) \end{aligned} \tag{6}$$

By projecting $\tilde{z}_i$ back to the original dimension of z_i, we finally obtain the C_1-sparse coefficient vector for each target image x_i. Note that in this approach, the selection of C_1 similar images is determined by the initial kNN step; but the degrees of similarity z_i are derived with the approximation objective and could be quite different from the distance-based measures.

By concatenating the coefficient vectors derived for all target image $\{x_i : i = 1, ..., N\}$, we obtain the coefficient matrix $Z \in \mathbb{R}^{N \times N}$ (illustrated in Figure 3b). The affinity matrix A is then computed as $(|Z| + |Z^T|)/2$, to have symmetric measure of similarities between image pairs. Spectral clustering is then performed on A to partition X into clusters.

2.1.2. Hierarchical Clustering

We apply the proposed spectral clustering approach in a hierarchical manner to generate K subclusters. At the first level, the affinity matrix A is computed for the entire image set X; and the number of clusters is the same as the number of classes. At the subsequent level $l > 1$, a cluster from the previous level, indexed by k_{l-1}, is sub-clustered by computing the affinity matrix for the images within this cluster. The number of sub-clusters is set to $\lfloor N_{k_{l-1}}/(\eta C_1/l) \rfloor$, where $N_{k_{l-1}}$ denotes the number of images in cluster k_{l-1} and η is a scaling constant. If the number of sub-clusters is not larger than 1, this cluster k_{l-1} is not sub-clustered. Assume we choose to have L levels of hierarchy, and a total of K subclusters $\{S_k : k = 1, ..., K\}$ are generated at the last level L (excluding those generated at the previous levels). These K subclusters are then the final outputs from this stage.

2.2. Basis Representation

With the subclusters generated, given a test image x, the second stage of our LSRE method is to obtain the basis representations of x based on the subclusters $\{S_k : k = 1, ..., K\}$. In other words, we would like to approximate x by using each subcluster S_k as a reference dictionary (illustrated in Figure 4a-c). A subcluster-based approximation, i.e. basis representation, can then be considered as a base classifier for x. While there are many sparse representation algorithms (Wright et al., 2010) well suited for this purpose, we adopt the LLC model mainly due to its efficiency.

Formally, we formulate the following objective function for basis representation from the subcluster S_k:

$$\begin{aligned} &\min_{p_{x,k}} \|x - \bar{S}_{x,k} p_{x,k}\|^2 + \lambda \|d_{x,k} \odot p_{x,k}\|^2 \\ &s.t. \;\; \mathbf{1}^T p_{x,k} = 1, \;\; \|p_{x,k}\|_0 \leq C_2 \end{aligned} \tag{7}$$

where $\bar{S}_{x,k} \in \mathbb{R}^{H \times N_{x,k}}$ is the reference dictionary, constructed by concatenating the feature vectors of images in S_k that are from different subjects as x, and $N_{x,k}$ denotes the number of images in $\bar{S}_{x,k}$. Images from the same subject as x are excluded to ensure complete separation between the test and reference data. Here $d_{x,k} \in \mathbb{R}^{N_{x,k}}$ contains the pairwise Euclidean distances between $\bar{S}_{x,k}$ and x to incorporate the locality constraints. The coefficient vector $p_{x,k} \in \mathbb{R}^{N_{x,k}}$ is the basis representation of x from the subcluster S_k. It is expected to be C_2-sparse derived from the top C_2 similar reference images, and is derived analytically in the same way as our solution for Eq. (1). The approximation output of x is $\bar{S}_{x,k} p_{x,k}$ from the subcluster S_k. If using the basis representation $p_{x,k}$ as a base classifier, x can be classified by finding the class of images assigned the highest total weight:

$$\underset{y}{\text{argmax}} \;\; p_{x,k}^T I_{x,k,y} \tag{8}$$

where $y \in \{1, ..., Y\}$ denotes the class label with Y as the number of classes. $I_{x,k,y} \in \mathbb{R}^{N_{x,k}}$ indicates the indices of reference images of class y, with value 1 in the corresponding elements and 0 elsewhere.

2.3. Representation Fusion

With the set of basis representations $\{p_{x,k} : k = 1, ..., K\}$, the third stage of our LSRE method is to fuse the basis representations to obtain the class label of the test image x. Our design motivation for this stage is as follows. Intuitively we consider weighted combination of the outputs from the base classifiers as the final classification result. The problem is then how to determine the weights. In our setting, we hypothesize that by approximating the test image x from the basis representations, the approximation coefficient can be used as the weights for fusion. We also expect sparsity in the approximation so that only the top related basis representations would contribute to the fusion. This is conceptually similar to a subtype of the multi-atlas algorithms that the multiple atlases are fused using various sparse representation techniques (Zhang et al., 2012; Liao et al., 2013; Song et al., 2014a). However,

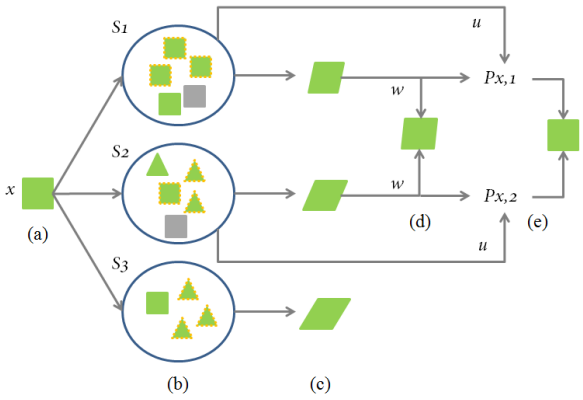

Figure 4: Illustration of basis representation and representation fusion. (a) indicates a test image x of class 1. (b) shows three subclusters. The first two subclusters each contains an image (gray square) from the same subject as the test image. They are thus removed from the reference dictionaries. The images with dashed outline are the most similar ones to the test image from each subcluster, and are used to generate the basis representation. (c) shows the approximation outputs. A larger distortion, i.e. larger difference between the approximated image and the test image, is expected with a more distant subcluster. (d) represents the approximation of x from the top two similar basis representations. (e) illustrates the computation of classification probabilities $P_{x,1}$ and $P_{x,2}$ by fusing the basis representations based on the approximation- and distribution-based fusion weights w and u. The final class label then corresponds to the class with the highest probability.

there seems no existing approach adapting the LLC model for fusion. Since LLC is highly efficient and naturally supports sparse approximation from top similar bases, we incorporate it into our design of the approximation-based weight vector. Furthermore, while some basis representations might produce very good approximation of x, the corresponding labels derived using Eq. (8) might actually be incorrect. We would thus like to estimate the reliabilities of the base classifier outputs from the individual basis representations. Our hypothesis is that subclusters containing a balanced mixture of image classes would be less discriminative than subclusters with images mainly a single class, and hence the corresponding base classifier output will be less reliable. We design a distribution-based weight vector to represent this reliability estimation. Illustration of this stage is shown in Figure 4d-e.

2.3.1. Approximation-based Fusion Weights

Given the basis representations $\{p_{x,k} : k = 1, ..., K\}$, we formulate the following objective function to approximate x:

$$\begin{aligned} &\min_{w_x} \|x - V_x w_x\|^2 + \lambda \|d_x \odot w_x\|^2 \\ &s.t. \;\; \mathbf{1}^T w_x = 1, \;\; \|w_x\|_0 \leq C_3 \end{aligned} \tag{9}$$

where the reference dictionary $V_x \in \mathbb{R}^{H \times K}$ is constructed by concatenating all approximations of x from the K basis representations: $V_x = \{\bar{S}_{x,k} p_{x,k} : k = 1, ..., K\}$. The pairwise Euclidean distances between the approximations V_x and x are contained in $d_x \in \mathbb{R}^K$, incorporating the locality constraints. The coefficient $w_x \in \mathbb{R}^K$ represents the weights of fusion, and is obtained analytically in the same way as our solution for Eq. (1).

In the objective function, the first term encourages a close approximation of x, and the second term penalizes higher weights to basis representations that are more distant from x. We also choose to impose a C_3-sparsity constraint on w_x so that only the top C_3 similar approximations of x would be involved. With this formulation, we expect that higher weights would be assigned to basis representations that are more related to x. In other words, assume x exhibits similar image features to the subcluster S_k. We would expect x to obtain a good approximation from S_k, i.e. a highly related basis representation $p_{x,k}$. Thereafter, this basis representation is also expected to have a high contribution towards approximating x during fusion, hence a large weight value in w_x.

2.3.2. Distribution-based Fusion Weights

We define a second weight vector, $u_{x,y}$, to estimate the reliabilities of the basis representations $\{p_{x,k} : k = 1, ..., K\}$ in classifying x as class y, based on the distribution of image classes in the subclusters. Consider that a subcluster S_k normally contains a mixture of images from different classes. The base classifier output of class y using $p_{x,k}$ would be quite reliable if the images in S_k (to be exact, $\bar{S}_{x,k}$) are mostly from class y. The reliability would degrade gradually as the number of images from the other classes increases. In addition, if the images in S_k (i.e. $\bar{S}_{x,k}$) are equally distributed among the Y classes, it would imply that S_k represents a localized region in the feature space that different classes are indistinguishable. The reliability of classification using $p_{x,k}$ would thus be lower than the case where the images belong to only a small number of classes.

With these considerations, we define the kth element $u_{x,k,y}$ of the weight vector $u_{x,y}$ as:

$$(\mathbf{1}^T I_{x,k,y}/N_{x,k}) \log(1 + \sigma\{\mathbf{1}^T I_{x,k,y}\}_{y=1}^{Y}) \tag{10}$$

where the first term indicates the percentage of class y images in $\bar{S}_{x,k}$, and $\log(\cdot)$ measures the standard deviation σ among the numbers of images belonging to the various Y classes. With a higher percentage and larger standard deviation in the distribution of image classes, the reliability of the basis representation $p_{x,k}$ to classify x as class y would be higher.

2.3.3. Weighted Classification

We finally compute the classification probabilities with a weighted fusion approach. Specifically, the classification probability of x belonging to class y is defined as:

$$P_{x,y} = \sum_{k=1}^{K} p_{x,k}^T I_{x,k,y} w_{x,k} u_{x,k,y} \tag{11}$$

where $p_{x,k}^T I_{x,k,y}$ is the classification probability derived with the basis representation $p_{x,k}$ as a base classifier (described in the previous section). Such K probabilities are then weighted combined with the approximation- and distribution-based weights $w_{x,k}$ and $u_{x,k,y}$. The test image x is then classified to the class with the highest probability: $\text{argmax}_y P_{x,y}$.

3. Dataset and Implementation

We used the ILD database (Depeursinge et al., 2012a) in this study. The database contains 113 HRCT images, and altogether 2062 2D regions-of-interest (ROIs) manually annotated with 17 ILD tissue class. The annotation was performed by two radiologists with 15 and 20 years of experience. Following the setup in (Depeursinge et al., 2012a,b; Song et al., 2013, 2014c), we selected the ROIs belonging to five major ILD tissue classes: normal (NM), emphysema (EM), ground glass (GG), fibrosis (FB) and micronodules (MN). We divided the axial slices into a grid of half-overlapping image patches with 31×31 pixels. The image patches with centroids inside the annotated ROIs were included in our experimentation. Our dataset thus comprised a total of 23131 image patches from 93 HRCT images / subjects, with 6438 NM, 1474 EM, 2974 GG, 4396 FB, and 7849 MN image patches. The numbers of ROIs belonging to the various tissue classes are 135, 54, 353, 386 and 265, annotated in 12, 5, 35, 35 and 16 images, respectively.

Each image patch was thus an "image" to be classified to one of the five ILD tissue classes ($Y = 5$). The texture-intensity-gradient (TIG) feature vector (Song et al., 2013) was used to describe each image. The feature vector is 176-dimensional ($H = 176$) containing three types of information: rotation-invariant local binary patterns based on Gabor-filtered images, intensity histogram, and histogram of oriented gradients with multiple coordinates. This feature vector was specifically designed for the ILD classification problem and showed good performance previously (Song et al., 2013, 2014c,b). To use a consistent test setup with our previous studies (Song et al., 2014c,b) for convenient performance comparison, we divided the dataset sequentially into four subsets of similar numbers of subjects and a leave-one-subject-out testing scheme was then performed for each subset. Note that due to the small number of subjects of the EM class, three of the five subjects were duplicated so that each subset contained two EM subjects.

In our experiments, we set the following parameters: the parameter balancing the approximation term and distance term $\lambda = 1e-2$, the sparsity constants $C_1 = C_2 = 5$ and $C_3 = 20$, the number of levels of subclustering $L = 6$, and the scaling factor to determine the number of clusters $\eta = 20$. During our empirical study, we experimented with various possible parameter settings on each subset. The set of values that provided good classification for all subsets was then used as the best parameters to evaluate our method performance. Our design choice mainly involved the ranges of possible parameter settings. In particular, the default value of λ used in existing LLC-based studies was $1e-4$; we thus experimented with $\lambda = 1e-1$ to $1e-4$. The sparsity constants C_1, C_2 and C_3 were related to sparse approximation of image features, and we found that with the standard sparse representation classifier, a sparsity constant of 10 provided the best performance; we thus experimented with values between 5 and 25. The parameters L and η were simply set according to C_1 and C_2, with $L = C_1+1$ and $\eta = 4C_2$, so that each subcluster at the final level would contain enough number (varying around $4C_2$) images for sparse approximation.

It is worth mentioning that the selection of the five tissue classes was motivated by the fact that they were the most common tissue patterns in ILDs (Depeursinge et al., 2012a) and the existing studies for this database focused on these five tissue classes. We adopted the same aim of study so that we could compare with the state-of-the-art directly. The remaining twelve tissue classes include consolidation (12 subjects), bronchial wall thickening (1 subject), reticulation (10 subjects), macronodules (5 subjects), cysts (1

Table 1: Confusion matrix of ILD classification.

Ground Truth	Prediction				
	NM	EM	GG	FB	MN
NM	**0.885**	0.045	0.010	0.007	0.054
EM	0.182	**0.796**	0.022	0.000	0.000
GG	0.069	0.000	**0.800**	0.068	0.064
FB	0.007	0.028	0.059	**0.854**	0.053
MN	0.034	0.000	0.046	0.048	**0.872**

subject), peripheral micronodules (5 subjects), bronchiectasis (5 subjects), air trapping (1 subject), early fibrosis (1 subject), increased attenuation (2 subjects), tuberculosis (1 subject), and pcp (2 subjects). It would be interesting to see how our method would extend to more classes, especially consolidation and reticulation.

4. Results and Discussion

4.1. Overall Performance

Table 1 shows the confusion matrix of the classification results. Most of the tissue classes obtained higher than 80% classification rates. 18.2% of EM images were misclassified as NM while few EM images were misclassified as the remaining three classes. This can be explained by the large inter-class ambiguity, i.e. visual similarity, between the EM and NM images, as shown in Figure 1. The NM images exhibited high similarity with both EM and MN images, hence the misclassification of NM images was mainly among the EM and MN images. Another observation is that GG, FB and MN images were rarely misclassified as EM. This could be explained by the small number of EM images compared to the other classes. The small number implies a lower probability of EM images selected for basis representation and a lower distribution-based weight of the EM class, and hence a lower probability of labeling the other classes as EM. Table 2 summarizes the classification recall, precision and F-score of each tissue class. Overall, the results show relatively balanced performance among the different tissue classes. Note that while the rates of misclassifying the other classes as EM were low, the precision of EM was low affected by the small number of EM images.

Table 2: Recall, precision and F-score of ILD classification.

	NM	EM	GG	FB	MN
Recall (%)	88.5	79.6	80.0	85.4	87.2
Precision (%)	89.1	70.0	79.0	85.2	89.3
F-score (%)	88.7	74.5	79.5	85.3	88.3

As described in Section 3, we set the balancing parameter $\lambda = 1e-2$, sparsity constants $C_1 = C_2 = 5$ and $C_3 = 20$, number of levels $L = 6$ and scaling factor $\eta = 20$ for subclustering. We found that the classification performance decreased gradually with smaller λ. The value of λ was especially important for representation fusion, and a small $\lambda = 1e-4$ caused 4.7% reduction in average F-score. The effects of λ were smaller for subcluster generation and basis representation, with reduction of 1.4% and 2.1%, respectively. Our method performance was quite insensitive to the sparsity constants C_1, C_2 and C_3, with a maximum of 1.1% decrease in average F-score when various values between 5 and 25 were tested. Changes in the number of levels L also resulted in minor performance differences with 0.8% decrease in average F-score when $L = 3$. When the subclusters were rather small with $\eta = 10$, the average F-score was reduced by 2.2%, indicating the need of having relatively large number of images in each subcluster.

The performance of our LSRE model was compared with the existing methods reported for ILD classification: (i) LF (Depeursinge et al., 2012b), which used localized features with the SVM classifier; (ii) PASA (Song et al., 2013), which was based on sparse representation with reference adaptation; (iii) BMSR (Song et al., 2014c), which was based on sparse representation in an AdaBoost construct; and (iv) LMLE (Song et al., 2014b), which was based on sparse representation in a sub-categorization model and was the state-of-the-art in patch-wise ILD tissue classification. The latter three approaches used the same TIG feature vector as in this study, hence the comparison with them demonstrated the effect of our LSRE model. The comparison with LF reflected the performance difference of the overall framework. We note that the results of LF were obtained directly from the paper (Depeursinge et al., 2012b), which used a slightly different selection of images from our dataset. For a fair comparison, PASA and BMSR were rerun to follow the same leave-one-subject-out test setup as this study; and the parameter settings followed

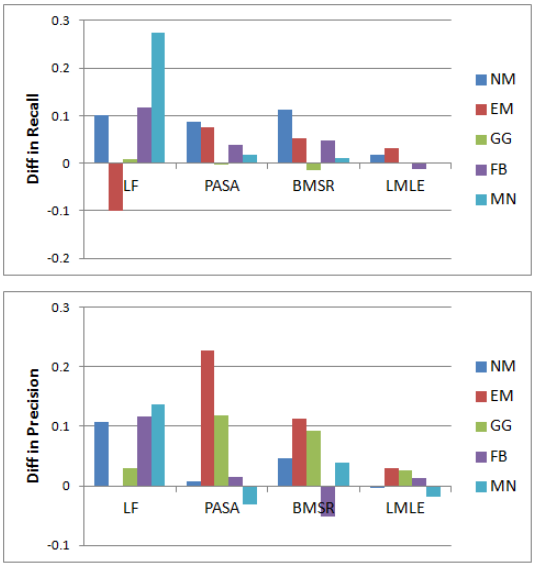

Figure 5: Differences in recall and precision between our LSRE model and the other methods reported for ILD classification.

those reported in these studies.

Figure 5 shows the differences in recall and precision between our LSRE model and the compared approaches, in which positive numbers indicate improvement of LSRE over the other approaches. The comparison shows that LSRE performed the best among the compared approaches. Compared to LMLE, although LSRE obtained slightly lower recall for the FB class, there were also fewer images misclassified as FB (including 11.1% of EM images). With a lower precision for MN, LSRE achieved more balanced results among the five tissue classes (6.1% standard deviation in F-score for LSRE vs. 7.8% for LMLE). LSRE also achieved more balanced results compared to BMSR and PASA, with 3.6% and 6.4% reductions in standard deviation of F-score. Recall that LMLE uses large margin learning for fusion of base classifiers. The advantage over LMLE suggests that while LSRE does not involve discriminative learning, it could actually obtain higher performance with the subcluster-based representation and fusion. The subclusters generated by LSRE can be used as base classifiers with sparse representation and a relatively uncomplicated fusion algorithm is required. On the other hand, LMLE is based on sub-categorization of individual classes separately; and the sub-

categories would not provide discriminative information between classes and the fusion component is thus particularly important for classification.

BMSR is also an ensemble classifier, however, its subclusters are created with random partition and the number of subclusters is much smaller (about 3% of that in LSRE); its base classifier is the L0 regularized sparse representation; and the fusion of base classifiers is based on a boosting algorithm. The improvement of LSRE over BMSR thus demonstrates the advantage of our overall method design. PASA is similar to the standard sparse representation algorithm, but involves adaptation of the reference data to the test images. The advantage over PASA implies that our ensemble model was more effective than using a global sparse representation classifier. Finally, our method is completely different from LF, which uses a different feature set and the SVM classifier. This comparison could however be biased since we used a different subset of the ILD database from LF. In particular, the two datasets contained different numbers of images but had very similar distributions of images among the five tissue classes. We would thus like to refer the readers to Figure 6 for comparison between SVM and LSRE, and show the comparison with LF as a general view of our method performance in the area of ILD tissue classification.

We also compared with the standard classifiers that are popular in medical imaging, including the kNN based on Euclidean distance, LMNN, SVM, SVM-KNN, sparse representation classifier (SRC) based on L0 regularization, the LLC model, and the random forest (RF) classifier. For all these approaches, we used the same TIG feature vector and leave-one-subject-out test setup as our LSRE model. The best performing parameters were set for each classifier: three nearest neighbors for kNN and LMNN, polynomial kernel for SVM and SVM-KNN (order of 3 and regularization parameter $C = 1.6$), 50 nearest neighbors for SVM-KNN, 10-sparsity for both SRC and LLC, the balancing parameter for LLC as $1e-2$, and 150 trees for RF.

As shown in Figure 6, LSRE achieved large improvement over the compared approaches. The kNN and LMNN classifiers are nearest neighbor-based methods. The kNN classifier did not work well since an image could appear similar to images of different classes, due to large inter-class ambiguity. While LMNN includes a learning-based distance metric, the learning algorithm was monolithic and affected by the large number of contradicting constraints. SRC and LLC are sparse representation models. SRC did not gain advantage over kNN, mainly due to the feature space complexity causing selection of reference images from the wrong class for sparse approximation.

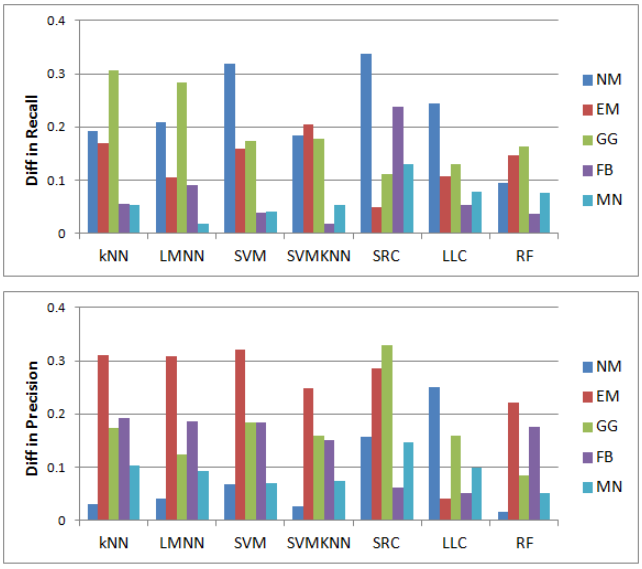

Figure 6: Differences in recall and precision between our LSRE model and the other popular classifiers.

LLC performed better than kNN and SRC, implying that incorporating the locality constraints into the approximation objective helped to accommodate the feature space complexity. The main difference between LSRE and LLC is that we used LLC at the subcluster-level to compute the basis representations and the LLC-based base classifiers are fused to obtain the final classification. The advantage of LSRE over LLC thus implies that our ensemble classifier was more effective than using a single LLC based on the entire reference set.

RF is a popular ensemble classifier based on tree bagging. Different from LSRE, RF creates the subclusters based on random sampling, uses decision tree as the base classifiers, and applies majority voting to obtain the fused classification output. RF also involves the additional random selection of feature subsets for further improvement. LSRE outperformed RF largely, and we suggest the essential cause was that the subclusters in our model were generated by clustering. The base classifiers built on these subclusters adapted to the local distributions in the feature space, and fusing these local classifiers could provide more accurate results than those created based on random subsets. Among the compared approaches, SVM is the most different classifier from our LSRE model. While SVM is typically a highly discrimi-

native classifier, with large intra-class variation and inter-class ambiguity, it could become over-fitted to the training data. The overfitting problem could be partially addressed by training local classifiers based on the nearest neighbors of the test data as in SVM-KNN, which obtained higher performance over the standard SVM. Classification with SVM-KNN however involved a single local classifier. The advantage of LSRE over SVM-KNN thus indicates the benefit of using an ensemble of base classifiers.

We further evaluated the statistical significance of performance improvement between our LSRE method and the compared approaches (excluding LF). A label vector containing 0 and 1 was computed from the classification outputs of each method, with 1 denoting correct classification and 0 otherwise. The label vector of LSRE was paired with that of each compared approach to compute the p-value using one-tailed paired t-test. The null hypothesis was that LSRE produced the same classification accuracy as the paired approach. We obtained p-value of 0.0155 when compared with LMLE, and $< 10^{-25}$ for comparisons with all the other approaches. The p-values were thus all less than 0.05 and indicated that our method achieved statistically significant improvement. Note that LF was not included in this evaluation, since we benchmarked with the reported results in Depeursinge et al. (2012b) directly and did not have the classification outputs of the images.

We would like to mention that the drawback of our LSRE method is the relative complexity. Our method contains three components: subcluster generation, basis representation, and representation fusion. Such an ensemble classifier design is notably more complicated than kNN and SVM, and the global sparse representation-based classifiers (SRC, LLC and PASA). On the other hand, BMSR, LMLE and RF are also ensemble classifiers, and they can be considered as containing three components of similar purposes but different algorithms. Compared to these ensemble approaches, our LSRE method involves a more complicated design of subcluster generation based on spectral clustering. However, LSRE does not require a training step to fuse the basis representations; and the training in LMLE can be quite slow for large datasets. At test time, LSRE is also faster than BMSR and LMLE mainly due to the incorporation of LLC, requiring about 0.06 second to classify an image vs. 0.93 and 0.4 seconds.

4.2. Evaluation of Subcluster Generation

Our subcluster generation method was compared with four other approaches for reference partition: (i) SSC (Elhamifar and Vidal, 2009), based

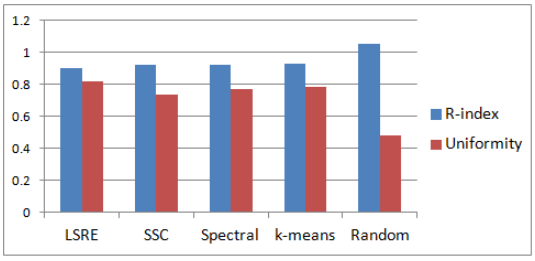

Figure 7: The R-index and uniformity values of the various clustering approaches.

on spectral clustering with the affinity matrix derived using L1-regularized sparse representation with C_1-sparsity; (ii) the standard spectral clustering; (iii) the k-means clustering; and (iv) random partition of disjoint clusters. All these approaches were conducted in the same hierarchical structure as in our subclustering method. Our method is most related to SSC since both involve customized affinity matrix with spectral clustering; and the main difference is our LLC-based approximation. The random partition is included since it is often used in ensemble classification.

First, we computed the R-index (Himberg et al., 2004) to measure the clustering qualities, which used the ratio between the within-cluster distances and the minimum between-cluster distances to quantify the compactness of and separation between clusters. A smaller value indicated a better clustering. As shown in Figure 7, our method obtained the lowest (best) R-index. The other three clustering methods obtained similar R indices. The advantage of LSRE over SSC was mainly attributed to the additional locality constraints incorporated into the sparse approximation. The random partition did not explore the feature space characteristics and resulted in the highest R-index.

The R-index, however, did not include the class label as a factor. While we expected that a subcluster could contain a mixture of different classes, it would help the classification performance if the subclusters were relatively uniform, meaning a subcluster contained images mostly from a single class. We thus computed another uniformity index, as the average ratio between the number of images of the majority class and the number of images in each subcluster. A larger uniformity value was better. As shown in Figure 7, our method obtained the highest uniformity. It was interesting that SSC provided the lowest uniformity while k-means obtained the highest, among

Table 3: Average F-score of ILD classification and execution time in seconds, of the various clustering approaches.

	LSRE	SSC	Spectral	k-means	Random
F-score	0.859	0.813	0.835	0.827	0.683
Time (s)	268	1900	223	9.8	0.4

the other three clustering algorithms. This could be accordant with our results in the previous section that SRC did not gain advantage over kNN. With large intra-class variation and inter-class ambiguity, the basic sparse representation would not be more effective than direct distance computation. Finally, as expected, the random partition resulted in the lowest uniformity.

Table 3 summarizes the resultant classification F-score averaged among the five tissue classes, and the execution time of subcluster generation. In all compared approaches, only the subcluster generation stage was replaced and the same processing as LSRE was applied for stages two and three. Our LSRE model obtained the highest F-score compared to the other four approaches. The benefit of having clustering-based reference partition over the random approach was evident. SSC was particularly slow due to the L1-regularized sparse approximation, and it was less effective when compared to the standard spectral and k-means clustering. Note that the subcluster generation process was required to run only once per subset of data (recall we divided the dataset into four subsets for faster testing) even in a leave-one-subject-out setting. This was because this stage was unsupervised without involving any class labels, hence all images could be included during clustering. Consequently, although our method was much slower than k-means, the impact on the overall classification efficiency was low.

4.3. Evaluation of Basis Representation

The basis representation stage was evaluated by comparing our method with kNN, SRC with L0 regularization (SRC-0) and L1 regularization (SRC-1). With kNN, equal weights were assigned to the top C_2 similar reference images from each subcluster as the basis representation. With SRC-0 and SRC-1, our LLC-based approximation method was replaced with the standard sparse approximation with C_2-sparsity. The subcluster generation and representation fusion stages were kept the same as our LSRE model.

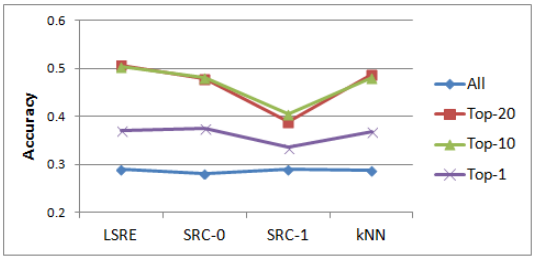

Figure 8: The average classification accuracies of various subsets of basis representations.

We first measured the average classification accuracy of the base classifiers based on various subsets of the subclusters. Specifically, four subsets were measured: *All* meaning the results from all subclusters, i.e. all basis representations, were averaged; *Top-20*, *Top-10* and *Top-1* meaning that the results from the subclusters producing the top 20, 10, or 1 best approximations were averaged. Note that the top approximations were selected individually for each test image based on the Euclidean distances between the approximation outputs and the test image.

As shown in Figure 8, Top-10 and Top-20 obtained similar accuracies, while the performance dropped with Top-1 and the lowest accuracies were obtained with All. This suggests that fusing the top 10 or 20 basis representations would provide better classification than using the top 1 or all of the basis representations. The classification accuracies of Top-10 and Top-20 were however quite low, implying that simply averaging the base classifiers would not achieve good classification. Another finding was that the accuracies of kNN and SRC-0 were close to our basis representation method while SRC-1 was much less accurate. This suggests that by fixing the number of neighbors in the approximation constraints (our method and SRC-0) and distance computation (kNN), the base classifiers were more effective than having varying number of nearest neighbors (SRC-1).

Table 4 lists the classification F-score averaged among all tissue classes, and the average time required to compute the basis representation for each test image. Our LSRE model shows clear advantage in classification performance compared to the other approaches. Different from the average accuracies of the basis representations (Figure 8), SRC-1 achieved the second highest F-score. This was attributed to the representation fusion stage, which combined the basis representations based on approximation- and distribution-

Table 4: Average F-score of ILD classification and execution time in seconds, of the various basis representation approaches.

	LSRE	SRC-0	SRC-1	kNN
F-score	0.859	0.764	0.786	0.775
Time (s)	0.05	0.13	0.82	0.05

based weights, rather than simply averaging them. For the same reason, our method achieved larger performance gain over the compared approaches than the average accuracies shown in Figure 8. In addition, the advantage of our LSRE model over SRC-0 and SRC-1 indicates the benefit of incorporating locality constraints into the sparse approximation for basis representation. Another advantage of our method was that the LLC-based formulation could be efficiently solved analytically. This was evident by the similar execution time between our method and kNN.

4.4. Evaluation of Representation Fusion

We compared our representation fusion method with the standard and highly related fusion techniques, including: (i) kNN, with which equal weights were assigned to the top C_3 basis representations for fusion; (ii) SRC-0, which computed the fusion weights based on sparse approximation with L0 regularization; and (iii) SRC-1, which used L1 regularization to generate the fusion weights. These compared approaches were used to replace our approximation-based fusion weights only and the distribution-based weights were still incorporated. In addition, we also evaluated the classification performance with approximation-based weights only (Approx), to analyze the effect of the distribution-based fusion weights.

For ensemble classification to achieve higher accuracy than using single classifiers, one of the main criteria is that the base classifiers need to provide diverse performance. In other words, it is more desirable if the base classifiers provided different classification probabilities rather than a uniform prediction, and different base classifiers would perform best for different subsets of the data. We thus measured the diversity using the following two metrics. First, we computed the entropy (Kuncheva and Whitaker, 2003) to quantify the level of differences in classification outputs between the base classifiers. Second, we checked the distribution of base classifiers that provided accurate

Table 5: Entropy and standard deviation measuring the diversity of base classifiers.

	LSRE	SRC-0	SRC-1	kNN
Entropy	0.098	0.050	0.085	0.098
Std	0.0069	0.0070	0.0078	0.0069

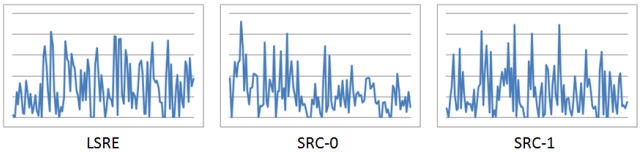

Figure 9: Distributions of selected base classifiers with accurate classifications, using LSRE, SRC-0, and SRC-1. The distribution of kNN is the same as (a). The x-axis represents the index of base classifiers / subclusters. The y-axis is the number of images accurately classified by the corresponding base classifier.

classification outputs. The evaluations were limited to the base classifiers that were selected for fusion. Note that our method and kNN would obtain the same entropy and distributions since our model selected the top C_3 base classifiers based on the kNN outputs.

Table 5 lists the entropy values of the various approaches. A larger value implied that the base classifiers were more independent and hence more diverse. The results show that our method produced the largest entropy while SRC-0 produced the lowest. In addition, as shown in Figure 9, the selection of base classifiers exhibited quite different distributions among the compared approaches. Table 5 lists the standard deviations of the distributions, with the cumulative distribution normalized to 1. A lower standard deviation means a more scattered distribution of base classifiers with accurate outputs, hence a higher diversity. Overall, the results show that our method provided the highest diversity among the compared approaches (same as kNN). This diversity property would then contribute to better classification performance achievable by our LSRE model compared to SRC-0 and SRC-1.

Table 6 shows the average F-score, and the average time required to derive representation fusion for each test image, with the various fusion techniques. The lower F-score using Approx indicates the benefit of including distribution-based fusion weights. The advantage of LSRE over kNN suggests

Table 6: Average F-score of ILD classification and execution time in seconds, of the various representation fusion approaches.

	LSRE	SRC-0	SRC-1	kNN	Approx
F-score	0.859	0.705	0.793	0.787	0.767
Time (s)	0.014	0.020	0.022	0.013	0.014

the effects of including the approximation-based weights. In addition, SRC-0 and SRC-1 obtained low F-scores. This is because with sparse representation, good approximation could be achieved by combining basis representations of different classes, and fusing them could often lead to misclassification. The improvement of LSRE over SRC-0 and SRC-1 suggests that by incorporating the locality constraints, our fusion method was more effective in identifying basis representations that were indeed representative of the test image. In addition, representation fusion was quite fast with any of these compared approaches, since the reference dictionary was small with K vectors (K being the number of subclusters).

4.5. Evaluation of ROI Classification

Our LSRE model was applied to classify image patches of 31×31 pixels, hence the previous evaluations were conducted at the image patch-level to demonstrate the effectiveness of LSRE. To further analyze the clinical relevance, we also evaluated the classification performance at the ROI-level. An ROI was classified by summing the classification probabilities, $P_{x,y}$ in Eq. (11), of the divided patches, and choosing the class with the highest probability. While we expect incorporating spatial relationships (Song et al., 2014a) or high-level feature descriptions (Lu et al., 2011, 2014) would help to improve the classification accuracy, in this study, we used the simple majority voting to focus our method design on the LSRE model.

Table 7 shows the classification results at the ROI-level. Note that although some measures are lower than the corresponding values in Table 2, it does not mean that more image patches were misclassified with the majority voting step. In fact, we found that with the majority voting, higher recall, precision and F-score were obtained for all five tissue classes at the image patch-level, with on average 5.6% improvement in F-score for each class. The ROIs were however of varying sizes, and the ROI-level performance measures

Table 7: Recall, precision and F-score of ROI classification.

	NM	EM	GG	FB	MN
Recall (%)	91.9	88.9	76.8	86.8	73.2
Precision (%)	74.3	87.3	77.9	85.2	84.3
F-score (%)	82.1	88.1	77.3	86.0	78.4

would be affected by small and misclassified ROIs.

We also compared with the state-of-the-art result of ROI-level ILD tissue classification (Asherov et al., 2014), which is based on bag of visual words feature representation and SVM classifier. The compared study was conducted on a different subset (91 subjects with 1018 ROIs) of the ILD database from ours (93 subjects with 1193 ROIs), and the two sets of ROIs were distributed differently among the five tissue types. Nevertheless, this comparison gave a general view of the effectiveness of our method. Our method achieved higher F-scores for all five tissue types with on average 2.6% improvement of each type. An exception was that lower precision was obtained for the EM type (-6.9%), mainly because there were fewer EM ROIs in our dataset.

We note that to assess if a method can be of real clinical interest, typically the classification results are analyzed based on intra- and inter-observer agreements. However, such statistics are not available for the ILD dataset. A similar study of ILD tissue classification (Sluimer et al., 2006) reported intra- and inter-observer agreements of 89% and 77% classification accuracy. Although a different dataset was used, we think that these statistics could be similarly applicable to our study given the similar problem domain. Our average classification accuracy at the ROI-level was 81.5%, which was thus comparable to the expert readings. In addition, when creating the ILD database, the radiologists spent on average one hour per case to identify ILD cases with high confidence (Depeursinge et al., 2012a). Although our method was applied to the annotated ROIs rather than the entire image, it required only about 16 seconds per case at runtime. Overall, considering the effectiveness and efficiency of our method, we suggest that our method can be useful as a second opinion to assist radiologists in decision making.

5. Conclusions and Future Work

We present a Locality-constrained Subcluster Representation Ensemble (LSRE) model to classify HRCT lung image patches of five ILD tissue classes. LSRE is a new ensemble classification model, with three main differences from the existing ensemble-based approaches. First, data subsets are generated using hierarchical spectral clustering with a sparse approximation-based affinity matrix. Second, the base classifiers are fused with data-adaptive approximation- and distribution-based weights. Third, the locality constraints are incorporated into each stage of our model to obtain effective sparse approximation and improve the model efficiency. Our ensemble-based design helps to tackle the difficulties in accurate classification caused by the intra-class variation and inter-class ambiguity in the feature space. We evaluated our method on a large ILD database, and demonstrated good performance improvement over the often used classifiers.

Our results show that the clustering-based subcluster generation is very important to the classification performance. We thus plan to investigate if further enhancing the clustering method, possibly by incorporating multi-way clustering (Ng et al., 2001) or the affinity propagation algorithm (Frey and Dueck, 2007; Zhan et al., 2009), will improve the classification performance in our future work. Our results also show that the LLC-based method is much more effective than the standard L0 and L1 regularized sparse approximation algorithms for basis representation and fusion. Our another future work is thus to investigate improving the LLC-based method, possibly with more advanced distance functions. Finally, besides customized parameter settings, our model does not involve any application-specific design and is generally applicable to multi-class classification. We will investigate applying the LSRE model to other medical image classification problems, such as the differentiation of various stages of dementia, in our future work.

6. Acknowledgement

This work was supported in part by Australian Research Council (ARC) grants. Heng Huang was partially supported by US NSF-IIS 1117965, NSF-IIS 1302675, NSF-IIS 1344152, and NSF-DBI 1356628.

References

Allen, D., Lu, L., Yao, J., Liu, J., Turkbey, E., Summers, R.M., 2014. Robust automated lymph node segmentation with random forests. SPIE Med. Imaging , 90343X.

Asherov, M., Diamant, I., Greenspan, H., 2014. Lung texture classification using bag of visual words. SPIE Med. Imaging , 90352K.

Breiman, L., 2001. Random forests. Machine Learning 45, 5–32.

Chatelain, P., Pauly, O., Peter, L., Ahmadi, S., Plate, A., Botzel, K., Navab, N., 2013. Learning from multiple experts with random forests: application to the segmentation of the midbrain in 3d ultrasound. MICCAI , 230–237.

Criminisi, A., Robertson, D., Konukoglu, E., Shotton, J., Pathak, S., White, S., Siddiqui, K., 2013. Regression forests for efficient anatomy detection and localization in computed tomography scans. Medical Image Analysis 17, 1293–1303.

Depeursinge, A., Vargas, A., Platon, A., Geissbuhler, A., Poletti, P.A., Muller, H., 2012a. Building a reference multimedia database for interstitial lung diseases. Comput. Med. Imaging Graph. 36, 227–238.

Depeursinge, A., de Ville, D.V., Platon, A., Geissbuhler, A., Poletti, P.A., Muller, H., 2012b. Near-affine-invariant texture learning for lung tissue analysis using isotropic wavelet frames. IEEE Trans. Inf. Technol. Biomed. 16, 665–675.

Dong, J., Xia, W., Chen, Q., Feng, J., Huang, Z., Yan, S., 2013. Subcategory-aware object classification. CVPR , 827–834.

Elhamifar, E., Vidal, R., 2009. Sparse subspace clustering. CVPR , 2790–2797.

Escalera, S., Tax, D.M.J., Pujol, O., Radeva, P., Duin, R.P.W., 2008. Subclass problem-dependent design for error-correcting output codes. IEEE Trans. Pattern Anal. Mach. Intell. 30, 1041–1054.

Feulner, J., Zhou, S.K., Hammon, M., Hornegger, J., Comaniciu, D., 2013. Lymph node detection and segmentation in chest CT data using discriminative learning and a spatial prior. Medical Image Analysis 17, 254–270.

Frey, B.J., Dueck, D., 2007. Clustering by passing messages bewtween data points. Science 315, 972–976.

Gorelick, L., Veksler, O., Gaed, M., Gomez, J.A., Moussa, M., Bauman, G., Fenster, A., Ward, A.D., 2013. Prostate histopathology: learning tissue component histograms for cancer detection and classification. IEEE Trans. Med. Imag. 32, 1804–1818.

Himberg, J., Hwarinen, A., Esposito, F., 2004. Validating the independent components of neuroimaging time series via clustering and visualization. Neuroimage 22, 1214–1222.

Huang, X., Dione, D.P., Compas, C.B., Papademetris, X., Lin, B.A., Bregasi, A., Sinusas, A.J., Staib, L.H., Duncan, J.S., 2014. Contour tracking in echocardiographic sequences via sparse representation and dictionary learning. Medical Image Analysis 18, 253–271.

Jacobs, C., Sanchez, C.I., Saur, S.C., Twellmann, T., de Jong, P.A., van Ginneken, B., 2011. Computer-aided detection of ground glass nodules in thoracic CT images using shape, intensity and context features. MICCAI , 207–214.

Kuncheva, L.I., Whitaker, C.J., 2003. Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. Machine Learning 51, 181–207.

Lee, G., Ali, S., Veltri, R., Epstein, J.I., Christudass, C., Madabhushi, A., 2013. Cell orientation entropy (core): predicting biochemical recurrence from prostate cancer tissue microarrays. MICCAI , 396–403.

Li, H., Wang, Y., Liu, K.J.R., Lo, S.B., Freedman, M.T., 2001. Computerized radiographic mass detection - part I: lesion site selection by morphological enhancement and contextual segmentation. IEEE Trans. Med. Imag. 20, 289–301.

Liao, S., Gao, Y., Lian, J., Shen, D., 2013. Sparse patch-based label propagation for accurate prostate localization in CT images. IEEE Trans. Med. Imag. 32, 419–434.

Liu, G., Lin, Z., Yu, Y., 2010. Robust subspace segmentation by low-rank representation. ICML , 663–670.

Liu, M., Lu, L., Ye, X., Yu, S., Huang, H., 2011a. Coarse-to-fine classification via parametric and nonparametric models for computer-aided diagnosis. CIKM , 2509–2512.

Liu, M., Lu, L., Ye, X., Yu, S., Salganicoff, M., 2011b. Sparse classification for computer aided diagnosis using learned dictionaries. MICCAI , 41–48.

Lu, C., Min, H., Zhao, Z., Zhu, L., Huang, D., Yan, S., 2012a. Robust and efficient subspace segmentation via least squares regression. ECCV , 347–360.

Lu, C., Zheng, Y., Birkbeck, N., Zhang, J., Kohlberger, T., Tietjen, C., Boettger, T., Duncan, J.S., Zhou, S.K., 2012b. Precise segmentation of multiple organs in CT volumes using learning-based approach and information theory. MICCAI , 462–469.

Lu, L., Bi, J., Wolf, M., Salganicoff, M., 2011. Effective 3D object detection and regressioni using probabilistic segmentation features in CT images. CVPR , 1049–1056.

Lu, L., Devarakota, P., Vikal, S., Wu, D., Zheng, Y., Wolf, M., 2014. Computer aided diagnosis using multilevel image features on large-scale evaluation. MCV , 161–174.

von Luxburg, U., 2007. A tutorial on spectral clustering. Stat. Comput. 17, 395–416.

Ng, A.Y., Jordan, M.I., Weiss, Y., 2001. On spectral clustering: analysis and an algorithm. NIPS , 849–856.

Parrado-Hernandez, E., Gomez-Verdejo, V., Martinez-Ramon, M., Shawe-Tylor, J., Alonso, P., Pujol, J., Menchon, J.M., Cardoner, N., Soriano-Mas, C., 2014. Discovering brain regions relevant to obsessive-compulsive disorder identification through bagging and transduction. Medical Image Analysis 18, 435–448.

Rokach, L., 2010. Ensemble-based classifiers. Artif. Intell. Rev. 33, 1–39.

Ryu, J.H., Olson, E.J., Midthun, D.E., Swensen, S.J., 2002. Diagnostic approach to the patient with diffuse lung disease. Mayo Clin. Proc. 77, 1221–1227.

Shi, J., Malik, J., 2000. Normalized cuts and image segmentation. IEEE Trans. Pattern Anal. Mach. Intell. 22, 888–905.

Sluimer, I.C., Prokop, M., Hartmann, I., van Ginneken, B., 2006. Automated classification of hyperlucency, fibrosis, ground glass, solid and focal lesions in high-resolution CT of the lung. Med. Phys. 33, 2610–2620.

Song, Y., Cai, W., Huang, H., Wang, X., Zhou, Y., Fulham, M., Feng, D., 2014a. Lesion detection and characterization with context driven approximation in thoracic FDG PET-CT images of NSCLC studies. IEEE Trans. Med. Imag. 33, 408–421.

Song, Y., Cai, W., Huang, H., Zhou, Y., Feng, D., Chen, M., 2014b. Large margin aggregation of local estimates for medical image classification. MICCAI , 196–203.

Song, Y., Cai, W., Huang, H., Zhou, Y., Wang, Y., Feng, D., 2014c. Boosted multifold sparse representation with application to ILD classification. ISBI , 1023–1026.

Song, Y., Cai, W., Zhou, Y., Feng, D., 2012. Thoracic abnormality detection with data adaptive structure estimation. MICCAI , 74–81.

Song, Y., Cai, W., Zhou, Y., Feng, D., 2013. Feature-based image patch approximation for lung tissue classification. IEEE Trans. Med. Imag. 32, 797–808.

Srinivas, U., Mousavi, H.S., Monga, V., Hattel, A., Jayarao, B., 2014. Simultaneous sparsity model for histopathological image representation and classification. IEEE Trans. Med. Imag. 33, 1163–1179.

Tong, T., R.Wolz, Coupe, P., Hajnal, J.V., Rueckert, D., ANDI, 2013. Segmentation of MR images via discriminative dictionary learning and sparse coding: application to hippocampus labeling. NeuroImage 76, 11–23.

Tourassi, G.D., 1999. Journey toward computer-aided diagnosis: role of image texture analysis. Radiology 213, 317–320.

Tu, Z., 2005. Probabilistic boosting-tree: learning discriminative models for classification, recognition, and clustering. ICCV , 1589–1596.

Wang, J., Yang, J., Yu, K., Lv, F., Huang, T., Gong, Y., 2010. Locality-constrained linear coding for image classification. CVPR , 3360–3367.

Wang, L., Shi, F., Gao, Y., Li, G., Gilmore, J.H., Lin, W., Shen, D., 2014. Integration of sparse multi-modality representation and anatomical constraint for isointense infant brain MR image segmentation. NeuroImage 89, 152–164.

Webb, W.R., Muller, N.L., Naidich, D.P., 2008. High-resolution CT of the lung. Lippincott Williams Wilkins.

Weinberger, K., Saul, L., 2009. Distance metric learning for large margin nearest neighbor classification. Journal of Machine Learning Research 10, 207–244.

Weiss, N., Rueckert, D., Rao, A., 2013. Multiple sclerosis lesion segmentation using dictionary learning and sparse coding. MICCAI , 735–742.

Wright, J., Ma, Y., Mairal, J., Sapiro, G., Huang, T.S., Yan, S., 2010. Sparse representation for computer vision and pattern recognition. Proc. IEEE 98, 1031–1044.

Wu, Y., Liu, G., Huang, M., Guo, J., Jiang, J., Yang, W., Chen, W., Feng, Q., 2014. Prostate segmentation based on variant scale patch and local independent projection. IEEE Trans. Med. Imag. 33, 1290–1303.

Xing, F., Yang, L., 2013. Robust selection-based sparse shape model for lung cancer image segmentation. MICCAI , 404–412.

Xu, Y., Gao, X., Lin, S., Wong, D.W.K., Liu, J., Xu, D., Cheng, C., Cheung, C.Y., Wong, T.Y., 2013. Automatic grading of nuclear cataracts from slit-lamp lens images using group sparsity regression. MICCAI , 468–475.

Yaqub, M., Javaid, M.K., Cooper, C., Noble, J.A., 2014. Investigation of the role of feature selection and weighted voting in random forests for 3-d volumetric segmentation. IEEE Trans. Med. Imag. 33, 258–271.

Yu, G., Feng, Y., Miller, D.J., Xuan, J., Hoffman, E.P., Clarke, R., Davidon, B., Shih, I., Wang, Y., 2010. Matched gene selection and committee classifier for molecular classification of heterogeneous disease. Journal of Machine Learning Research 11, 2141–2167.

Zhan, Y., Dewan, M., Zhou, X.S., 2009. Cross modality deformable segmentation using hierarchical clustering and learning. MICCAI , 1033–1041.

Zhang, H., Berg, A.C., Maire, M., Malik, J., 2006. SVM-KNN: discriminative nearest neighbor classification for visual category recognition. CVPR , 2126–2136.

Zhang, P., Wee, C., Nieghammer, M., Shen, D., Yap, P., 2013. Large deformation image classification using generalized locality-constrained linear coding. MICCAI , 292–299.

Zhang, S., Zhan, Y., Zhou, Y., Uzunbas, M., Metaxas, D.N., 2012. Shape prior modeling using sparse representation and online dictionary learning. MICCAI , 435–442.

Zhao, Q., Okada, K., Rosenbaum, K., Kehoe, L., Zand, D.J., Sze, R., Summar, M., Linguraru, M.G., 2014. Digital facial dysmorphology for genetic screening: hierarchical constrained local model using ICA. Medical Image Analysis 18, 699–710.