Stacked Autoencoders for Unsupervised Feature Learning and Multiple Organ Detection in a Pilot Study Using 4D Patient Data

Hoo-Chang Shin, *Student Member*, *IEEE*, Matthew R. Orton, David J. Collins, Simon J. Doran, and Martin O. Leach

Abstract—Medical image analysis remains a challenging application area for artificial intelligence. When applying machine learning, obtaining ground-truth labels for supervised learning is more difficult than in many more common applications of machine learning. This is especially so for datasets with abnormalities, as tissue types and the shapes of the organs in these datasets differ widely. However, organ detection in such an abnormal dataset may have many promising potential real-world applications, such as automatic diagnosis, automated radiotherapy planning, and medical image retrieval, where new multimodal medical images provide more information about the imaged tissues for diagnosis. Here, we test the application of deep learning methods to organ identification in magnetic resonance medical images, with visual and temporal hierarchical features learned to categorize object classes from an unlabeled multimodal DCE-MRI dataset so that only a weakly supervised training is required for a classifier. A probabilistic patch-based method was employed for multiple organ detection, with the features learned from the deep learning model. This shows the potential of the deep learning model for application to medical images, despite the difficulty of obtaining libraries of correctly labeled training datasets and despite the intrinsic abnormalities present in patient datasets.

Index Terms—Edge and feature detection, object recognition, pixel classification, machine learning, biomedical image processing

1 Introduction

Medical image analysis remains one of the less studied areas of computer vision. Unlike the frequently used scene images, for which the features are often well defined [1], [2], [3] and where the aim is to recognize an object in a 2D image, medical datasets and the objects contained within them are often 3D, with recognition performed on the component 2D slices. Moreover, while scene images are familiar to us and there are "enough" images with ground-truth provided [4], [5] for the training of machine learning algorithms, medical images are harder to obtain, and the ground-truth labels require substantially more specialist knowledge to define. By the same token, the time-consuming nature of the labeling task provides a strong impetus for the development of automated methods such as those described here. This is especially the case for patient data because of the abnormalities arising from disease. Both the shape and contrast properties of an organ with disease might look significantly different from the corresponding normal tissue. Furthermore, the majority of medical images—including all those containing the pathology that is the likely target and motivation for segmentation studies—are obtained from patients rather than healthy volunteers. This presents significant problems in making test datasets widely available, problems which are rooted both in the "data reuse" clauses of the ethical approvals under which a study has been conducted and in the nondisclosure arrangements imposed by the pharmaceutical companies that often sponsor the trials. Multimodal and so-called "functional" images can provide additional diagnostic information about the tissues being imaged to supplement the standard morphological images. However, the relatively recent introduction of such techniques, together with the cost of the extra imaging, mean that appropriately labeled functional datasets are rare and often are available only for small patient cohorts. Dynamic contrast-enhanced magnetic resonance imaging (DCE-MRI) [6] is a typical example: It has become an important tool for cancer diagnosis and assessment of therapeutic outcomes as it provides information on blood perfusion dynamics and vascular permeability of tissues, but it is uncommon to obtain DCE-MRI from a healthy subject because there are significant ethical restrictions on the use of contrast agents in nonpatients. A 4D DCE-MRI study is comprised of serial 3D datasets obtained during administration of a contrast agent.

We believe that "deep learning" can provide a promising approach to machine learning in patient datasets and might be a useful component of a diagnostic support platform. Unsupervised deep learning of hierarchical features fits well with the situation described above of medical image analysis using limited patient datasets and limited access to high-quality labeled data. It is our hypothesis that when hierarchical features are learned unsupervised, they represent characteristics of the object classes appearing in the

• *The authors are with the Institute of Cancer Research and Royal Marsden NHS Foundation Trust, Sutton, United Kingdom. E-mail: {hoo.shin, matthew.orton, david.collins, simon.doran, martin.leach}@icr.ac.uk.*

Manuscript received 1 Apr. 2012; revised 29 Sept. 2012; accepted 17 Dec. 2012; published online 28 Dec. 2012.
Recommended for acceptance by S. Bengio, L. Deng, H. Larochelle, H. Lee, and R. Salakhutdinovg.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number TPAMISI-2012-04-0240.
Digital Object Identifier no. 10.1109/TPAMI.2012.277.

dataset, and therefore only "rough guidance" is required from the human operator to train a classifier. Raina et al. [8] coined the term "self-taught learning" for this process, and our experiments show that features of the objects can be learned effectively from unlabeled data, with better representations being learned when the original data contain richer information from multiple modalities (in this case temporal), rather than simple visual data alone.

Following the approach of [8], we compare our new procedure to principal component analysis (PCA), which serves as a baseline method for unsupervised feature learning and, in addition, to a single-convolutional neural network (1-CNN) to demonstrate the effect of pretraining. We also compare our algorithm with two established feature-learning methods for image and time-series data: histogram of oriented gradients (HOG) [3] and a discrete Fourier transform (DFT) approach.

We show that a deep learning with a stacked sparse autoencoder model can be effectively used for unsupervised feature learning on a complex dataset for which it is difficult to obtain labeled samples. It makes minimal assumptions about the model describing the given data, and a similar model can be applied to new kinds of dataset with minimal redesign. Previous studies have shown that reducing the number of assumptions about the data and annotations can improve performance on action-classification tasks in multimodal data [17]. Furthermore, an "open" model (i.e., one for which the characteristics of the features learned can be controlled by its hyperparameters [9]) can be extended to "context-specific" feature learning. In our case, a typical context is the binary classification of an entity in the dataset as belonging to a particular organ class, and this type of learning is an approach in which *different* sets of features, each set being specific to a given context, can be learned by the *same* base-model. Finally, we demonstrate a probabilistic part-based method for object detection, which is used for localization of multiple organs, with the multimodal features learned from an unlabeled 4D DCE-MRI dataset.

The remainder of the paper is organized as follows: In Section 2, we review related works in the literature. Section 3 introduces the 4D patient data used in our study. Section 4 introduces the single-layer sparse autoencoder and reviews a preliminary study of our system applying the sparse autoencoder [10]. The concept of stacked sparse autoencoders, a deep architecture of the single-layer sparse autoencoder applied with max-pooling, is introduced in Section 5, together with analysis and comparison with other methods. Multi-organ detection with the stacked sparse autoencoders and probabilistic part-based object detection are covered in Section 6, followed by a discussion and conclusion.

2 Related Research

Our overall aim is to learn the object classes in a minimally labeled dataset: In other words, only a weakly supervised training is required to train a classifier. So-called "part-models" for the self-learning of object classes were studied for 2D images in [11], [12], [13], [14], and [15] to achieve object detection in such weakly supervised settings. In our work, a deep network model is used to learn features and part-based object class models in an unsupervised setting.

In [16], Ji et al. used 3D CNN to perform human-action recognition in video sequences. In this case, the CNNs were trained with labeled datasets and a large number of labeled examples were required. Furthermore, the action recognition was performed on a subwindow within a video sequence, which had to be preselected by a tracking algorithm, and the performance of the action-recognition was dependent on the tracking algorithm. By contrast, in [17] a generative model for learning latent information was applied for action recognition, and it did not require a tracking algorithm to recognize a human action, where the spatiotemporal features were learned from video sequences in an unsupervised manner. Based on the learned spatio-temporal features, "interest-points" were detected within a video sequence, and multiple actions could be recognized in a single video, based on those interest points. In a similar manner, we use a deep learning model to learn the latent information in a 4D medical image dataset.

Deep learning has attracted much interest recently, and has been used in a number of application areas. Many studies have shown how hierarchical structures in images can be learned using deep architectures with application to object recognition [18], [19], [20], [21], [22], [23]. Object recognition and tracking in videos with deep networks was shown in [24], where graphical model was used in addition to unsupervised feature learning by a restricted Boltzmann machines (RBM) [25]. Deep neural networks for classification of fMRI brain images were studied in [26], where RBMs were used to classify the stage and action of a volume while the images were taken.

Deep learning of multimodal features was recently studied in [27]. Our approach is similar, and we use stacked autoencoder model structure for separately learning both visual and temporal features. Independent subspace analysis, a deep neural network model for unsupervised multimodal feature learning, was suggested in [28], whereas in [29] and the many previous action-recognition studies appearing in [28], the objective was to recognize the action a video sequence represents. This also applies to [27], where the objective was to use multimodal feature learning to classify the whole video sequence as a single category. In our study, we aim to use unsupervised feature learning to recognize several objects within a given multimodal dataset.

Previous studies of automated object detection in medical images have tended to concentrate on brain images, especially detecting brain tumors. This is largely because both the shape and properties of the brain are more homogeneous across individuals than is the case for other parts of the body; for example, segmentation of MS lesions is reported in [30], [31], [32], and [33]. In all these cases, the disease tends to change the overall shape of the brain relatively little, whereas substantial shape changes can be observed with diseased abdominal organs. Moreover, a tumor is not an organ type but is a collection of abnormal tissues, which makes the approach to tumor segmentation different from object detection with a pattern recognition approach. Some of the complex tumor types represented by

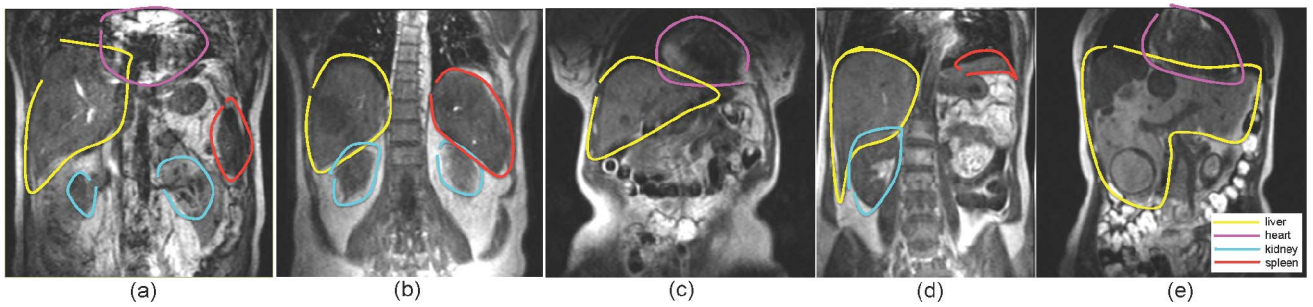

Fig. 1. The shapes of the organs vary substantially and the shape of a liver with metastases can be very abnormal (e). Regions were labeled as described in the main text. Note how the exact outline of the organs is not always clear. Uncertainty in identifying the spleen was high as it is difficult to distinguish from the other nearby organs, for example, in (d).

features learned with a sparse autoencoder in our dataset can be seen later in Section 4.1 and Fig. 4. In previous work [34], we suggested an approach for brain tumor segmentation in which a single-layer sparse autoencoder was used to learn the features present in the variation of image brightness in multiparametric MR images, followed by spatial clustering and logistic regression to segment edema and tumor. While this result indicates the potential of applying sparse autoencoders for medical image classification, the methods in [34] require additional elements to enable abdominal organ detection and classification to be performed at the same time.

The abdominal region contains many important organs and therefore has great potential to be useful for automated diagnosis and radiotherapy planning. Multi-organ detection was demonstrated in computed tomography (CT) images in [35], in contrast-enhanced abdominal CT images in [36], and in whole-body MR Dixon sequences in [37]. In all of these cases, a clearly labeled training dataset was required. Multi-organ segmentation on CT images using active learning with a minimal supervisory training set was demonstrated in [38], although in this study, a clinical expert's presence was required for the consecutive labeling during the active learning process. Also, the organs in the dataset in the studies are not largely abnormal as is the case in our data with tumors.

To our knowledge, there has not yet been an application of unsupervised feature learning with a deep learning approach to object recognition in medical images with large heterogeneous datasets. We demonstrate multi-organ detection in 4D DCE-MRI patient data, using the hierarchical multimodal features learned from an unlabeled subset of datasets with 78 patient scans.

The training, cross validation (CV), and test dataset are anonymized patient data from different studies of diseases, and our results show that the proposed method successfully learns features that lead to good classification performance in complex and variable datasets with low image resolution and noisy ground truth labels.

3 Dataset

Our 4D dataset consists of a time series of 3D DCE-MRI scans from two studies of liver metastases and one study of kidney metastases:

- *Dataset A.* Forty-six scans of patients with liver metastases, each containing 7-12 contiguous coronal slices with image size 256×256, repeated at $T = 40$ time points.
- *Dataset B.* Three scans of patients with kidney metastases, each containing 14 contiguous coronal slices with size 256×267, repeated at $T = 40$ time points
- *Dataset C.* Twenty-nine scans of patients with liver metastases from a clinical trial, each containing 14 contiguous coronal slices with image size 209×256, repeated at $T = 40$ time points.

These scans were acquired using a sequential breath-hold protocol [39], where two image volumes were obtained during consecutive 6 second breath-holds. Each breath-hold was followed by a 6 second interval where the patient was instructed to take a single breath. The images are obtained from different patients, and the slice locations (anterior-posterior positions are represented on the y-axis of the image volumes) are different between patients, which in consequence make the shapes of some organs in the images differ substantially. Moreover, the uncertainties in the locations and boundaries of the organs in coronal DCE-MRI images are large because many organs are located closely together, image resolution is relatively low (2.5 mm in-plane), and because of the pattern of contrast uptake in some organs. Furthermore, when an organ is in a late stage of a disease, its shape can be grossly abnormal. Examples are shown in Fig. 1. A contrast agent is injected into the patient during the DCE-MRI acquisition so that the contrast of the successive images changes according to the blood perfusion dynamics and vascular permeability of the tissues observed. DCE-MRI images of a liver patient scan and a time series of a liver voxels brightness change are shown in Fig. 2.

Subsets of dataset A were used for training, subsets of dataset B for CV, and dataset C was used for the final visualization and test, respectively. "Rough" outlines encompassing the labeled tissues, as shown in Fig. 1, were drawn by a nonexpert, and subsequently adjusted and confirmed by a radiologist. These outlines are used for supervised training in the training dataset and performance evaluation in the CV dataset.

4 Single-Layer Sparse Autoencoder

An autoencoder is a symmetrical neural network to learn the features of a dataset in an unsupervised manner. This is

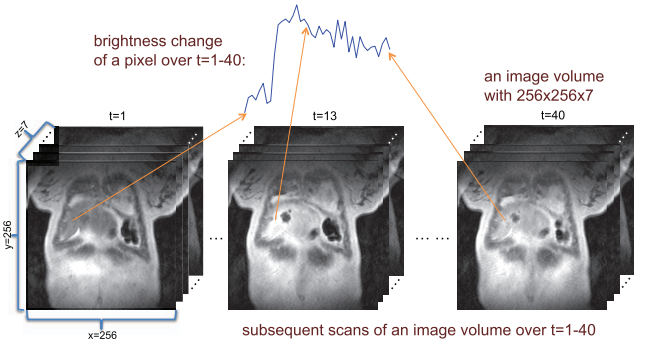

Fig. 2. A 4D DCE-MRI scan of a liver patient for a time course $1 \leq t \leq 40$ with volume size of $256 \times 256 \times 7$. Each pixel of an image slice in a volume gives a time series of its brightness over 40 images. The time series represents the perfusion status of the tissue in the voxel and will vary with tissue types.

done by minimizing the reconstruction error between the input data at the encoding layer and its reconstruction at the decoding layer so that the correlation between the input features is learned in an EM-like fashion [19], [40] in the mapping weight vectors.

Encoding of an input vector $\mathbf{x} \in \mathbb{R}^{D\times 1}$ is done by applying a linear mapping and a nonlinear activation function to the network:

$$\mathbf{a} = sigm(\mathbf{W}\mathbf{x} + \mathbf{b}_1), \tag{1}$$

where $\mathbf{W} \in \mathbb{R}^{N\times D}$ is a weight matrix with N features, $\mathbf{b}_1 \in \mathbb{R}^N$ is an encoding bias, and $sigm(x)$ is the logistic sigmoid function $(1 + \exp(-x))^{-1}$. Decoding of $\mathbf{a}$ is performed using a separate decoding matrix:

$$\mathbf{z} = sigm(\mathbf{V}^T\mathbf{a} + \mathbf{b}_2), \tag{2}$$

where $\mathbf{b}_2$ is a decoding bias and the decoding matrix is $\mathbf{V} \in \mathbb{R}^{N\times D}$. Features in the data are learned by minimizing the reconstruction error of the likelihood function $L(\mathbf{X},\mathbf{Z}) = \frac{1}{2}\sum_{i=1}^{m} \|\mathbf{z}_i - \mathbf{x}_i\|_2^2$, where $\mathbf{X}$ and $\mathbf{Z}$ are all the training and reconstructed data, respectively, and the features are encapsulated in $\mathbf{W}$.

While an autoencoder has a close relationship to PCA by usually performing a dimensionality reduction, an "overcomplete" (larger than the input dimension) nonlinear mapping of the input vector $\mathbf{x}$ can be made by applying sparsity to the target activation function, that is, a sparse autoencoder [41], [42], [43], [44], [45]. To achieve this, the objective in the sparse autoencoder learning is to minimize the reconstruction error with a sparsity constraint:

$$L(\mathbf{X}, \mathbf{Z}) + \beta \sum_{j=1}^{N} \mathrm{KL}(\rho \| \hat{\rho}_j), \tag{3}$$

where β is the weight of the sparsity penalty, N is the number of features in the weight matrix, ρ is the target average activation of $\mathbf{a}$, and $\hat{\rho}_j = \frac{1}{m}\sum_{i=1}^{m}[a_j]_i$ is the average activation of jth input vector a_j over the m training data. The Kullback-Leibler divergence [46] is given by

$$\mathrm{KL}(\rho \| \hat{\rho}_j) = \rho \log \frac{\rho}{\hat{\rho}_j} + (1-\rho) \log \frac{1-\rho}{1-\hat{\rho}_j}, \tag{4}$$

which provides the sparsity constraint—a nonredundant overcomplete feature set will be learned when ρ is small, as in sparse coding [47].

The model is trained by optimizing the objective function (see (3)) with respect to $\mathbf{W}$, $\mathbf{V}$, $\mathbf{b}_1$, and $\mathbf{b}_2$, where we used backpropagation [48] and L-BFGS [49] to train the model. It is generally accepted that classification performance is improved by increasing the number of learned features (N), and the effect of the number of features on classification performance using single-layer networks has been studied in more detail in [50].

Our DCE-MRI data have both temporal and spatial domains. Temporal features are learned from the organ-specific changes in intensity that occur over time, as the contrast agent is differentially absorbed. Following the intensity of each 3D voxel in a set of n_y coronal slices of matrix size $n_x \times n_z$ through T time-points provides a set of $n_x n_y n_z$ voxel "contrast uptake curves." Features in the spatial domain are identified in our work by sampling 2D image "patches" as described below.

4.1 Application of Single-Layer Sparse Autoencoders to Temporal Feature Learning

Approximately 1.3×10^4 time series signals were randomly sampled from the complete set of contrast uptake curves in the training dataset, excluding the background and regions affected by breathing motion. A pixel is regarded as background or a region affected by breathing motion when its image intensity falls below 10 percent of the maximum intensity in the image within the imaging time-course ($T = 40$ image volumes). Temporal features are learned by the single-layer sparse autoencoder (see (3)) from the samples, where each time series is a 40 element input vector and the N temporal features are the individual weights $\mathbf{w}_j \in \mathbb{R}^{40\times 1}$ (rows) of the weight matrix $\mathbf{W} \in \mathbb{R}^{N\times 40}$.

Certain vascular characteristics of a tissue can be represented by its time series in the 4D DCE-MRI image dataset so that the temporal features alone may be sufficient for unsupervised tissue type classification. Unsupervised tissue type classification using a single-layer sparse autoencoder was evaluated and visualization was previously reported [10]. In this work, we 1) performed dimensionality reduction in the temporal space with a single layer autoencoder network, 2) did vector quantization of the features with a sigmoid activation function, and 3) mapped the result of the vector quantization into RGB space. Three examples of these results are shown in Fig. 4.

Different tissue types are represented in different colors—liver in cyan and blood vessels in green. Heart and kidney are represented as a number of different colors, but the color pattern is consistent. Liver tumors appear as a complex pattern of different classes. With this method some tissues of different organs appear labeled as being of the same class, for example liver, spleen, and part of the heart and aorta all appear as the same cyan color. This approach does not use the spatial information in the data, so although an organ with more constant tissue characteristic can be detected and segmented, an organ which consists of a combination of different tissue types is not detected as a single entity. Our aim here is to solve these problems by using deep architectures with spatial-pooling

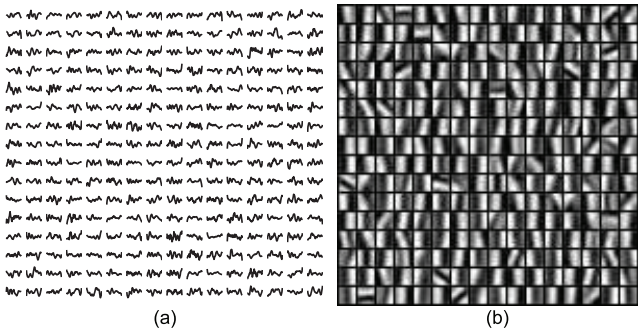

Fig. 3. Two hundred fifty-six overcomplete (a) temporal and (b) 8×8 size visual feature set learned by unsupervised sparse feature learning.

so that feature learning can incorporate progressively larger spatial regions.

4.2 Application of Single-Layer Sparse Autoencoders to Visual Feature Learning

There are many existing reports on the application of deep learning to classification using the purely spatial features found in 2D images. We describe these as "visual features," to have a clear distinction from temporal features with spatial pooling (see Section 5). In later sections we compare visual and temporal features separately, and shallow combined representation of multimodal features [27] as an augmented input to an organ classifier. We learned 2D visual features from approximately 1.3×10^4 image patches randomly sampled from the first image slice in each time-series (before the contrast agent is injected), also excluding patches containing background or voxels affected by breathing motion. For image patches of size $m \times m$ the visual features learned by the autoencoder are given by weight vectors $\mathbf{w}_j \in \mathbb{R}^{m^2 \times 1}$ and N vectors combine to give the weight matrix $\mathbf{W} \in \mathbb{R}^{N \times m^2}$.

Temporal and visual features for our data are shown in Fig. 3. They represent an overcomplete set of 256 temporal features, and there is no obvious redundancy or repetition of trivial signals (see Fig. 3a). The visual bases in Fig. 3b are learned from 8×8 image patches and show Gabor-like edge detectors of different orientations and locations, which are coherent with the results of the previous studies [9], [41]. We apply these features of different input modalities to build a part-based model for multiple organ detection (see Section 6).

5 Stacked Sparse Autoencoders

Stacked sparse autoencoders—a deep learning architecture of sparse autoencoders—are built by stacking additional unsupervised feature learning layers, and can be trained using greedy methods for each additional layer [42]. By applying a pooling operation [18], [52] after each layer, features of progressively larger input regions are essentially compressed, and this approach is used to build a part-based model for multiple organ detection.

5.1 Part-Based Model with Spatial Feature Matching

We want our model to learn object parts so that we can perform part-based object detection as was done for 2D images in [11], [12], [13], [14], [15], and for action recognition

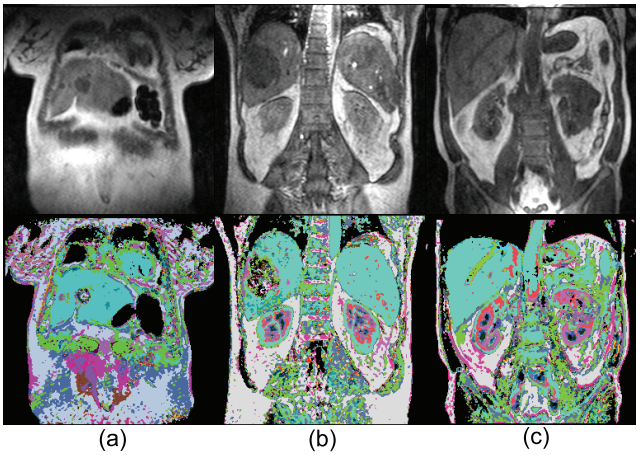

Fig. 4. Visualization of dimensionality reduction with a single-layer sparse autoencoder, where the size of the DCE-MRI temporal dimension is reduced from 40 to 16 elements. Different tissue types are visualized in different colors, and a liver tumor is represented as a complex pattern within liver (a), (b). Ambiguities in identifying some tissue types of different organs remain, with some subregions of the aorta, heart, liver, and spleen being represented as the same cyan color.

in video sequences in [17], where in [17] the parts were called the "interest-points." The visual feature learning layer already captures a certain spatial region (patch), which could then correspond to a part of an organ. But the temporal feature learning layer captures only a pixel in the spatial domain, and so a larger spatial proximity of temporal features (e.g., a part of an organ represented by temporal features) is captured by applying max-pooling:

$$\mathbf{y} = \max\{|\mathbf{W}\mathbf{x}_1, \mathbf{W}\mathbf{x}_2, \ldots, \mathbf{W}\mathbf{x}_R|\}, \tag{5}$$

where $\mathbf{W}$ is the encoding matrix (distinct for each layer), $\mathbf{x}_1, \ldots, \mathbf{x}_R$ are the input vectors to the max-pooling operation, and the max and modulus functions are applied element-wise. For the application of max-pooling after the temporal feature learning layer, $\mathbf{W}$ is the temporal feature set and $\mathbf{x}_i$ is a time-series signal. For max-pooling on an $M \times M$ patch, $R = M^2$.

As an example, a conceptual visualization of 3×3 max-pooling for 3×3 size patches in a 2D feature space using the visualization of Fig. 4c is shown in Fig. 5a, and for a 3×3 patch in 3D temporal feature space is visualized in Fig. 5b. It can be seen how the patches located on different regions of a kidney can capture the same temporal feature from a given patch size.

Features of next level spatial hierarchy—object parts with a larger region of spatial invariant feature set—are captured by a successive unsupervised feature learning on the max-pooled output of the features to learn the features of larger input region, based on what it has learned for a smaller input region. We examine in the following sections whether useful features are learned in the second feature learning layer. Examples of our model of two-layer stacked sparse autoencoder networks for learning hierarchical visual features and temporal features, each with a classifier network as the final layer, are shown in Fig. 6. Max-pooling in the visual feature learning network is applied such that each layer captures the same size of 2D spatial area as the temporal feature learning layer.

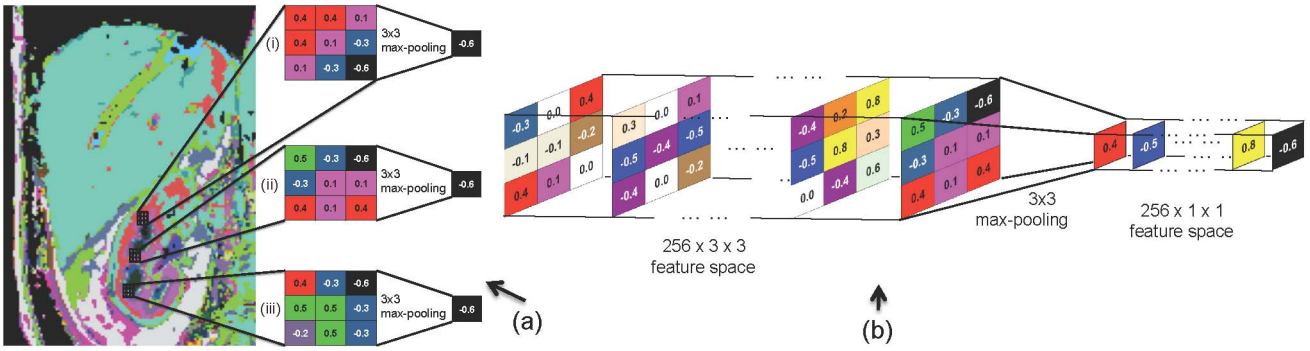

Fig. 5. (a) A conceptual visualization of max-pooling on a 2D feature space showing how it can capture the same feature for the patches at different locations in the kidney in Fig. 4c. (b) A conceptual visualization of 3×3 max-pooling on a 3D temporal feature space with 256 temporal features.

This can be compared with the "bag-of-words" model for image classification [53], [54], where the application of convolutional unsupervised feature learning for the learning of successive layers is conceptually similar to the spatial pyramid matching model [15], [55]. In our case, we are dealing with a "bag of spatial and temporal words." Some of the differences of the stacked sparse-autoencoder approach in this study to the works cited are that: 1) Spatial and temporal features are used in our work as opposed to visual features only, such as SIFT [1] and HOG [3], 2) orientation of the features is predefined in [15], and 3) next-level features with the previous hierarchy features as priors are learned in the successive autoencoder learning, whereas the same feature is used for all of the spatial pyramid hierarchies.

We compare our unsupervised learning method using sparse autoencoders with more popular, predefined features for vision and time-series: HOG and DFT. We also compare PCA as a baseline method for unsupervised feature learning, and see whether PCA can be applied to successive hierarchical feature learning as well.

5.2 Analysis and Comparison with Other Methods

The extent to which the learned features can represent our object classes is evaluated by the patch-wise classification accuracy of organs, based on the labels obtained from the roughly drawn regions of interest as shown in Fig. 1. Since the labeled regions include voxels from outside the intended organ, the accuracy cannot be 100 percent, even with perfect classification. However, as the labeled regions

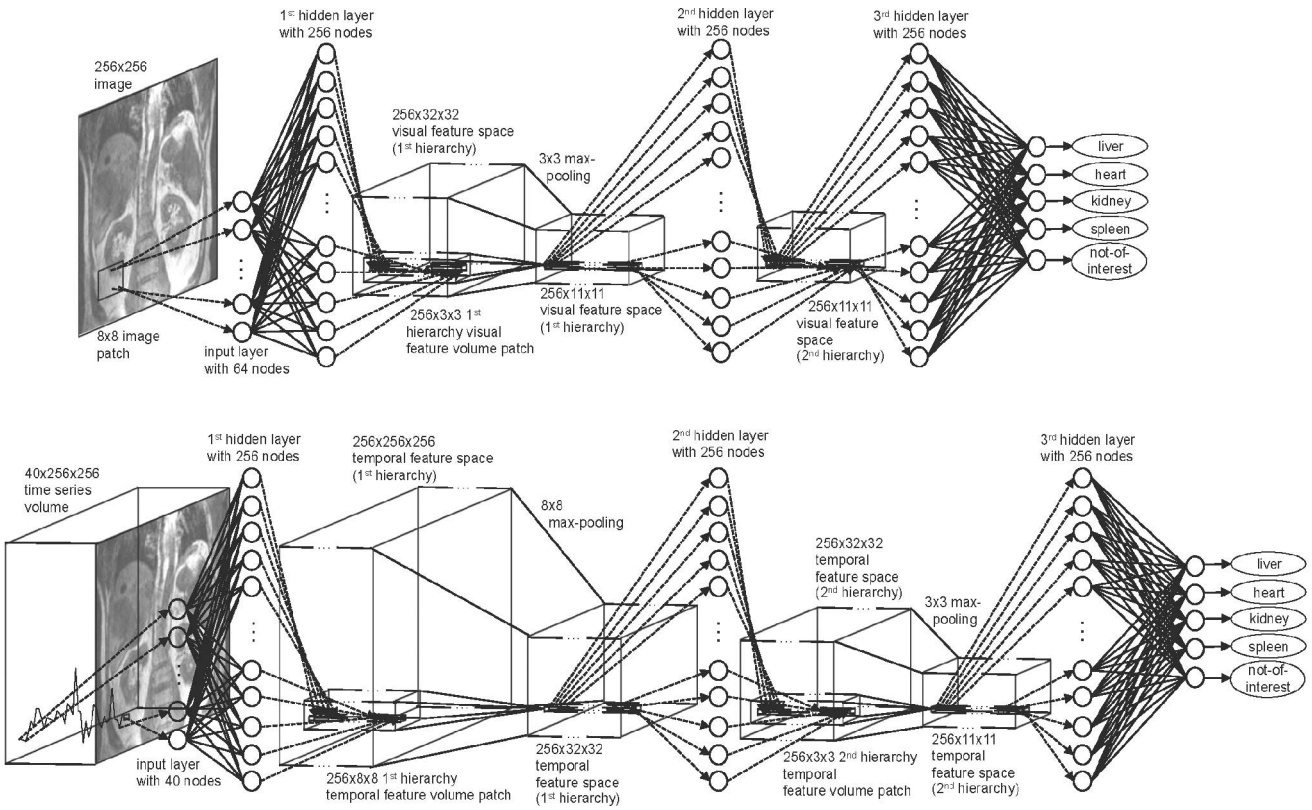

Fig. 6. The overall architecture of the visual feature learning networks (top) and temporal feature extraction networks (bottom). The first and second hidden layers are unsupervised feature learning networks and the third hidden layer is a classification network which is trained with supervision to classify patches of different organs.

TABLE 1
Part-Based Classification Accuracies and the Stacked Sparse Autoencoder (SAE) Hyperparameters (Including Patch and Pooling Sizes) Used with: First- and Second-Level Hierarchy (L1 and L2); Visual and Temporal Features (VI and TE); 16 and 256 Learned Features

method	ρ	λ	β	mps_1	mps_2	acc	$\mathrm{F1}_{\mathrm{avg}}$
VI L1 SAE-16	0.3	0.003	1	16	n/a	36.00	0.30
VI L2 SAE-16	0.1	0.0001	3	12	3	33.24	0.30
TE L1 SAE-16	0.01	0.003	1	16	n/a	51.71	0.53
TE L2 SAE-16	0.01	0.0001	1	6	3	30.68	0.23
VI L1 SAE-256	0.3	0.0001	27	16	n/a	36.37	0.33
VI L2 SAE-256	0.03	0.0001	1	16	3	36.12	0.38
TE L1 SAE-256	0.003	0.0001	1	16	n/a	51.68	0.56
TE L2 SAE-256	0.1	0.0001	1	6	3	53.10	0.54

method	mps_1	mps_2	acc	$\mathrm{F1}_{\mathrm{avg}}$
VI L1 PCA	16	n/a	36.09	0.33
VI L2 PCA	16	6	29.70	0.33
TE L1 PCA	16	n/a	42.33	0.48
TE L2 PCA	9	6	35.74	0.42
VI HOG	6	2	28.28	0.30
TE DFT	12	n/a	41.90	0.50
VI 1-CNN	12	n/a	35.46	0.34
TE 1-CNN	16	n/a	45.57	0.53

Baseline models are compared with their average classification accuracies for organs (acc), and average F1-score ($\mathrm{F1}_{\mathrm{avg}}$) of each individual object class's score is shown for comparison later in Section 5.4 with Table 2.

contain more correct voxels than incorrect voxels, we assume that higher accuracy corresponds to a better classification performance.

We compare our unsupervised feature learning methods using stacked autoencoders (SAE) to PCA, as PCA is a popular unsupervised feature learning method. We also evaluate whether PCA can learn hierarchical features when applied successively with max-pooling, where we use 16 principal components for projection of the input data. We test SAE with both 16 (SAE-16) and 256 (SAE-256) learned features to make a fair comparison with PCA using 16 features, and to see the effect of the number of SAE learned features and overcompleteness on classification performance.

A single layer convolutional network (1-CNN) is also tested to see the effect of pretraining on the features, and in addition, a single convolutional network using HOG visual features and DFT temporal features. Classification is done with a single-layer classifier network, where the parameters for the training are chosen by a CV test on small subsets of the data. With the best parameters so derived, the final accuracy is reported after additional training and CV using larger subsets of the dataset. Unsupervised feature learning and classifier training uses only dataset A, and classification is performed only with dataset B to show the applicability of the features learned unsupervised to an unseen dataset.

Deep networks are known to be difficult to train, and this certainly applies to stacked sparse autoencoder training where there are many hyperparameters affecting the behavior of the model. The hyperparameters required for training the sparse autoencoders are the target mean activation ρ, the weight of the sparsity penalty β, and the weight decay for the backpropagation optimization λ. We used a coordinate ascent-like method to optimize these for each layer, together with the patch and pooling sizes. Coordinate ascent consists of optimizing each parameter while the others are fixed, and repeating this process for a certain number of iterations until the performance converges.

It was optimized on the classification accuracy of the entire object classes overall, and the optimized average accuracy is shown in Table 1, with the corresponding hyperparameters and patch/pooling sizes. During the optimization process, the classification performance of each individual object class is recorded as well, but in F1-score rather than in accuracy because in this case the true/false label is biased for each class to the others, where F1-score is defined as $\mathrm{F1} = 2 \cdot \frac{\mathrm{precision} \cdot \mathrm{recall}}{\mathrm{precision} + \mathrm{recall}}$, and $\mathrm{precision} = tp/(tp + fp)$, $\mathrm{recall} = tp/(tp + fn)$, tp: true positives, fp: false positives, fn: false negatives. The averaged F1-score $\mathrm{F1}_{\mathrm{avg}}$ over each class's F1-score for the optimal overall classification accuracy and the corresponding hyperparameter settings are shown in Table 1 as well, and will be discussed later in Section 5.4.

As was mentioned earlier, typical recognition tasks on medical images are to recognize objects in each component 2D slice in a 3D volume, and correspondingly an object category can have many appearances on its 2D slices. However, an organ type can contain different tissue types, which are represented in the characteristics of the DCE-MRI time-series signal, and the constant characteristic penetrates through the 3D slices. We can also see that second hierarchy features do not necessarily give better results and can be worse for some features. This was also seen in [9] for SAE, and we will discuss this further in Section 5.4.

5.3 Unsupervised Learning of Object Classes

Patches of organ classes filtered with each of the features compared are shown in Fig. 7 as 2D scatter plots. From the training dataset, 1,500 patches are sampled randomly for each organ category, filtered with the features, and the dimension of the patches is reduced to two using PCA to aid visualization. It is noticeable that the object classes are very well captured by the 16 temporal features learned by single layer unsupervised sparse feature learning (TE-L1-SAE-16). The object classes are reasonably well separated with 1-convolutional DFT temporal features (TE-DFT), but less so with 1-convolutional PCA (TE-L1-PCA) or 1-convolutional temporal features alone (TE-1-CNN). It is not obvious from the plots whether the overcomplete methods (with 256 features) have learned features that better discriminate the organ classes and may therefore be expected to give better classification performance. This is due to the dramatic dimensionality reduction needed for visualization (256 to 2), as the classification performance with those features shows good results in Table 1. Overall, temporal features have better classification performance than visual-only features.

Fig. 7 appears to show nearly perfect categorization of self-learned features for the TE-L1-SC-16 approach, but the reason the classification accuracy is not higher than that in Table 1 can be seen in Fig. 8. Fig. 8 shows 1,500 randomly sampled patches of a new subset of the training dataset

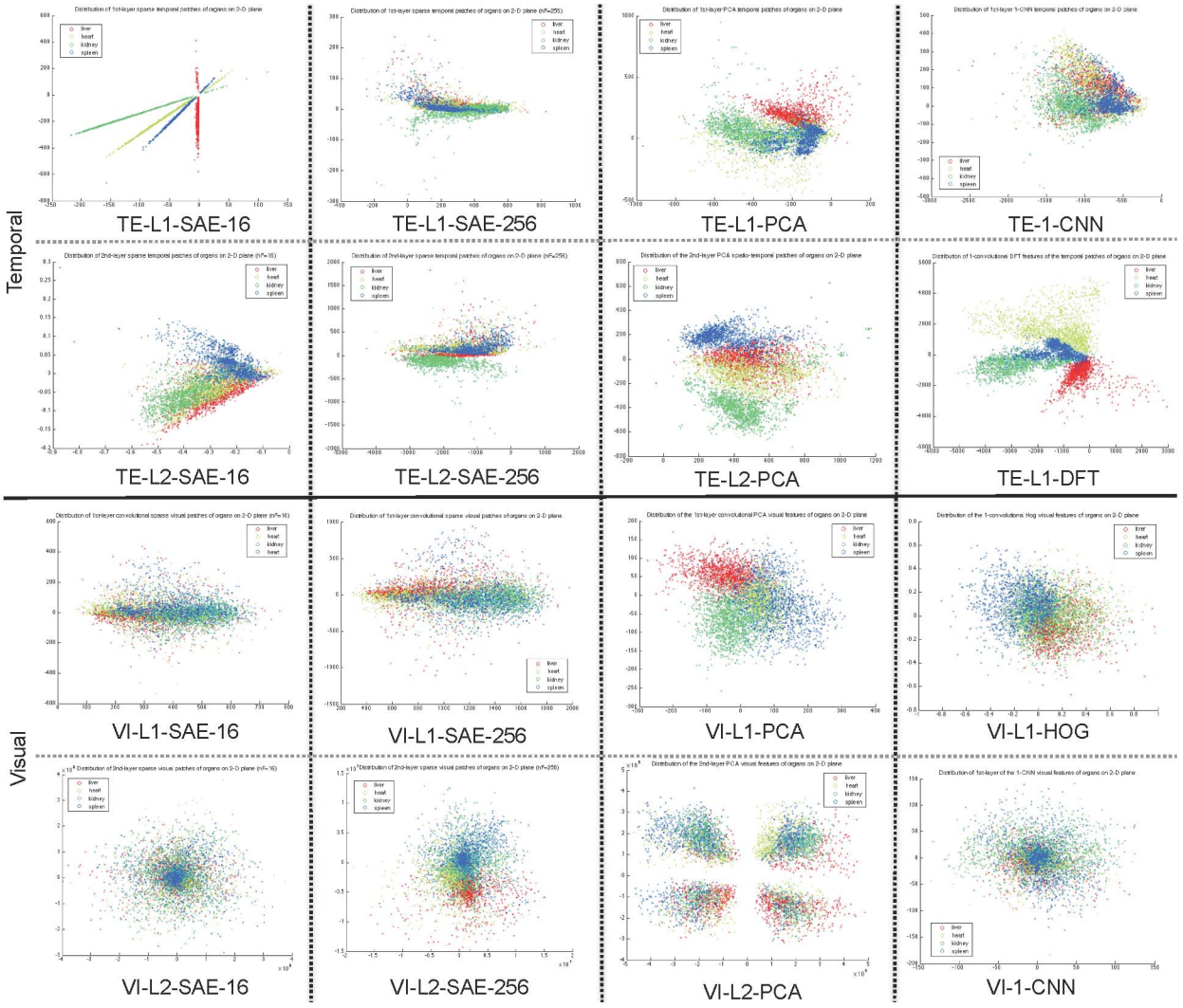

Fig. 7. Scatter plots showing 1,500 randomly sampled patches of the organ object classes (red: liver, yellow: heart, green: kidney, blue: spleen) in the training dataset with each of the feature learning methods and projected onto 2D space using PCA.

(liver patient dataset) and CV dataset (kidney patient dataset), each filtered with the same temporal features used in Fig. 7. Although the TE-L1-SAE-16 features separate the organ classes very well in the new subset of the training dataset, the separation is not nearly as clear on the CV dataset. This performance reduction could be mitigated by unsupervised feature learning on a larger subset of (more heterogeneous) training data, but it is technically challenging to train on a very large scale dataset. In Fig. 8, patches of the CV dataset filtered with TE-L1-DFT and TE-L2-PCA—which showed good separation of the organ classes with the training data—are also shown, and they too show less clear separation in the CV dataset.

5.4 Context-Specific Feature Learning

Different organ classes have different properties, and therefore it seems reasonable to suppose that the task of separating a given organ class from all other classes might be best achieved by learning the optimal feature for that particular organ, rather than by training on the average separation performance for all classes. Applying this in the context of action recognition as in [17] the question would be: Can one obtain better performance with a feature learning model optimized specifically for "hand waving," for example, rather than using the same feature learning model that simultaneously tries to classify, say, "running" and all the other different actions studied?

It is normally time consuming and difficult to design a new feature-learning model for every object class, but deep learning requires very little modification. In our study, we applied a model with the same basic design for both visual and temporal feature learning. Moreover, features of different characteristics can be learned by tuning the hyperparameters in the learning model, as was studied in [9].

In principle one would optimize the hyperparameters in Table 1 separately for each object class, but the computational resources required to do this exceeded what was available for this study. Instead, during the hyperparameter optimization process in Section 5.2 and Table 1, we picked parameter sets with the best F1-scores of each object class along the trajectory of optimization process. The best F1-scores for each object class along the optimization process of overall classification accuracies in Table 1 are shown in Table 2 as $\mathrm{F1}_{\mathrm{tmp}}$, with their corresponding hyperparameter sets.

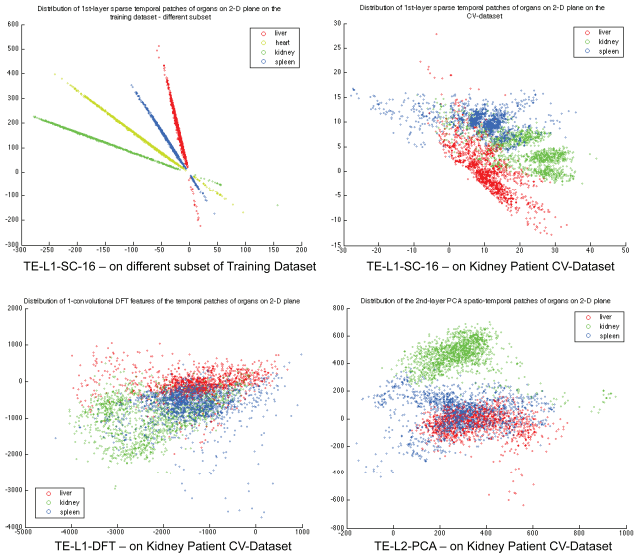

Fig. 8. Scatter plots showing 1,500 randomly sampled patches of a different subset of the training dataset and the CV liver patient dataset. The patches are processed and displayed in the same way as for Fig. 7. Since the scans of the CV dataset are focused on the kidneys, heart does not appear in those images due to its anatomical location; therefore heart is absent in the CV dataset.

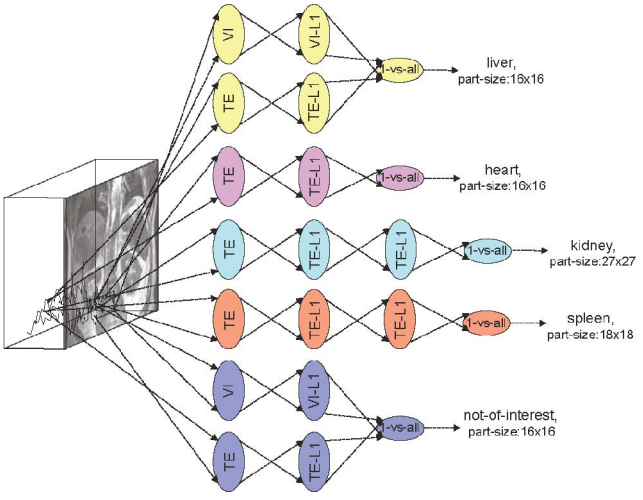

Fig. 9. A conceptual visualization of the usage of context-specific features with SAE in classification. Patches of different modalities are sampled from the dataset and go through different feature networks to be classified as an object part of an organ category.

We then train one-versus-all classifiers by logistic regression using the parameter sets with the same inputs as the multiclass classifier from the convolutional network. The one-versus-all classification accuracies with equal number of true/false labels, optimized for each object class, is shown as $\mathrm{acc}_{\mathrm{opt}}$ in Table 2. For the results using visual and temporal features only, the accuracy with first autoencoder layer is denoted by $\mathrm{acc}_{l+/-}$ if $\mathrm{acc}_{\mathrm{opt}}$ was achieved by second autoencoder layer, whereas $\mathrm{acc}_{l+/-}$ represents the accuracy with second autoencoder layer if the $\mathrm{acc}_{\mathrm{opt}}$ was achieved by first autoencoder layer. A shallow combined representation of multimodal features [27] is examined as well for both hierarchy autoencoders, and their classification accuracies both with first (acc_{l_1}) and second (acc_{l_2}) autoencoder layers are shown in the right-hand section of the table. The final classifier networks with the context-specific feature learning model for each object class is shown in Fig. 9.

Here we observe that:

1. Organ-specific parameter selections generally give improved performance, although not always by a large amount.
2. Temporal features alone give good classification performance for heart, kidney, and spleen. (Note: The heart was classified on a different dataset because of its absence on the validation set—see the caption to Fig. 8.
3. Generally, the second-layer visual features showed better performance than the first-layer features, while this was not evident for the temporal features, which had a worse performance for liver and heart. We hypothesize that this may be related to the fact that the second-layer visual features capture a larger region (e.g., 48×48) than the second-level temporal features (e.g., 24×24).
4. Shallow combination of visual and temporal features showed better performance than features of either modality alone for liver and "not-of-interest" (NOI) tissues, although the increase in accuracy for liver was small.

It is also possible to draw some conclusions from these results about the parameter settings for deep network models with SAE: 1) The optimized sparsity ρ in the second layer tends to be lower than that in the first layer to capture fewer and larger size higher hierarchy features. 2) The weight decay λ and regularization β affect the behavior of the autoencoder less than the sparsity parameter ρ. 3) In the temporal domain, a larger feature set (256) was selected for organs than for NOI tissues (16 features)—this is probably

TABLE 2
Model and Hyperparameters of Visual and Temporal Features for Each Organ Class for the Context-Specific Feature Learning

visual features									
organ	model	ρ	λ	β	mps_1	mps_2	$\mathrm{F1}_{\mathrm{tmp}}$	$\mathrm{acc}_{\mathrm{opt}}$	$\mathrm{acc}_{l+/-}$
liver	L2-256	0.03	0.0001	1	16	3	0.53	58.10	52.57
heart	L2-256	0.03	0.0001	1	16	3	0.43	63.25	55.23
kidney	L1-256	0.3	0.0001	27	16	n/a	0.32	52.97	49.82
spleen	L2-256	0.03	0.0001	1	16	3	0.51	73.21	61.45
NOI	L2-256	0.03	0.0001	1	16	3	0.35	60.38	59.88

temporal features									
organ	model	ρ	λ	β	mps_1	mps_2	$\mathrm{F1}_{\mathrm{tmp}}$	$\mathrm{acc}_{\mathrm{opt}}$	$\mathrm{acc}_{l+/-}$
liver	L1-256	0.3	0.0001	27	16	n/a	0.50	64.78	57.46
heart	L1-256	0.3	0.0001	1	16	n/a	0.81	84.88	84.54
kidney	L2-256	0.3	0.0001	1	9	3	0.72	81.82	79.16
spleen	L2-256	0.01	0.0001	1	6	3	0.54	78.44	73.13
NOI	L1-16	0.3	0.0001	1	8	3	0.58	58.09	52.01

combined		
organ	acc_{l_1}	acc_{l_2}
liver	66.80	62.62
heart	68.35	65.59
kidney	68.28	79.41
spleen	66.97	63.44
NOI	70.61	62.01

The classification accuracy with the chosen model for each organ class $\mathrm{acc}_{\mathrm{opt}}$ is shown for the CV dataset for all organs except for heart (which does not appear in the CV data and so is tested on a subset of training dataset). Accuracy with higher/lower autoencoder layer ($\mathrm{acc}_{l+/-}$) is compared with that of the first/second layer giving the optimal accuracy ($\mathrm{acc}_{\mathrm{opt}}$). The average F1-score in picking the parameters in the optimization process in Table 1 is also shown: $\mathrm{F1}_{\mathrm{tmp}}$ (NOI = Not of interest).

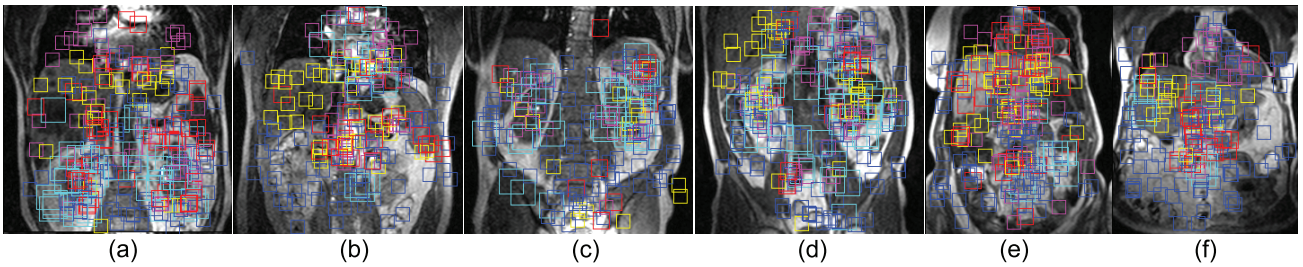

Fig. 10. Classification results of part-based organ detection (yellow: liver, magenta: heart, cyan: kidney, red: spleen, blue: NOI). (a), (b): Liver patient training dataset, (c), (d): kidney patient CV dataset, (e), (f): liver patient from a clinical trial. The patch size for liver and heart is 16×16, for kidney 27×27, spleen 18×18, and for NOI 24×24. The various parameters, including the patch sizes for each organ class, are chosen based on the results shown in Table 2.

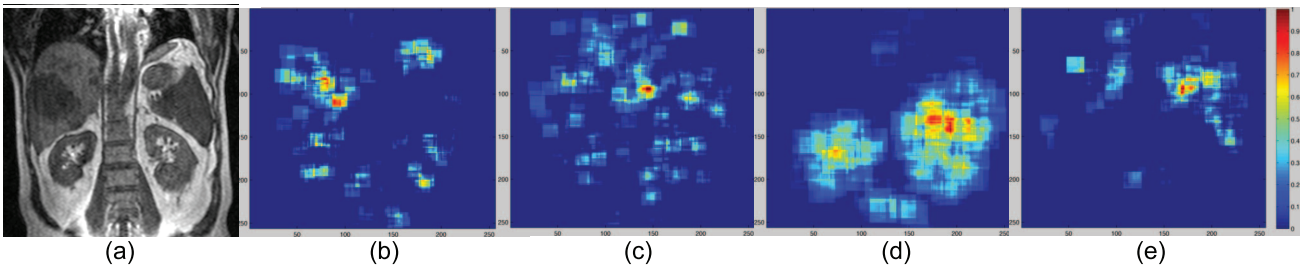

Fig. 11. Source image of a liver tumor patient (a), with probability maps for: (b) liver, (c) heart, (d) kidney, and (e) spleen.

because NOI does not represent a specific object class, and so the fewer features are used, the less prone the model is to overfitting specific background entities in the training data.

6 Part-Based Multi-Organ Detection

From the results in Table 2, we chose: TE-L1-256 features for heart, TE-L2-256 for kidney and spleen, the combined L1-16 features for NOI, and the combined L1-256 features for liver. Some of the part-based organ detection results in training, CV, and the test dataset are shown in Fig. 10. As might be expected, the results reflect the organs at which the image datasets themselves were targeted. Thus, liver is better recognized on the scans whose purpose was to image liver tumors, and kidney in renal cell carcinoma patients.

6.1 Probabilistic Part-Based Organ Detection

As more patches are classified correctly than incorrectly to their corresponding category of organ, we perform probabilistic part-based organ detection, first by generating a probability map for each organ, and then by selecting a threshold to generate a binary mask, using the features selected in Table 2. The probability map is generated using 1,000 randomly sampled patches, where the sampling location is on the nonbackground regions and the regions not affected by the breathing motion. Consider the probability map for organ A. We iterate through all patches and for each pixel location (i, j) of the map, we increase its score for being an organ A by 1 unit if the location (i, j) is in a patch classified as the organ A. On the other hand, if (i, j) is in the patch but the classifier returns a different class, we subtract 0.2 from the score. The final scores after all patches have been considered is normalized by dividing by the maximum score in the image. The importance of including the NOI class in all our analyses is thus clear. An example of a probability map is shown in Fig. 11.

For the final organ detection, we perform a number of simple postprocessing steps. From the probability map, we obtain the largest, contiguous region for which the probability is larger than a preset threshold, except for kidney, where we get two such regions using our prior anatomical knowledge that there are more likely to be two kidneys than one. There are cases where only one kidney appears in the image, and these cases are accounted for by ignoring any regions that are smaller than 200 pixels. The thresholds are organ-specific (see Table 3) and were selected by examining the pixel-wise precision and recall on the CV dataset. We then apply convex hull processing to the final regions for each object category to outline the regions smoothly.

Some examples of the final visualization of multi-organ detection are shown in Fig. 12. Organs are generally well detected, but the performance of the algorithm varies from patient to patient. Encouragingly, even unusually large livers and those with metastases are correctly assigned to the liver organ class. Notice that this method of performing organ recognition does not lead to mutually exclusive regions, something which is a consequence of independently generating and processing the probability maps for each organ.

TABLE 3
Selected Threshold and Pixel-Wise Precision/Recall with the Threshold for the Organ Classes on the CV Dataset, Except Heart, Which Was Validated on the Training Dataset

	threshold	pixel-prec./recall	object-prec./recall
liver	0.1	0.86/0.32	0.96/0.91
heart	0.2	0.83/0.25	0.57/0.67
kidney	0.4	0.45/0.31	0.46/0.91
spleen	0.3	0.45/0.37	0.40/0.72

Object-wise precision/recall was validated on the test dataset of a clinical trial.

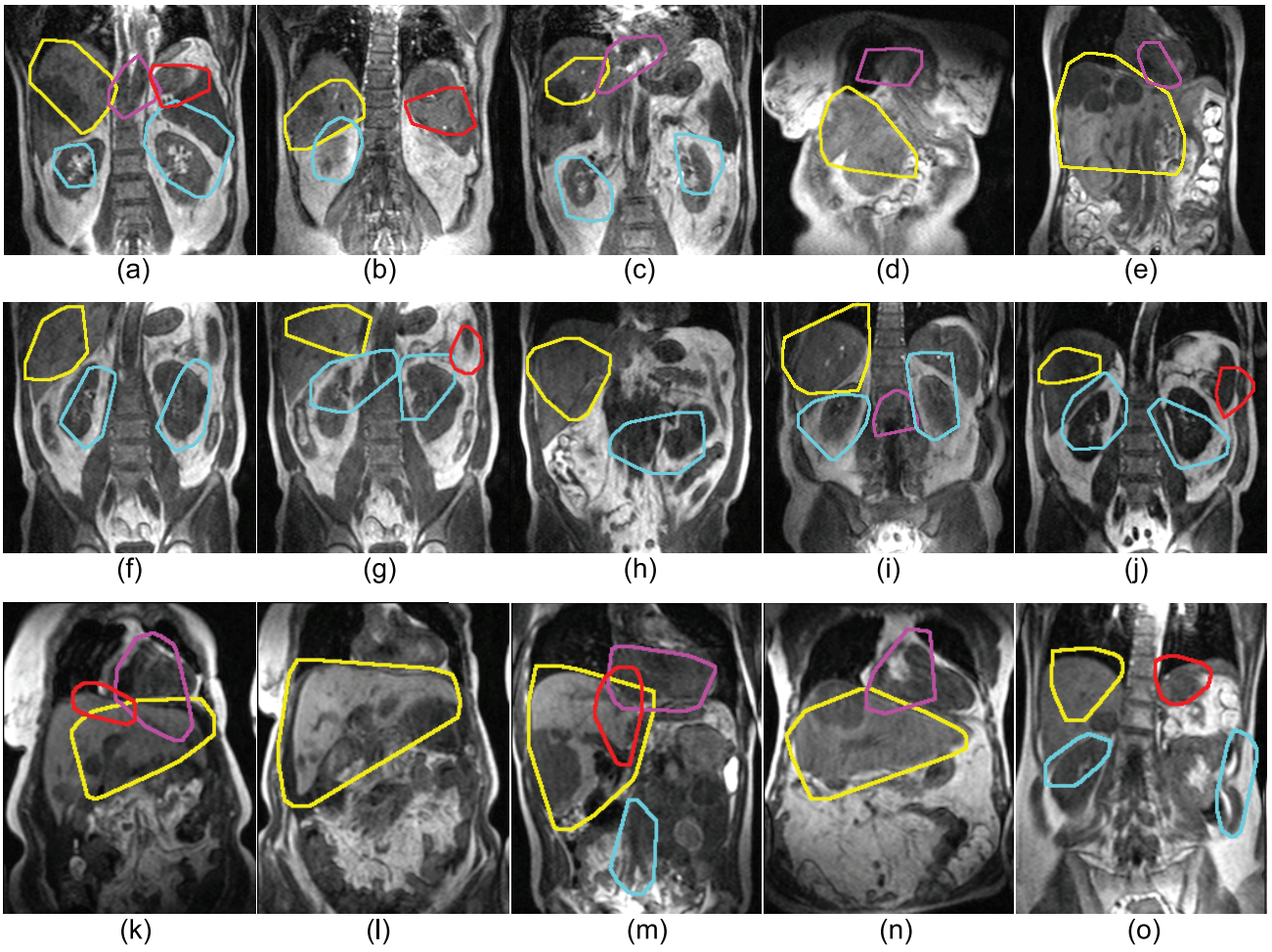

Fig. 12. Some examples of the final multiorgan detection (yellow: liver; magenta: heart; cyan: kidney; red: spleen) on the training dataset (first row), CV dataset (second row), and test dataset (third row). Liver and kidney are well detected, whereas spleen is less well detected. In some images, the detected heart region also contains aorta ((a) and (i)), which is probably because the signal uptake pattern in the aorta and the heart are similar. The liver class detected includes both normal appearing tissues and tumor tissues. Liver tumor is seen in most of the liver patient images (first and third row), with some largely abnormal liver shapes ((e) and (m)).

Pixel-wise precision and recall scores on the CV kidney patient dataset, and object-wise precision and recall scores on the test liver patient dataset from a clinical trial are shown in Table 3. All of the organs types are well detected, with average 0.60 in precision and 0.80 in recall.

7 Discussion and Future Work

Although our model showed good performance, it is likely that even better results would be achieved if the training were performed on a larger dataset. Future work collecting a large dataset for full training, CV, and test on various patient studies should lead to interesting results. A further area of research would be to investigate the balance between tightly controlling the variability in the input data (for example, by standardizing the MR sequence parameters and/or preselecting the patients) and providing the algorithms with more heterogeneous data. The former case would presumably lead to better classification in test images that resembled the training data, while the latter might lead to models that are more widely applicable but have poorer performance in individual cases.

Optimization was performed with a coordinate ascent-like paradigm using average classification accuracies, and the best model for each organ category was chosen by examining the trajectories of the performance evaluation scores for each organ class. Even though we started from an empirically "good" set of hyperparameters, it is possible that the coordinate ascent ended up in a local optima. It would be interesting to do a full parameter search for each organ category or use some recently introduced methods for optimizing hyperparameters without a full search [56], [57].

PCA was chosen for the main baseline method of unsupervised feature learning for comparison as it is the most widely used one. Although, even when not as widely used as PCA, sparse coding [47] or Independent Component Analysis [58] with max-pooling have somewhat closer relation to the suggested model as these allow learning an overcomplete feature set. Therefore, it will also be meaningful to compare these to the suggested method, as well as with some of the other deep learning architectures for unsupervised feature learning, such as deep belief networks [18], [21], [41], [44] (with RBM) or convolutional networks [16], [59] (possibly initializing each layer using autoencoders).

For the reasons alluded to in the introduction, the datasets used for this research cannot be made public in the foreseeable future. However, a study performing multimodal brain

tumor analysis using a related approach has been described in [34], as part of a brain tumor segmentation challenge. There, we used a combined approach of unsupervised feature learning, clustering, and one-versus-all classification with logistic regression. The data are publicly available and the result with this approach is compared with a number of the other popular methods.

8 Conclusion

Visual and temporal hierarchical features have been learned from roughly labeled DCE-MRI images of patients with different types of tumor, using a deep learning approach. By contrast with the usual object-detection environment, the challenges for object detection in patient datasets are as follows: 1) The organs with diseases are sometimes grossly abnormal, 2) the shape of the organs shown by slices in a 3D medical image differ between slices in ways that are sometimes challenging even for a trained radiologist, and 3) it is hard to obtain many training datasets and the ground truth is hard to define. With unsupervised hierarchical feature learning, organ classes are learned without detailed human input, and only a "roughly" labeled dataset was required to train the classifier for multiple organ detection.

Part-based multi-organ detection was performed on a heterogeneous patient dataset of three independent studies with different disease foci. Training was done only in one dataset and object recognition was done on an unseen dataset with good performance. This method can accommodate a range of organ types, including those with metastases and very abnormal shapes. To the best of our knowledge, there have been no previous studies using deep learning for organ detection in heterogeneous MRI datasets from patients, and the results of this pilot study are promising. Further applications may be developed from this, with additional segmentation algorithms being combined with our deep learning approach.

Acknowledgments

The authors would like to acknowledge the support received from the CRUK and EPSRC Cancer Imaging Centre in association with the MRC and the Department of Health (England) grant C1060/A10334 and also NHS funding to the NIHR Biomedical Research Centre and the NIHR Clinical Research Facility.

References

[1] D.G. Lowe, "Object Recognition from Local Scale-Invariant Features," *Proc. IEEE Int'l Conf. Computer Vision,* vol. 2, pp. 1150-1157, 1999.

[2] H. Bay, T. Tuytelaars, and L.V. Gool, "Surf: Speeded Up Robust Features," *Proc. European Conf. Computer Vision,* pp. 404-417, 2006.

[3] N. Dalal and B. Triggs, "Histograms of Oriented Gradients for Human Detection," *Proc. IEEE Conf. Computer Vision and Pattern Recognition* pp. 886-893, 2005.

[4] L. Fei-Fei, R. Fergus, and P. Perona, "One-Shot Learning of Object Categories," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 29, no. 4, pp. 594-611, Apr. 2006.

[5] G. Griffin, A. Holub, and P. Perona, "The Caltech-256 Object Category Dataset," technical report, California Inst. of Technology, 2007.

[6] D.J. Collins and A.R. Padhani, "Dynamic Magnetic Resonance Imaging of Tumor Perfusion," *IEEE Eng. in Medicine and Biology Magazine,* vol. 23, no. 5, pp. 65-83, Sept./Oct. 2004.

[7] A.R. Padhani, G. Liu, D. Mu-Koh, T.L. Chenevert, H.C. Thoney, T. Takahara, A. Dzik-Jurasz, B.D. Ross, M.V. Cauteren, D. Collins, D.A. Hammoud, R.JS. Gordon, T. Bachir, and C.L. Peter, "Diffusion-Weighted Magnetic Resonance Imaging as a Cancer Biomarker: Consensus and Recommendations," *Neoplasia,* vol. 11, pp. 102-125, 2009.

[8] R. Raina, A. Battle, H. Lee, B. Packer, and A.Y. Ng, "Self-Taught Learning: Transfer Learning from Unlabled Data," *Proc. Int'l Conf. Machine Learning,* pp. 759-766, 2007.

[9] I. Goodfellow, Q. Le, A. Saxe, H. Lee, and A. Ng, "Measuring Invariances in Deep Networks," *Advances in Neural Information Processing Systems,* vol. 22, pp. 646-654, 2009.

[10] H. Shin, M. Orton, D.J. Collins, S. Doran, and M.O. Leach, "Autoencoder in Time-Series Analysis for Unsupervised Tissues Characterisation in a Large Unlabelled Medical Image Data Set," *Proc. IEEE Int'l Conf. Machine Learning and Application,* pp. 259-264, 2011.

[11] M. Weber, M. Welling, and P. Perona, "Towards Automatic Discovery of Object Categories," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* pp. 101-108, 2000.

[12] R. Fergus, P. Perona, and A. Zisserman, "Object Class Recognition by Unsupervised Scale-Invariant Learning," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* pp. 264-271, 2003.

[13] E.J. Bernstein and Y. Amit, "Part-Based Statistical Models for Object Classification and Detection," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* pp. 734-740, 2005.

[14] A. Torralba, K.P. Murphy, and W.T. Freeman, "Sharing Visual Features for Multiclass and Multiview Object Detection," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 29, no. 5, pp. 854-869, May 2007.

[15] P.F. Felzenszwalb, R.B. Girshick, D. McAllester, and D. Ramanan, "Object Detection with Discriminatively Trained Part-Based Models," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 32, no. 9, pp. 1627-1645, Sept. 2010.

[16] S. Ji, W. Xu, M. Yang, and K. Yu, "3D Convolutional Neural Networks for Human Action Recognition," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 35, no. 1, pp. 221-321, Jan. 2013.

[17] J.C. Niebles, J. Wang, and L. Fei-Fei, "Unsupervised Learning of Human Action Categories Using Spatial-Temporal Words," *Int'l J. Computer Vision,* vol. 79, pp. 299-318, 2008.

[18] H. Lee, R. Grosse, R. Ranganath, and A. Ng, "Unsupervised Learning of Hierarchical Representations with Convolutional Deep Belief Networks," *Comm. ACM,* vol. 54, no. 10, pp. 95-103, 2011.

[19] M.A. Ranzato, F.J. Huang, Y.L. Boureau, and Y. LeCun, "Unsupervised Learning of Invariant Feature Hierarchies with Applications to Object Recognition," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* pp. 1-8, June 2007.

[20] M.D. Zeiler, G.W. Tayler, and R. Fergus, "Adaptive Deconvolutional Networks for Mid and High Level Feature Learning," *Proc. IEEE Int'l Conf. Computer Vision,* pp. 2018-2025, Nov. 2011.

[21] K. Sohn, D.Y. Jung, H. Lee, and A.O. Hero, "Efficient Learning of Sparse, Distributed, Convolutional Feature Representations for Object Recognition," *Proc. IEEE Int'l Conf. Computer Vision,* pp. 2643-2650, 2011.

[22] X. Glorot, A. Bordes, and Y. Bengio, "Domain Adaptation for Large-Scale Sentiment Classification: A Deep Learning Approach," *Proc. 28th Int'l Conf. Machine Learning,* pp. 513-520, 2011.

[23] K. Yu, W. Xu, and Y. Gong, "Deep Learning with Kernel Regularization for Visual Recognition," *Advances in Neural Information Processing Systems,* vol. 21, pp. 1889-1896, 2008.

[24] L. Bazzani, N. Freitas, H. Larochelle, V. Murino, and T. Jo-Anne, "Learning Attentional Policies for Tracking and Recognition in Video with Deep Networks," *Proc. 28th Int'l Conf. Machine Learning,* pp. 937-944, 2011.

[25] D.E. Rumelhart and J.L. McClelland, "Parallel Distributed Processing: Psychological and Biological Models," *Information Processing in Dynamical Systems: Foundations of Harmony Theory,* vol. 1, pp. 194-281, 1986.

[26] T. Schmah, G.E. Hinton, R. Zemel, S.L. Small, and S. Strother, "Generative versus Discriminative Training of RBMs for Classification of fMRI Images," *Advances in Neural Information Processing Systems,* vol. 21, pp. 1409-1416, 2009.

[27] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A.Y. Ng, "Multimodal Deep Learning," *Proc. Int'l Conf. Machine Learning,* 2011.

[28] Q.V. Le, W.Y. Zou, S.Y. Yeung, and A.Y. Ng, "Learning Hierarchical Invariant Spatio-Temporal Features for Action Recognition with Independent Subspace Analysis," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* pp. 3361-3368, 2011.

[29] L. Li and B.A. Prakash, "Time Series Clustering: Complex Is Simpler!" *Proc. Int'l Conf. Machine Learning,* pp. 185-192, 2011.

[30] E. Geremia, B. Menze, O. Clatz, E. Konukoglu, A. Criminisini, and N. Ayache, "Spatial Decision Forests for MS Lesion Segmentation in Multi-Channel MR Images," *Proc. Medical Image Computing and Computer-Assisted Intervention,* pp. 111-118, 2010.

[31] J. Corso, E. Sharon, S. Dube, S. El-Saden, U. Sinha, and A. Yuille, "Efficient Multilevel Brain Tumor Segmentation with Integrated Bayesian Model Classification," *IEEE Trans. Medical Imaging,* vol. 27, no. 5, pp. 629-640, May 2008.

[32] M. Clark, L. Hall, D. Goldgof, R. Velthuizen, F. Murtagh, and M. Silbiger, "Automatic Tumor Segmentation Using Knowledge-Based Techniques," *IEEE Trans. Medical Imaging,* vol. 17, no. 2, pp. 187-201, Apr. 1998.

[33] A. Farhangfar, R. Greiner, and C. Szepesvá'ri, "Learning to Segment from a Few Well-Selected Training Images," *Proc. Int'l Conf. Machine Learning,* pp. 305-312, 2009.

[34] H. Shin, "Hybrid Clustering and Logistic Regression for Multi-Modal Brain Tumor Segmentation," *Proc. Workshops and Challenges in Medical Image Computing and Computer-Assisted Intervention,* 2012.

[35] T. Okada, K. Yokota, M. Hori, M. Nakamoto, H. Nakamura, and Y. Sato, "Construction of Hierarchical Multi-Organ Statistical Atlases and Their Application to Multi-Organ Segmentation from CT Images," *Proc. Medical Image Computing and Computer-Assisted Intervention,* pp. 502-509, 2008.

[36] M.G. Linguraru and R.M. Summers, "Multi-Organ Automatic Segmentation in 4D Contrast-Enhanced Abdominal CT," *Proc. IEEE Int'l Symp. Biomedical Imaging,* pp. 45-48, 2008.

[37] O. Pauly, B. Glocker, A. Criminisi, D. Mateus, A. Möller, S. Nekolla, and N. Navab, "Fast Multiple Organ Detection and Localization in Whole-Body MR Dixon Sequences," *Proc. Medical Image Computing and Computer-Assisted Intervention,* pp. 239-247, 2011.

[38] J. Iglesias, E. Konukoglu, A. Montillo, Z. Tu, and A. Crimisini, "Combining Generative and Discriminative Models for Semantic Segmentation of CT Scans via Active Learning," *Information Processing in Medical Imaging,* vol. 6801, pp. 25-36, 2011.

[39] M.R. Orton, K. Miyazaki, D.M. Koh, D.J. Collins, D.J. Hawkes, D. Atkinson, and M.O. Leach, "Optimizing Functional Parameter Accuracy for Breath-Hold DCE-MRI of Liver Tumours," *Physics in Medicine and Biology,* vol. 54, pp. 2197-2215, 2009.

[40] A.P. Dempster, N.M. Laird, and D.B. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm," *J. Royal Statistical Soc.,* vol. 39, pp. 1-38, 1977.

[41] R. Marc'Aurelio, L. Boureau, and Y. LeCun, "Sparse Feature Learning for Deep Belief Networks," *Advances in Neural Information Processing Systems,* vol. 20, 2007.

[42] Y. Bengio, P. Lamblin, D. Popovici, H. Larochelle, and U. Montreal, "Greedy Layer-Wise Training of Deep Networks," *Advances in Neural Information Processing Systems,* vol. 19, p. 153, 2007.

[43] P. Vincent, H. Larochelle, Y. Bengio, and P.A. Manzagol, "Extracting and Composing Robust Features with Denoising Autoencoders," *Proc. 25th Int'l Conf. Machine Learning,* pp. 1096-1103, 2008.

[44] H. Lee, C. Ekanadham, and A. Ng, "Sparse Deep Belief Net Model for Visual Area V2," *Advances in Neural Information Processing Systems,* vol. 20, pp. 873-880, 2008.

[45] H. Larochelle, Y. Bengio, J. Louradour, and P. Lamblin, "Exploring Strategies for Training Deep Neural Networks," *The J. Machine Learning Research,* vol. 10, pp. 1-40, 2009.

[46] S. Kullback and R.A. Leibler, "On Information and Sufficiency," *The Annals of Math. Statistics,* vol. 22, pp. 79-86, 1951.

[47] B.A. Olshausen and D.J. Field, "Sparse Coding with an Overcomplete Basis Set: A Strategy Employed by VI?" *Vision Research,* vol. 37, pp. 3311-3325, 1997.

[48] D.E. Rumelhart, G.E. Hinton, and R.J. Williams, "Learning Representations by Back-Propagating Errors," *Nature,* vol. 323, pp. 533-536, 1986.

[49] D.C. Liu and J. Nocedal, "On the Limited Memory BFGS Method for Large Scale Optimization," *Math. Programming,* vol. 45, pp. 503-528, 1989.

[50] A. Coates, H. Lee, and A.Y. Ng, "An Analysis of Single-Layer Networks in Unsupervised Feature Learning," *Proc. Int'l Conf. Artificial Intelligence and Statistics,* vol. 15, pp. 215-223, 2011.

[51] G.E. Hinton, S. Osindero, and Y.W. Teh, "A Fast Learning Algorithm for Deep Belief Nets," *Neural Computation,* vol. 18, no. 7, pp. 1527-1554, 2006.

[52] D. Hubel and T. Wiesel, "Receptive Fields and Functional Architecture in Two Nonstriate Visual Areas (18 and 19) of the Cat," *J. Neurophysiology,* vol. 28, pp. 229-89, 1965.

[53] J. Sivic, B.C. Russel, A.A. Efros, A. Zisserman, and W.T. Freeman, "Discovering Objects and Their Location in Images," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* pp. 370-377, 2005.

[54] E. Nowak, F. Jurie, and B. Triggs, "Sampling Strategies for Bag-of-Features Image Classification," *Proc. European Conf. Computer Vision* pp. 490-503, 2006.

[55] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* pp. 2169-2178, 2006.

[56] J. Bergstra and Y. Bengio, "Random Search for Hyper-Parameter Optimization," *J. Machine Learning Research,* vol. 13, pp. 281-305, 2012.

[57] J. Snoek, H. Larochelle, and R.P. Adams, "Practical Bayesian Optimization of Machine Learning Algorithms," *Proc. Advances in Neural Information Processing Systems Conf.,* 2012.

[58] Q.V. Le, A. Karpenko, J. Ngiam, and A.Y. Ng, "ICA with Reconstruction Cost for Efficient Overcomplete Feature Learning," *Advances in Neural Information Processing Systems,* vol. 24, pp. 1017-1025, 2011.

[59] Y. LeCun and Y. Bengio, "Convolutional Networks for Images, Speech, and Time Series," *The Handbook of Brain Theory and Neural Networks,* pp. 255-257, MIT Press, 1995.

Hoo-Chang Shin received the BS degree in computer science and mechanical engineering from Sogang University, Korea, in 2004 and the Diplom Ingenieur degree in electrical engineering from the Technical University of Munich, Germany, in 2008. He is working toward the PhD degree at the Institute of Cancer Research, University of London. His research interests include machine learning, computer vision, signal processing, and medical applications of artificial intelligence. He is a student member of the IEEE and the IEEE Computer Society.

Matthew R. Orton received the MEng and PhD degrees from the University of Cambridge, United Kingdom, in 1999 and 2004, respectively. He is a staff scientist at the Institute of Cancer Research, University of London, United Kingdom. His main research interests include modeling functional medical image data, and robust estimation methods for applying these models to clinical trials and clinical practice.

David J. Collins is a consultant clinical scientist employed at the Institute of Cancer Research and the Royal Marsden Foundation Trust. His interests include quantitative functional imaging methodologies employed in clinical trials. Current focus areas include whole body diffusion imaging and perfusion imaging using magnetic resonance. He is an author or coauthor of more than 140 papers in magnetic resonance.

Simon J. Doran received the PhD degree from the University of Cambridge in the area of quantitative imaging of NMR relaxation times. Following postdoctoral research on ultrafast MRI in Grenoble, France, he became a lecturer in MRI in the Physics Department at the University of Surrey, Guildford, United Kingdom, where he remains a visiting fellow and directs the experimental program in optical computed tomography. In his current post of senior staff scientist at the Institute of Cancer Research, Sutton, United Kingdom, he leads the development of advanced imaging informatics.

Martin O. Leach is codirector of the Cancer Research UK and EPSRC Cancer Imaging Centre, deputy head of the Division of Radiotherapy and Imaging and Professor of Physics as Applied to Medicine at the Institute of Cancer Research and the Royal Marsden Hospital, University of London. He joined the Institute of Cancer Research and the Royal Marsden in 1978 after PhD research in physics at the University of Birmingham. Since 1986, he has led a program of translational research developing and applying imaging methods to improved detection, diagnosis, and evaluation of cancer. He has more than 300 publications, and has been awarded the Barclay Medal of the *British Journal of Radiology* and the Silvanus Thompson Medal of the British Institute of Radiology. He is a fellow of the Academy of Medical Sciences, the Institute of Physics, the Institute of Physics and Engineering in Medicine, the Society of Biology, the ISMRM, and the ESMRMB. He is also a National Institute of Health Research senior investigator. He is also a senior investigator at the National Institute of Health Research. He is currently the chair of the ECMC Imaging Steering Committee and was previously chair of the British Chapter of the ISMRM.