Medical Image Analysis 17 (2013) 790–804

Contents lists available at SciVerse ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

Neighbourhood approximation using randomized forests

Ender Konukoglu[a,*], Ben Glocker[b], Darko Zikic[b], Antonio Criminisi[b]

[a] *Athinoula A. Martinos Center for Biomedical Imaging, Massachusetts General Hospital/Harvard Medical School, MA 02129, USA*
[b] *Microsoft Research, Cambridge CB1 2FB, UK*

ARTICLE INFO

Article history:
Received 20 December 2012
Received in revised form 23 April 2013
Accepted 24 April 2013
Available online 10 May 2013

Keywords:
Approximate nearest neighbours
Random decision forests
Patch-based segmentation
Image-based regression
Supervised neighbour search

ABSTRACT

Leveraging available annotated data is an essential component of many modern methods for medical image analysis. In particular, approaches making use of the "neighbourhood" structure between images for this purpose have shown significant potential. Such techniques achieve high accuracy in analysing an image by propagating information from its immediate "neighbours" within an annotated database. Despite their success in certain applications, wide use of these methods is limited due to the challenging task of determining the neighbours for an out-of-sample image. This task is either computationally expensive due to large database sizes and costly distance evaluations, or infeasible due to distance definitions over semantic information, such as ground truth annotations, which is not available for out-of-sample images.

This article introduces *Neighbourhood Approximation Forests* (NAFs), a supervised learning algorithm providing a general and efficient approach for the task of approximate nearest neighbour retrieval for arbitrary distances. Starting from an image training database and a user-defined distance between images, the algorithm learns to use appearance-based features to cluster images approximating the neighbourhood structured induced by the distance. NAF is able to efficiently infer nearest neighbours of an out-of-sample image, even when the original distance is based on semantic information. We perform experimental evaluation in two different scenarios: (i) age prediction from brain MRI and (ii) patch-based segmentation of unregistered, arbitrary field of view CT images. The results demonstrate the performance, computational benefits, and potential of NAF for different image analysis applications.

1. Introduction

Computational methods exploiting annotated datasets for image analysis can yield high accuracy and robustness. Amongst various ways of integrating existing data into analysis tasks, one class of methods has shown significant potential: *Neighbourhood-based Approaches* (NbAs) (Boiman et al., 2008). These intuitive approaches perform analysis on a new out-of-sample image by propagating information from annotated images which are most "similar" to the new one. The similarity is based on an application-specific distance, and the set of the most similar images is denoted as the set of *nearest neighbours* with respect to this distance. Since the construction of an explicit model over the annotated database is avoided, such approaches are also called non-parametric or exemplar-based. Unlike model-based or parametric approaches, which analyse new images using a model distribution, NbAs can leverage the full variability present in a database. Despite their simplicity and descriptive power, so far the application of NbAs has been limited due to lack of annotated data and issues related to general applicability and computational costs. Motivated by the rapid increase in publicly available databases, this article focuses on the above computational issues of neighbourhood-based methods, aiming to make this approach practical for a large class of applications in image analysis.

Neighbourhood-based approaches have been gaining momentum in the community over the last few years. Recent works apply variations of the general framework in different settings. Patch-based techniques (Coupé et al., 2011; Rousseau et al., 2011) and multi-atlas methods (Jia et al., 2012; Aljabar et al., 2009; Artaechevarria et al., 2009; Isgum et al., 2009; van Rikxoort et al., 2010; Rohlfing et al., 2003), use nearest neighbour search for segmenting medical images by leveraging large databases containing voxel-wise annotations. Nonlinear "manifold"-based methods, used in a large variety of tasks ranging from image registration to semantic classification (Hamm et al., 2010; Wolz et al., 2010; Atasoy et al., 2012; Lian and Davatzikos, 2011; Guerrero et al., 2011; Cho et al., 2011), apply the same general framework. Here, subsequent analysis based on prior non-linear dimensionality reduction relies on the neighbourhood structure present in the data. Indeed, the goal of most dimensionality reduction methods is to preserve precisely this neighbourhood-structure in the embedding.

[*] Corresponding author.
E-mail address: ender.konukoglu@live.com (E. Konukoglu).

1361-8415/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.media.2013.04.013

Despite the fact that methods utilising NbAs have demonstrated performance with state-of-the-art accuracy and robustness, these approaches suffer from certain limitations. The main challenge is the retrieval of nearest neighbours for an out-of-sample image with respect to an existing database. This difficulty manifests itself in two layers:

1. **Cost of Distance Computation.** Depending on the nature of the distance metric and the size of the database, determining the neighbours can be computationally very expensive. In the most naive approach, one has to compute the distance between the out-of-sample image and all the database images. The cost of this exhaustive search can become impractical, especially if the size of the database is large, or if the distance computation involves costly operations (e.g. deformable registration). This issue is also noted for "manifold"-based techniques, particularly for the task of computing the manifold coordinates of an out-of-sample image, as pointed out by Aljabar et al. (2012).
2. **Feasibility of Distance Computation.** In many scenarios the neighbourhood structure of interest is best defined via an application-specific distance which relies on additional meta-information, such as diagnostic labels, voxel-based annotations, or other non-image characteristics. Since such semantic information is usually not available for a new image (and inferring this information might be the actual aim of the analysis task), the distance between the new image and the database images cannot be computed. So, even in cases where database size allows exhaustive search, it seems impossible to determine the neighbours without the semantic information. To overcome this problem, it is common to replace the original distance with a surrogate which is based only on image intensity information at the expense of introducing heuristics and potential modelling inaccuracies.

This article presents a novel algorithm, *Neighbourhood Approximation Forests*, which addresses the challenges of nearest neighbour retrieval in the general NbA framework. NAF is a supervised learning method for approximating the neighbours of an out-of-sample image within a database, where the neighbourhood is induced by an *application-specific* distance of arbitrary definition. During the learning phase, NAF uses a *training database*, which consists of images and pairwise distances, to learn to use appearance-based features to cluster the images approximating the neighbourhood structure induced by the application-specific distance. This process starts from high-dimensional representation of images that consist of appearance-based features that are extracted without any prior assumptions on their individual relevancy or their compactness. From this high-dimensional space, the algorithm automatically determines the discriminative features along with associated sequences of tests, which yields a clustering of the training database that approximates the distance-induced neighbourhood structure. As a variant of random decision forests (Breiman, 2001; Amit and Geman, 1997), the algorithm encodes the sequences of tests in an ensemble of decision trees. For an out-of-sample test image, NAF predicts its closest neighbours within the database in an efficient manner by traversing the constructed trees and counting co-occurrences of the test image and all database images that reached the same leaves of the trees. This prediction procedure does not require the evaluation of the distance since the trees only contain tests on features. Therefore, NAF addresses both the cost and the feasibility issues in the distance computation.

While a number of approaches for nearest neighbour retrieval has been proposed in the literature, to the best of our knowledge, NAF is unique in its built-in feature selection property and its generality. It leverages the automatic feature selection strategies of decision trees for approximating neighbourhood structures. As a result, NAF is able to handle truly high-dimensional representations of images, e.g. 10,000 features, in its learning phase. This capability is especially important in the context of medical image analysis, where determining the distance-specific features themselves is of interest to understand links between image-based markers and diseases (Sabuncu and Van Leemput, 2011; Gaonkar and Davatzikos, 2012). Additionally, through a generic formulation, NAF is able to work on completely arbitrary distance definitions, whether based on semantic or image-based information. Therefore, NAF can be directly applied to a wide variety of applications without modifying the core algorithm itself. In the following, we highlight the relation of NAF to previous work in more detail.

1.1. Related work

Neighbourhood-based approaches in medical image analysis mostly rely on appearance-based distances only (Allassonnière et al., 2007; Aljabar et al., 2009; Sabuncu et al., 2009; Coupé et al., 2011; Rousseau et al., 2011). Therefore, the focus of these methods is on reducing the costs of distance computation. In contrast, NAF is able to incorporate distances that are defined on semantic information and relate them to the space of appearance-based features. The computational efficiency of NAF, on the other hand, is directly inherited from the properties of decision trees.

Acknowledging the impracticality of exhaustive search, approaches from the field of medical image analysis use either heuristics for efficient search as in Aljabar et al. (2009), Coupé et al. (2011), and Rousseau et al. (2011) or unsupervised k-means clustering for multi-template constructions (Sabuncu et al., 2009; Allassonnière et al., 2007). Heuristic strategies are a set of application-specific ad hoc rules for speeding up the neighbourhood search. Therefore, they have to be explicitly designed for each application, which reduces their flexibility. In contrast, NAF is a learning technique that *automatically* adapts to the given application in a supervised manner. Strategies based on k-means achieve efficiency by aiming to determine representative centroids for the database reflecting different data clusters. The neighbours of an out-of-sample image are then obtained from the set of centroids. The difficulty here is the optimal choice of the number of centroids. Too few centroids will not correctly reflect the neighbourhood structure, and too many centroids increase computational time but also might yield an overfitted representation. Also, the centroids computed in an unsupervised way might not reflect the same neighbourhood structure which is induced by the original distance. In contrast, NAF takes into account each individual image in the database for predicting the neighbours, remains efficient thanks to its inherited properties, and uses supervision for clustering the data points.

In the area of machine learning there are numerous works which tackle the problem of (approximate) nearest neighbour search for an out-of-sample image. The first group of such techniques are *supervised metric learning methods*, which aim to approximate an arbitrary distance via a surrogate metric learned from compact representations of images. The learned metric should preserve the structure induced by the original distance, (Xing et al., 2002; Goldberger et al., 2005; Globerson and Roweis, 2006; Weinberger and Saul, 2009; Chechik et al., 2010; Der and Saul, 2012). Given an initial set of image descriptors, these methods rely on dimensionality reductions (such as PCA) and the new metric is defined in the lower dimensional space. However, these methods rely on the assumption that there is a distance-preserving transformation from high to low dimensional feature spaces. Finding appropriate transformations is a non-trivial task. Furthermore, the features in the low-dimensional space are not easily interpretable as they represent a complex linear or non-linear combination of the higher dimensions. In contrast to this, NAF is capable of handling directly high-dimensional input feature spaces removing the need of feature pre-selection or dimensionality reduction. The learned, discriminative features remain interpretable and, for instance, allow to generate feature relevance maps.

The second group of machine learning techniques for approximate nearest neighbours are *hashing methods*. These techniques also focus on finding compact representations of images, particularly in form of binary codes. Such codes allow efficient retrieval of neighbours from possibly very large databases, (Nister and Stewenius, 2006; Weiss et al., 2009; Muja and Lowe, 2009; Norouzi and Fleet, 2011). Similar to supervised metric learning these works also assume an appropriate mapping from high to low-dimensional features. However, finding appropriate mappings which are capable of representing the neighbourhood still remains a challenging problem.

The last class of approaches aims to find the manifold coordinates of a new image without reconstructing the embedding using non-linear dimensionality reduction techniques (Bengio et al., 2004; He and Niyogi, 2004). These methods focus on the dimensionality reduction part only and rely on the assumption that the distance between the new image and all other images in the database can be computed efficiently. Therefore, these methods do not focus on the computational issues which are addressed by NAF.

1.2. Article overview

The remainder of the article is structured as follows. Section 2 describes the method in detail. The properties and potential of NAF are demonstrated by experiments on two different applications in Section 3. We show an application for image-based age regression from brain MRI, and patch-based segmentation of varying field-of-view CT scans. Following the experimental evaluation, we conclude the article with a discussion and an outlook of further possible applications of NAF. Please note that the presented article is an extended version of the conference proceeding (Konukoglu et al., 2012), which was presented at MICCAI 2012. This article differs significantly in the description of the method as well as the relation to previous work, the experimental setup, analysis and results.

2. Neighbourhood approximation forests

The general framework of NbA leverages annotated data for various image analysis tasks. The underlying principle is to analyse an image using other "similar" images, for which prior information is available. The nature of the prior information can vary from semantic annotations to outputs of previously performed computationally expensive calculations.

In the most abstract sense, NbAs formulate the set of all images as a high-dimensional space $\mathcal{I}$, in which individual images are represented as points $I \in \mathcal{I}$. The set of images with available prior information is a finite subset within this space $\mathbf{I} = \{I_p\}_{p=1}^{P} \in \mathcal{I}$ and it is often called the *training dataset* or simply the database. The space $\mathcal{I}$ is equipped with a distance $\rho(I, J) : \mathcal{I} \times \mathcal{I} \to \mathbb{R}$ that quantifies a similarity between images. The definition of $\rho(\cdot,\cdot)$ is application-specific and not necessarily appearance-based. For instance, $\rho(\cdot,\cdot)$ may depend on certain meta information according to the goal of the application.

For analysing a new image J, for which prior information is not available, i.e. an out-of-sample image $J \notin \mathbf{I}$, neighbourhood-based approaches focus on the set of k images from $\mathbf{I}$, which are most similar to J. More precisely, they define the set of k images in $\mathbf{I}$ with the lowest ρ distance to J as the k nearest neighbour (kNN) images. Here we denote this set by $\mathbf{N}_{\rho}^{k}(J)$. After determining the kNNs, the NbA framework performs the subsequent analysis by propagating information from $\mathbf{N}_{\rho}^{k}(J)$ to J.

Finding $\mathbf{N}_{\rho}^{k}(J)$ within $\mathbf{I}$ is the essential step in NbA. As discussed in the introduction, this step is very challenging because the computation of the distance $\rho(\cdot,\cdot)$ can be infeasible or expensive, depending on the nature of the distance. In the following, we describe a learning algorithm to approximate $\mathbf{N}_{\rho}^{k}(J)$ that overcomes these challenges.

Neighbourhood Approximation Forests (NAFs) relies on the hypothesis that the neighbourhood structure constructed by $\rho(\cdot,\cdot)$ can be approximated using the information contained in the images. Based on this assumption, the algorithm learns to approximate the neighbourhood $\mathbf{N}_{\rho}^{k}(J)$ of an out-of-sample image J within $\mathbf{I}$ by using only image-based features, without the need to actually evaluate $\rho(\cdot,\cdot)$. Naturally, the set of features that are used for such an approximation depends on the distance $\rho(\cdot,\cdot)$. As a supervised learning method, NAF uses a training set to automatically discover distance-specific image-based features in an initially high-dimensional feature space.

NAF is a variant of random decision forests (Breiman, 2001; Criminisi et al., 2011), i.e. an ensemble of binary decision trees, where each tree is an independently learned predictor of neighbourhood. Being a supervised learning method, it has two phases: training (construction of trees) and prediction. The rest of this section describes these phases in detail starting from the prediction procedure.

2.1. Tree testing: predicting neighbourhood with a single tree

The function of a tree in NAF is to predict the nearest neighbours of an image J with respect to $\rho(\cdot,\cdot)$ using a high-dimensional representation of the image $\mathbf{f}(I) \in \mathbb{R}^{Q}$, which is composed of intensity-based features. Similar to conventional binary decision trees, (Amit and Geman, 1997), a NAF tree T is a graphical representation of a set of hierarchically nested binary tests. This representation is composed of nodes and directed edges as illustrated with an example in Fig. 1.

Given an image J, the prediction procedure starts at the root node of the tree, s_0. The binary test associated with s_0 is applied to $\mathbf{f}(J)$ and based on the result J is sent to either one of its two child nodes, i.e. the two nodes connected to s_0 via outgoing edges. At this new node, say s_1, the next associated test is applied to $\mathbf{f}(J)$ and J is sent further to one of s_1's children. This process is repeated for every new node until J reaches a node with no further outgoing edges, a *leaf node*. The path that image J traverses until it reaches a leaf node depends on the feature vector $\mathbf{f}(J)$ and the binary tests associated with each node in the tree.

At each leaf node the algorithm stores training images $I_n \in \mathbf{I}$ (represented by their indices) which have arrived at that particular node following the same procedure as a test image. The training images contained in the leaf node reached by J, have taken the same path as J, and therefore have features which are similar to a certain degree to those of J, according to the applied tests. This subset of training images, denoted as $\mathbf{N}_{T(\rho)}(J)$, is the prediction of tree T for the neighbourhood of J. The subscript $T(\rho)$ denotes the tree's dependence on $\rho(\cdot,\cdot)$, which we will explain in detail in Section 2.2.

Fig. 1 illustrates the testing procedure for a toy example where the similarity between images is defined by the proximity of the colours of their bounding boxes, e.g. yellow is more similar to orange than to red. As described above, the test image starts at the root node, follows a path based on the binary tests and ends up at a leaf node occupied by yellow images, which are the predicted neighbours of the test image for this tree.

The neighbours predicted by a single tree depend on the binary tests stored in the tree nodes. The next section describes how these

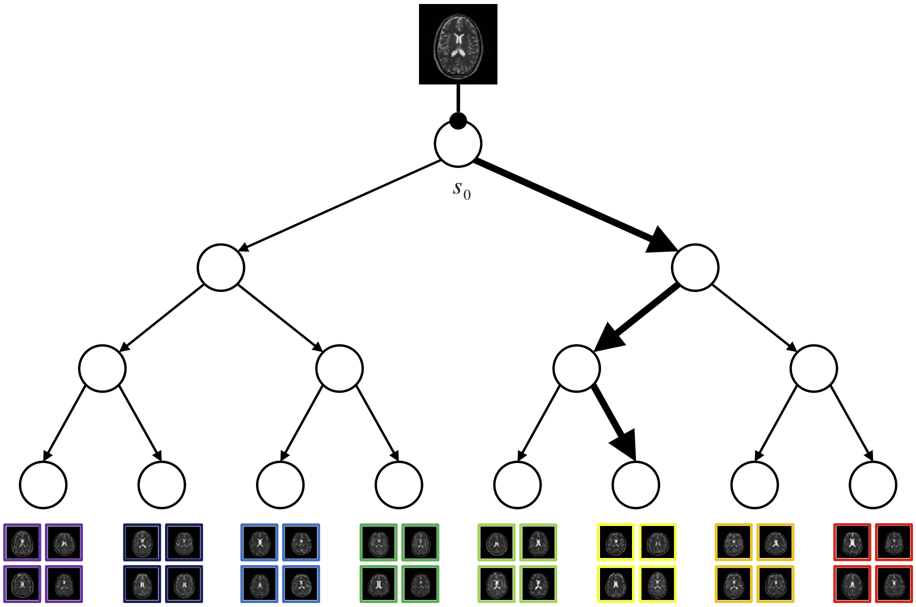


Fig. 1. Toy example illustrating the NAF testing procedure. For the out-of-sample image *J* (top), an analysis task is to be performed, based on similar images in the database (bottom). The similarity between the images is based on some meta information, represented by colour. This information is available for the database images, but not for the out-of-sample image. The trained NAF tree determines the neighbourhood for *J*, by hierarchically performing tests at each node, based on intensity-based feature description of *J*. The parameters for the tests are learned in the NAF training step (Section 2.2, Fig. 2) in a supervised manner, such that the approximated neighbourhood reflects the original similarity. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

tests are learned and how each tree is constructed using the training database.

2.2. Tree training

The goal of the training phase is to discover a small set of image features and associated binary tests that can be used to approximate the neighbourhood structure defined by $\rho(\cdot,\cdot)$ using the procedure described in Section 2.1. NAF achieves this in a supervised manner using a training set composed of an image set $\mathbf{I}$ and pairwise distances $\rho(I_n, I_m)$, $\forall I_n,\ I_m \in \mathbf{I}$. Furthermore, the training set also contains a high-dimensional representation of each image $\mathbf{f}(I)$. Unlike previous approaches, NAF does not assume that $\mathbf{f}(I)$ is a compact descriptor. $\mathbf{f}(I)$ can be of high dimensions, e.g. Q is higher than 10^4 in the experiments in Section 3, and may contain features that are both relevant and non-relevant for approximating $\rho(\cdot,\cdot)$. In fact the role of training is to discover the relevant features within $\mathbf{f}(I)$.

The training phase constructs a tree starting with its root node and progressively growing it deeper by adding new nodes. Tree construction corresponds to finding the set of nested binary tests on image features that sequentially partition $\mathbf{I}$ into spatially consistent and compact subsets with respect to $\rho(\cdot,\cdot)$. Assuming $\mathbf{I}$ is a representative database, the learned binary tests can be successfully applied to other, unseen images. Let us describe this procedure in detail.

The construction of a tree T starts by randomly choosing a subset of the entire feature vector $\mathbf{f}_T(I) \in \mathbb{R}^q$, $q < Q$, $\mathbf{f}_T(I) \subset \mathbf{f}(I)$. By injecting such randomness each tree works on a different subspace, which yields decorrelated trees. This allows individual trees to produce independent predictions and is known to improve generalisation as demonstrated in previous works (Amit and Geman, 1997; Breiman, 2001).

Using the randomly chosen feature set $\mathbf{f}_T(I)$ NAF performs tree construction by recursively determining the *optimal branching* at each node and for splitting the incoming data into two sets. Assume the algorithm is at a given node s and the set of training images at this node is $\mathbf{I}_s$. Branching of s and the partitioning of $\mathbf{I}_s$ into two sets is defined via a binary test

$$t_s(I_n; m, \tau) \triangleq \begin{cases} I_n \in \mathbf{I}_{s_R}, & \text{if } f_T^m(I_n) > \tau, \\ I_n \in \mathbf{I}_{s_L}, & \text{if } f_T^m(I_n) \leqslant \tau, \end{cases} \forall I_n \in \mathbf{I}_s, \tag{1}$$

where f_T^m denotes the mth component of $\mathbf{f}_T(I_n)$, $\tau \in \mathbb{R}$, and s_L and s_R are the children of s. Optimising this branching aims to estimate the parameters m and τ that yield the most compact partitioning of $\mathbf{I_s}$. To this end, we define the spatial compactness of a set $\mathbf{A}$ with respect to $\rho(\cdot,\cdot)$ as the *Cluster Size*:

$$C_\rho(\mathbf{A}) \triangleq \frac{1}{|\mathbf{A}|^2} \sum_{I_i \in \mathbf{A}} \sum_{I_j \in \mathbf{A}} \rho(I_i, I_j), \tag{2}$$

where $|\mathbf{A}|$ denotes the size of the set. Based on $C_\rho(\cdot)$, the gain in compactness for a specific set of parameters is determined by

$$G(\mathbf{I}_s, m, \tau) \triangleq C_\rho(\mathbf{I}_s) - \frac{|\mathbf{I}_{s_R}|}{|\mathbf{I}_s|} C_\rho(\mathbf{I}_{s_R}) - \frac{|\mathbf{I}_{s_L}|}{|\mathbf{I}_s|} C_\rho(\mathbf{I}_{s_L}), \tag{3}$$

where the weights $|\mathbf{I}_{s_R}|/|\mathbf{I}_s|$ and $|\mathbf{I}_{s_L}|/|\mathbf{I}_s|$ avoid bias towards unbalanced partitions. Based on $G(\mathbf{I}_s, m, \tau)$, the best possible binary test at node s is determined by solving the following optimisation problem

$$(m_s, \tau_s) = \arg_{m,\tau} \max\ G(\mathbf{I}_s, m, \tau) \text{ such that } |\mathbf{I}_{s_R}| \geqslant \Delta,\ |\mathbf{I}_{s_L}| \geqslant \Delta, \tag{4}$$

where Δ is a parameter denoting the minimum allowed number of samples in a node. The value of Δ prevents us from obtaining nodes where meaningful statistics cannot be drawn due to small sample size. This optimisation problem can be computationally expensive if one considers all m, especially if the feature vector $\mathbf{f}_T(\cdot)$ is of high dimension. The proposed algorithm does not take into account all m but chooses a small random subset of the components of $\mathbf{f}_T(\cdot)$ at each node. This corresponds to the common approach of randomised greedy optimisation in decision forests (Criminisi et al., 2011). For each value of m we exhaustively optimise over τ thresholds. Training is initialised by setting $\mathbf{I}_{s_0} = \mathbf{I}$. These steps are repeated for each new node in a recursive manner.

Tree growing is stopped at a node when (i) the algorithm can no longer find a test that creates a more compact partitioning, i.e. $\forall (m, \tau)$, $G < 0$, (ii) the number of training images within the node falls below a threshold, i.e. $|\mathbf{I}_s| < \Delta$ or (iii) the maximum allowed depth is reached (to limit memory requirements and computational cost of testing).

We illustrate the training process for a toy example in Fig. 2. Tree construction starts at the root node with the entire training database, where similarity between images is defined as the proximity of the colours of the bounding boxes for each image. At each node the training set is partitioned into smaller, more compact sets. Sequentially applying the node optimisation the algorithm produces leaf nodes where tree growing is stopped. Here, the remaining images in each leaf node are very similar to each other and no further partitioning is possible. For illustration purposes we assume that there are features that can perfectly partition the data and the optimisation at each node is able to determine these features.

In the training phase multiple trees are constructed independently and due to the randomness in construction they are all different. The final component of the NAF algorithm is the integration of individual tree predictions into a forest prediction for the k nearest neighbours for an out-of-sample image J. We describe this step in the following section.

2.3. NAF: predicting neighbourhood with the forest

To predict the closest neighbours of an image J, NAF combines the individual tree predictions and computes the affinity of J to each database image I_n by

$$\mathbf{w}_F(J, I_n) \triangleq \sum_{\forall T \in F} \mathbf{1}_{\mathbf{N}_{T(\rho)}(J)}(I_n), \tag{5}$$

where $\mathbf{1}_{\mathbf{A}}(x)$ is the indicator function, i.e. $\mathbf{1}_{\mathbf{A}}(x) = 1$ if $x \in \mathbf{A}$, and 0 if $x \notin \mathbf{A}$. $\mathbf{w}_F(J,I_n)$ simply counts the number of trees that predict I_n as the nearest neighbour of J. A high value of $\mathbf{w}_F(J,I_n)$ indicates that images J and I_n share many similar features based on the binary tests which were optimised with respect to $\rho(\cdot,\cdot)$. Therefore, it is reasonable to assume that I_n is within close vicinity of J. We would like to note that in Gray et al. (2011), Criminisi et al. (2011), and Fu et al. (2012) similar integration strategies have been used. However, in Gray et al. (2011) and Criminisi et al. (2011) the purpose is to *define* a neighbourhood structure using the forest rather than *approximating* a given one as proposed here. While in Fu et al. (2012) the affinities for an image are only considered in the specific setting of image annotation to improve the information propagation in the leaf node for this application. Compared to these works, NAF is a more general approach that can handle arbitrary distances and neighbourhoods. Thus, the above mentioned works can be seen as special instantiations within our framework.

Once the affinities $\mathbf{w}_F(J,I_n)$ are calculated for all $I_n \in \mathbf{I}$, the NAF computes the forest prediction of $\mathbf{N}^k_\rho(J)$ simply by determining the set of k training images with the largest $\mathbf{w}_F(J,I_n)$ values. We denote this set as $\mathbf{N}^k_{F(\rho)}(J)$. The subscript indicates dependency of the forest on $\rho(\cdot,\cdot)$.

2.4. Design parameters

NAF contains several design parameters that should be set based on the application. This section provides a list of these parameters and provides intuitive explanation for each one of them. A quantitative analysis of the most important parameters is provided using the age-regression experiment presented in Section 3.1.

2.4.1. Minimum samples per node Δ

The minimum sample size Δ is an important parameter that influences the training phase of the algorithm. It plays a role in the tree construction both by constraining the node optimisation in Eq. (4) and by setting a stopping criteria for the tree growth. The Δ parameter ensures that each leaf node can produce a statistically meaningful prediction. In addition, Δ is also related to overfitting of the learning algorithm. Setting Δ too low would force a tree to learn very fine details of the training set leading to overfitting. As a result the trained forest might not generalise to unseen images. On the other extreme, setting Δ too high would yield limit

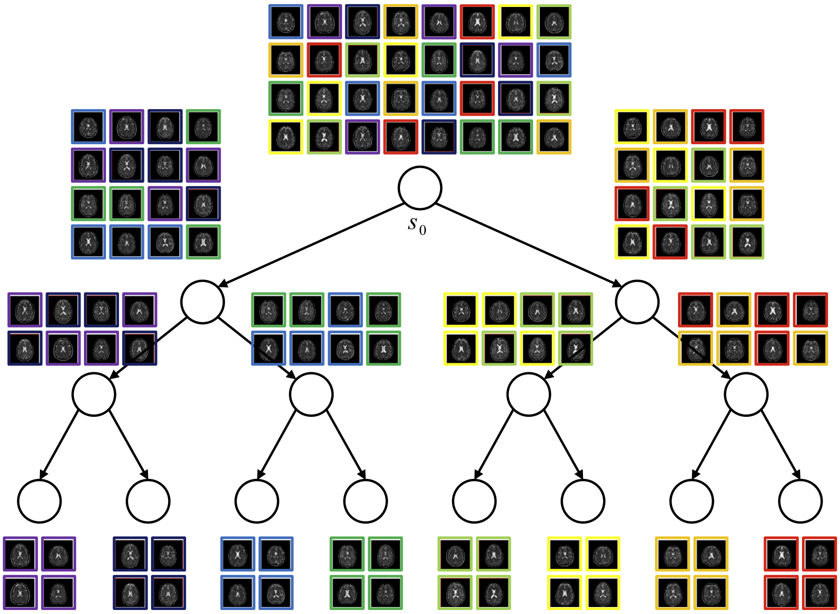


Fig. 2. Toy example illustrating the NAF training procedure. The similarity between images is defined as the proximity of the bounding box colours. The tree construction starts at the root node with the entire training database, and at each node the training set is partitioned into smaller sets. The partitioning is performed by optimising the parameters of the test function at each node according to Eq. (4), such that the meta-information-based similarity is captured by the intensity-based features.

the predictive power of the algorithm. In Section 3.1 we analyse the effect of this parameter for a neighbourhood-based regression problem.

2.4.2. Maximum tree depth

In conventional methods using decision trees, the maximum tree depth is an important parameter related to overfitting, see (Criminisi et al., 2011). In NAF its effect is dampened by the Δ parameter. If the maximum depth is set very high, tree growing would be stopped due to the Δ parameter before the maximum depth is reached. In contrast, a too low value for the maximum depth may result in leaf nodes which could have been split further into more compact partitions. Therefore a too low maximum depth may again limit the predictive power of a tree. Besides its influence on the approximation accuracy, the maximum tree depth also influences the memory requirements and the computational cost of prediction. A deeper tree requires more memory and will take longer to traverse.

2.4.3. Number of trees

Each tree works on a randomized subspace of the entire feature representation of an image. Increasing the number of trees almost always results in an increased accuracy since overall predictions are made with more information. This increase often levels off after a certain number of trees. The leveling off becomes important when one considers the fact that the number of trees affects the computational complexity of the algorithm both during training and testing. The point where the accuracy levels off, however, is application dependent and should be analysed. Section 3.1 provides such an analysis for the MRI based age-regression problem.

2.4.4. Feature space size Q and restriction q

In general, the discriminative set of features which will be relevant for approximating the neighbourhood structure induced by an arbitrary $\rho(\cdot,\cdot)$ is not known a priori. For certain image classification tasks and retrieval problems different authors rely on SIFT features or even handcrafted features. However, it is desirable to automatically detect the relevant features by learning them with respect to the training data and the distance. This is precisely what the NAF training aims for. To this end, the parameter Q is set fairly high trying to include as many features as possible into the training.

The q parameter influences the learning procedure in two ways. First, the ratio q/Q has an impact on the similarity between different trees in the forest. A value close to 1 yields trees which are more correlated as they are largely using similar regions of the feature space. In contrast, a low q/Q ratio will result in more decorrelated trees which is a desired property. Second, the absolute value of q relates to how much information each tree is exposed to and also impacts the computational complexity for the construction (due to Eq. (4)). A low q value may result in non-descriptive trees, while a high q value would result in descriptive trees but with high computational costs at training time. In summary the value of q should be selected considering Q and the computational budget. For instance, for the age-regression experiment presented in the next section the q/Q ratio is set to 0.1 while Q is set to relatively high values around 10^4. For the patch-based segmentation experiment on the other hand, q/Q is set to 1/3 and Q to 1500. Furthermore, in both experiments small variations around this value have not influenced the results.

3. Experiments

In the following, we demonstrate the properties and performance of NAF with experiments on two different applications: image-based age regression from MRI brain scans, and patch-based segmentation of varying field-of-view unregistered CT scans. These two applications have different goals, work with different imaging modalities, and use different definitions of image distances, $\rho(\cdot,\cdot)$. In both cases, we construct simple analysis methods based on the neighbourhood-based approach utilising NAF, and evaluate the resulting methods. Our evaluation focuses on quantifying the added benefits of NAF on the final analysis result: regression and segmentation accuracies. Particularly, we compare results obtained using NAF and obtained using the conventional appearance-based (unsupervised) clustering methods for nearest neighbour retrieval. Quantitative and qualitative results confirm the benefits of incorporating NAF for nearest neighbour search.

For each application, we first discuss the input data which consists of the image database $\mathbf{I}$, the distance measure $\rho(\cdot,\cdot)$, and the feature space $\mathbf{f}(\cdot)$, followed by explaining parameter settings, evaluation and results.

3.1. Age regression from brain MRI

In this first application the goal is to predict the age of a subject from its brain MRI, and a database of images for which the ages of the scanned subjects are known. The algorithm we devise here is to determine the closest neighbours of an image within the database using NAF, and propagate the age information from these images. This application demonstrates the use of NAF in a setting where the neighbourhood is based on meta information, i.e. age. In this context, we show the difference between NAF and appearance-based clustering in terms of regression accuracy. We further analyse why NAF achieves better accuracy by examining the set of features that are selected during training. The analysis shows that NAF automatically discovers features that are in accordance with the known physiological effects of ageing.

Additionally, we present three further analyses: (i) parameter analysis quantifying the effect of design parameters on the final prediction accuracy, (ii) comparison of the neighbourhoods estimated by NAF and appearance-based clustering with the underlying "true" neighbourhood (that could have been only determined if the semantic information was available at the test time), and (iii) quantitative and qualitative comparison of NAF with supervised-metric learning methods.

3.1.1. Dataset

We use 414 T1-weighted brain MR images from the publicly available OASIS database, which contains the age information for each subject (Marcus et al., 2007). Images are skull stripped, histogram equalised, and aligned to a common reference frame via affine registration. For computational efficiency, we use downsampled images at an isotropic voxel resolution of 2 mm.

3.1.2. Application-specific settings

Distance $\rho(\cdot,\cdot)$: For the aim of age regression, it seems reasonable to define similarity between two subjects in terms of their difference in age. Thus, for this application the pairwise distance defining the neighbourhood structure is simply $\rho(I,J) \triangleq |\text{age}(I) - \text{age}(J)|$.

Database $\mathbf{I}$: For the age regression experiment we perform leave-one-out tests, i.e. the set $\mathbf{I}$ consists of 413 images for each test. We train NAF on these images based on the age-related pairwise distance $\rho(I_n,I_m)$.

Feature space $\mathbf{f}(\cdot)$: The feature vector for each image consists of intensity values at Q = 10,000 randomly chosen voxels defined in the common reference frame. These intensity values are extracted from smoothed version of the images. We apply a mean filter with a physical size of 12 mm in each dimension (leading to a structural element of size 6 × 6 × 6 voxels) to obtain the smooth versions of the images.

NAF parameters: For each test we train NAF using the corresponding $\mathbf{I}$, $\mathbf{f}(\cdot)$ and $\rho(I_n, I_m)$, $\forall I_n, I_m \in \mathbf{I}$. We empirically determined the parameters for the training phase as $q = 10^3$, F contains 300 trees, $\Delta = 15$ and the maximum depth for a tree is 15, which in the experiment was not reached by any tree. Here, these parameters were not optimised. One could, however, consider using a validation set to find the optimum set of parameters. In order to provide an insight on these parameters, further analysis on the influence of Δ and number of trees is presented in Section 3.1.5 in more detail.

Regression details: The regression of the subject age starts by extracting the features from a test brain MR scan J. These features are used in the trained NAF to determine the k closest neighbours of the image within the database as well as the affinity values $\mathbf{w}(J, \cdot)$ for these training images (see Eq. (5)). Based on these the algorithm predicts the age using the weighted mean as follows:

$$\hat{\text{age}}(J) \triangleq \frac{\sum_{I_n \in \mathbf{N}^k_{F(\rho)}} \mathbf{w}(J, I_n)\ \text{age}(J)}{\sum_{I_n \in \mathbf{N}^k_{F(\rho)}} \mathbf{w}(J, I_n)}. \tag{6}$$

3.1.3. Evaluation and results

We evaluate the performance of NAF by comparing the real age of the test subject with the predicted one. Fig. 3a shows plots with the predicted age vs. real age for all 414 subjects (result of leave-one-out tests) for four different values of k = 1, 7, 15, 413. On each plot we report the r-value of the prediction as well as the root mean square error (RMS). The results are very similar to those obtained with relevance vector machines (RVM) (Franke et al., 2010) and relevance voxel machines (RvoxM) (Sabuncu and Van Leemput, 2011). Both have been applied to a slightly smaller subset of the same database in Sabuncu and Van Leemput (2011). RVM achieves an RMS error of 10 years and the two variations of RvoxM achieve RMS errors of 9 and 8 years.

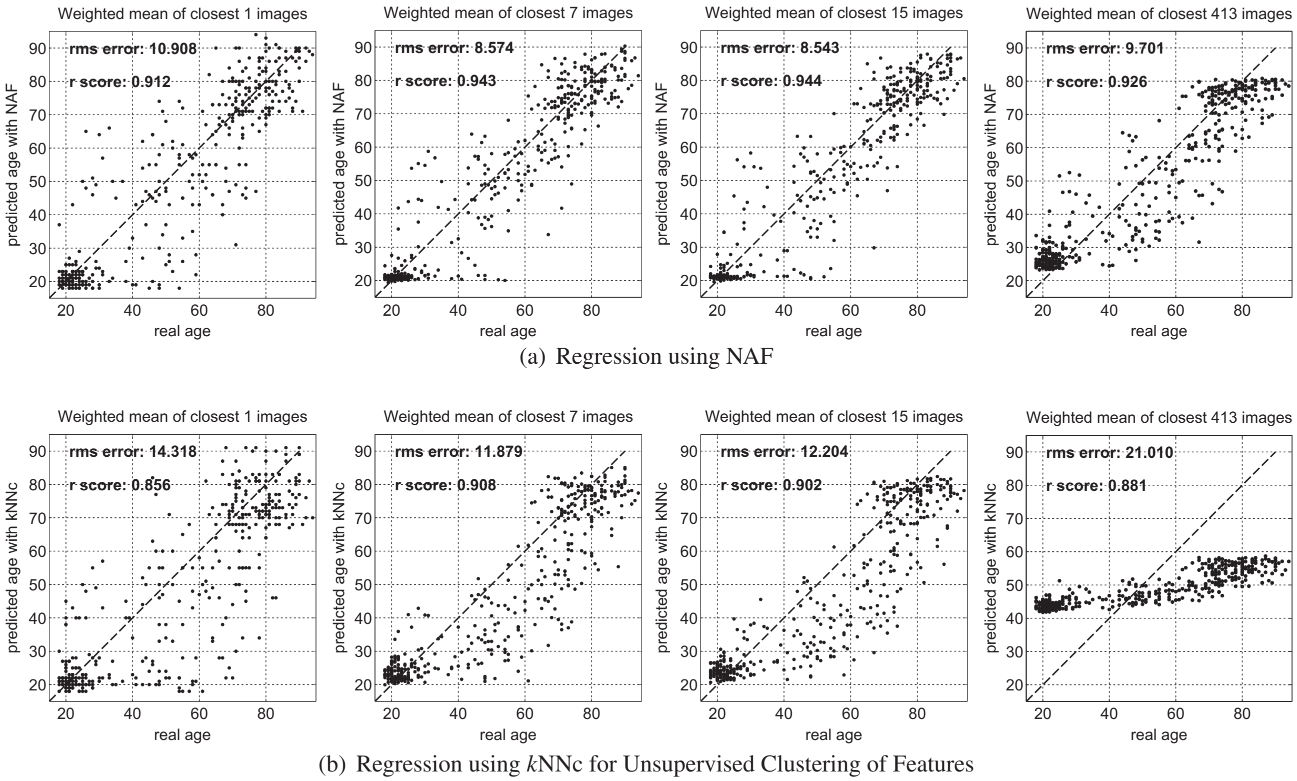
In terms of computation times, the regression for a test subject, for the leave-one-out tests, took on the average 2.8 s on an Intel Xeon processor running Matlab for 64-bits. The training for each leave-one-out test took 43.9 s.

In order to understand the added benefits of our supervised approach over unsupervised clustering, which is the current choice of researchers using NbA, we also perform age regression using an appearance-based k-nearest neighbour clustering method (kNNc) and compare its performance. Here, the neighbourhood is defined with respect to L_2 distances between the entire feature vectors $\mathbf{f}(I)$, i.e. $\rho_{k\text{NNc}}(I_n, I_m) \triangleq \|\mathbf{f}(I_n) - \mathbf{f}(I_m)\|_2$. In this case for each test image we exhaustively search for the k nearest neighbours of a test image J. The regression method is again defined as a weighted mean by

$$\tilde{\text{age}}(J) \triangleq \frac{\sum_{I_n \in \mathbf{N}_{k\text{NNc}}} w(J, I_n)\ \text{age}(J)}{\sum_{I_n \in \mathbf{N}_{k\text{NNc}}} w(J, I_n)} \tag{7}$$

$$w(J, I_n) \triangleq \frac{1}{\rho_{k\text{NNc}}(J, I_n)}. \tag{8}$$

Fig. 3b plots the regression results obtained using the kNNc approach. The results obtained by NAF show higher accuracy for all choices of k. Particularly, the difference at k = 413 when all the training images are used for the regression is striking. The regression method using NAF performs much better than the one using kNNc simply because it is able to use a neighbourhood definition devised specifically for the task of age regression. NAF learns which features to use to approximate this neighbourhood and allows nearest neighbour retrieval specific to age regression. In contrast, the method using kNNc simply assumes feature distances are correlated to age differences without any learning involved. Without



(a) Regression using NAF

(b) Regression using kNNc for Unsupervised Clustering of Features

Fig. 3. (a) Regression results obtained using the closest neighbours approximated by NAF: Graphs plot predicted age vs. real age for each image in the database computed using leave-one-out style. From left to right the predictions are obtained using k = 1, 7, 15, 413 closest neighbours respectively. In each plot the r value and the RMS error for the regression is provided. (b) Same plots as above displaying regression results obtained using the closest neighbours predicted using feature clustering.

knowing which features are relevant for age regression, the images *k*NNc brings as nearest neighbours are similar in overall appearance but not necessarily in age.

We would also like to provide some examples where the regression approach using NAF fails. Fig. 4 displays some images where the real age and the predicted age differs (left in each row) along with examples from the same age groups where predicted and real ages agree (middle and right). We note that in these failure cases the images have characteristics of a different age group compared to the actual age of the subject.

3.1.4. Relevant feature maps

In the previous part we noted that NAF is able to achieve better performance than an unsupervised *k*NNc due to the supervised feature selection. Here, we examine this result further. Specifically, we determine the features NAF selects during training as the most relevant ones for predicting age. Fig. 5a shows the frequency of the selected features in the first three levels of the trees throughout the forest. The values are normalised with respect to the total number of nodes in these levels. Furthermore, for visualisation purposes the raw frequency maps are smoothed with a Gaussian kernel of standard deviation 2.0 mm. We observe that NAF uses features extracted around the ventricles, prefrontal cortex, brainstem and hippocampal regions. All these structures are known to show substantial change with ageing (Luft et al., 1999; Gunning-Dixon et al., 2008).

For comparison we also examine the features that are selected if we define a pairwise image distance as the L_2 norm of the feature difference between images, i.e. $\rho(I_n,I_m) = \|\mathbf{f}(I_n) - \mathbf{f}(I_m)\|_2$. We train a NAF using this image distance and determine the most relevant features. We note that this corresponds to unsupervised clustering with respect to the entire feature vector, which kd-tree type algorithms would achieve (Arya et al., 1998; Muja and Lowe, 2009). Fig. 5b shows the same frequencies as (a) for this modified approach. We see that almost all selected features are around the ventricles where the largest intensity variation is present. This map is substantially different than what NAF selects when age is used for supervision through the distance definition. The quantitative results of the previous section reflect the effect of this difference in selected features on regression accuracies.

3.1.5. Parameter analysis

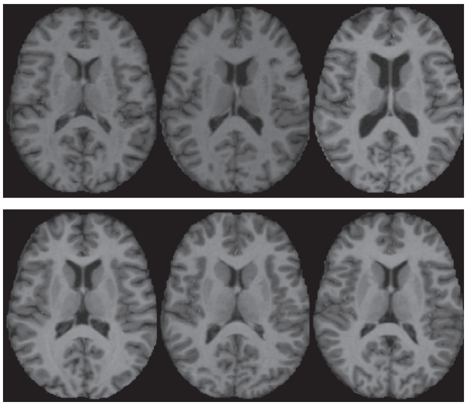
In this part, we analyse the effects of the most influential parameters of NAF on the regression accuracy. For computational reasons, we perform leave-18-out tests in which case $\mathbf{I}$ consists of 396 images. Fig. 6a plots the RMS error for age regression vs. $\varDelta$ while the other parameters are constant. The plots are given for different neighbourhood sizes *k*. We observe that for low values of $\varDelta$ the RMS error is rather high due to overfitting. For high values of $\varDelta$ we also observe an increase in RMS error due to decreasing discrimination power of NAF because of insufficient learning. Lastly we observe that there is a large parameter regime between $\varDelta = 15$ and $\varDelta = 30$ where the accuracy of the regression method is stable.

Fig. 6b plots the RMS error vs. number of trees. We observe that the increase in the number of trees is beneficial until 100 trees and beyond the additional increase in accuracy is rather small. In these plots we also observe the effect of *k* on the regression accuracy as well as the stability of the system with respect to parameters. We observe that with high *k* values the RMS errors are lower and the system is much more stable with respect to changing parameters.

3.1.6. Comparison with the true neighbourhood

In the previous parts we evaluated the performance of NAF through regression accuracies. Here, we compare the neighbours estimated by NAF to the neighbours computed using the original distance ρ: the *true* neighbours. In reality, the computation of the true neighbours is not possible in the test time because the meta-information is not available. However, to provide a comparison we assume that the meta-information is available at the test time. We then compute the true neighbours of a test image within the training database, denoted as $\mathbf{N}_{\text{real}}{}^{k}$, and compare this set with the ones estimated by NAF, $\mathbf{N}_{F(\rho)}$ and by unsupervised *k*NNc, $\mathbf{N}_{k\text{NNc}}$. This comparison blends two different aspects. First, it measures how well image information is able to capture the true neighbourhood defined by a non-image based information. Second, it measures how a clustering algorithm performs to estimate the neighbours. Unfortunately, there is no trivial way to separate these two aspects from each other. Nevertheless, the comparisons provide a feeling on how close the neighbourhood estimates get to the true one.

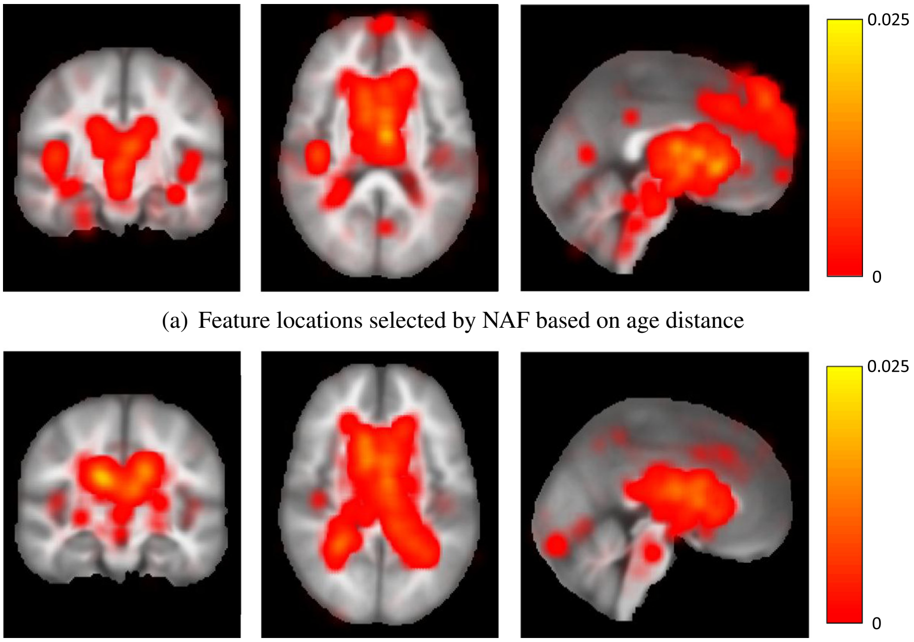
In Fig. 7a we plot the normalised size of the set of intersections between the true neighbourhood and the estimated ones for different *k*, i.e. $|\mathbf{N}^k_{\text{real}} \cap \mathbf{N}^k_{F(\rho)}|/k$ and $|\mathbf{N}^k_{\text{real}} \cap \mathbf{N}_{k\text{NNc}}|/k$, where $|\cdot|$ is the number of elements. The normalisation is with respect to *k*, which yields a value between 0, no common element, and 1, all the elements are the same. The bars correspond to mean statistics over 414 tests and errorbars are the standard errors. In Fig. 7b we plot the absolute difference between the average ages in the real neighbourhood and the estimated neighbourhoods for different *k* values, where $\bar{\text{age}}_{\mathbf{A}} = \sum_{I\in\mathbf{A}}\text{age}(I)/|\mathbf{A}|$. Once again the bars are mean statistics and errorbars are standard errors. We observe in (a) that for



	Age	Predicted Age
Left:	67	29.8
Middle:	60	60.1
Right:	71	72.8

	Age	Predicted Age
Left:	26	56.2
Middle:	22	21.4
Right:	26	30.1

Fig. 4. Examples demonstrating where the regression approach using NAF fails. Left: six different images from the database. Right: Table presenting the predicted ages and the actual ages for the images displayed on the left. We note that where the predicted age and real age differs are actually cases the images have characteristics of a different age group.



(a) Feature locations selected by NAF based on age distance

(b) Feature locations selected by NAF based on intensity distance

Fig. 5. Feature Selection Frequency: Maps show the frequencies at which each feature is selected during the training of NAF at the first three levels of the trees throughout the forest. (a) NAF is trained on the pairwise image distance $\rho(\cdot,\cdot)$ defined on subjects' ages. (b) NAF is trained on a distance defined as the difference between feature vectors of the images, i.e. appearance distance. We note that the selected features are significantly different. NAF selects features from anatomical areas that show substantial change with ageing (Luft et al., 1999; Gunning-Dixon et al., 2008) when $\rho(\cdot,\cdot)$ depends on the subjects' age. When the distance depends only on the appearance of images this is not the case.

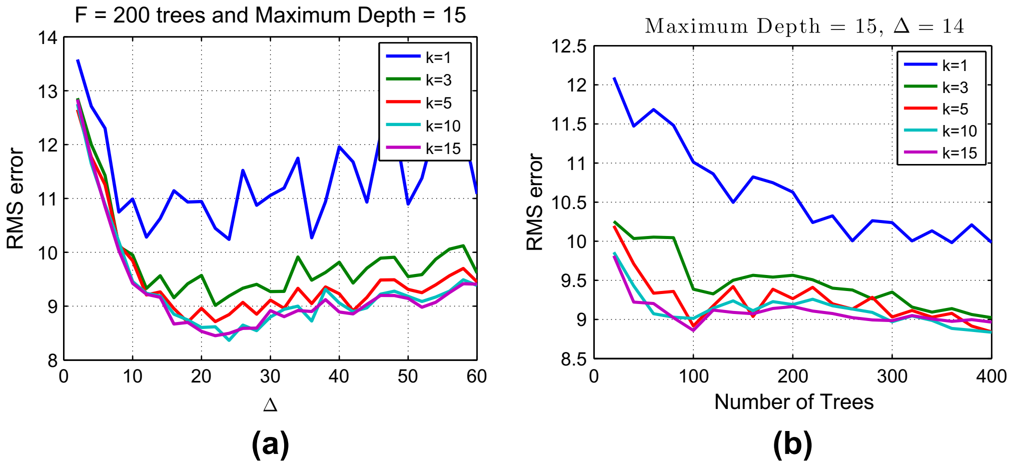


Fig. 6. (a) Regression accuracy as a function of Δ. Figure shows overfitting effects at low Δ and underlearning effects at high Δ. (b) Regression accuracy as a function of the number of trees. Figure shows that more trees in general improve accuracy however as the number of trees increase the marginal benefit of each additional tree becomes less.

small k the estimated neighbours differ from the true neighbours. As k increase the agreement between the true neighbourhood and the estimated one increases. In (b) we see that the mean ages in the estimated neighbourhood and the true one are similar for all k. This suggests that even though the estimated neighbours differ from the true neighbours, they are in the same age group, which is the important aspect for the analysis. We also see both in (a) and (b) that NAF produces a more accurate estimate of the neighbourhood compared to appearance-based clustering.

3.1.7. Comparison with supervised-metric learning

As we mentioned in the introduction, NAF is not the only option for estimating the neighbours of a test sample within a training data base, where neighbourhood is induced by an arbitrary distance. Supervised metric learning framework provides the tools for this purpose as well. Specifically, metric learning aims to reduce the dimensions of the feature space that data points are originally represented in, such that the Euclidean distance in the reduced dimension space approximates the neighbourhood structure between data points, which can be induced by an arbitrary distance. This is achieved by a globally linear or locally linear transformation of the original feature space. Metric learning methods use a training database to determine such a transformation. For a test image, they compute its out-of-sample projection on the reduced dimension space, where the distance between the test sample and the training samples provide an approximation of the closest neighbours.

In this last part, we compare NAF with four supervised metric learning methods: Neighbourhood Components Analysis (NCA)

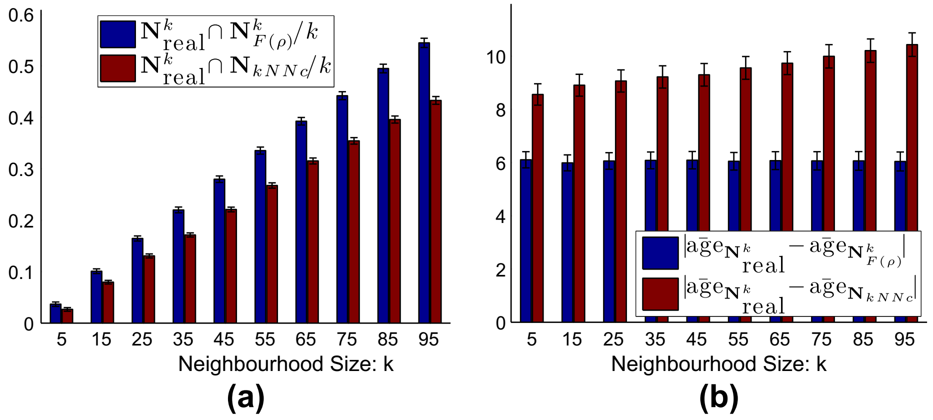


Fig. 7. Comparing the true neighbourhood with the estimates for the age-regression problem. (a) Normalised size of the set of intersections between the true neighbourhood and the estimated ones with different methods. (b) Absolute difference between the average ages in the true neighbourhood and the estimate ones. The true neighbourhood is computed using the distance ρ by assuming that the meta-information is available at the test time.

(Goldberger et al., 2004), Maximally Collapsing Metric Learning (MCML) (Globerson and Roweis, 2006), Large-Margin Nearest Neighbour (LMNN) (Weinberger et al., 2006) and Fisher's Linear Discriminant Analysis (LDA). We use the implementations of these algorithms in the publicly available.[1] Matlab Toolbox for Dimensionality Reduction (van der Maaten et al., 2009). Our evaluation is based on the regression accuracy of the neighbourhood-based approach that uses either of these algorithms.

In order to directly apply the available tools we convert age to categorical labels. For this we tested different bin sizes: (1, 2, 5, 10). Using each method we performed leave-one-out experiments. In each experiment we first reduced the dimension of the initial feature space from 10,000 to 75 and 100 using PCA. Running the metric learning methods on larger input space was not within our computational budget with the used implementation. Once again for each method we performed tests with both PCA dimensions. Lastly, we used different output space dimensions (10, 20, 50) for each algorithm and performed experiments for each value. Adding up all the parameter settings, we performed 24 leave-one-out experiments with each algorithm.

For each test image we compute its nearest neighbours and predict the age using the weighted mean as done previously for *k*NNc using Eqs. (7) and (8). This time instead of $\mathbf{N}_{k\text{NNc}}$ the neighbourhood found by the respective algorithm is used, such as $\mathbf{N}^k_{\text{NCA}}$, and instead of $\rho_{k\text{NNc}}(J,I_n)$ the respective Euclidean distance in the reduce dimension space is used, such as $\rho_{\text{NCA}}(J,I_n)$. Fig. 8 plots the RMS errors over 414 tests for different methods and different number of nearest neighbours used in the prediction process. For each supervised-metric learning method we show two bars that present the RMS errors corresponding to the best and the worst parameter settings. The plot shows that all of the algorithms perform similarly for this task. LDA and NAF produces slightly better error scores than NCA, LMNN and MCML.

Besides the quantitative differences one very important difference between NAF and the other methods is the way they use features. Supervised learning methods rely on determining a transformation of the original feature space. A transformation, however, mixes the effects of different features together. As a result, the dimensionality reduction comes at the expense of losing the interpretability of the features used by the algorithm. On the other hand, NAF, inheriting the advantages of the forest framework, uses the features directly without any transformations. As a result, it provides us the opportunity to further analyze the selected features and interpret them. This opens up new avenues and future research opportunities for tackling problems such as "can we find the relevant anatomical locations for a certain disorder?".

3.2. Efficient patch-based segmentation

In our second scenario, we build a patch-based segmentation system and apply it to multi-organ segmentation of unregistered, varying field-of-view CT scans. Here, NAF is trained to be able to retrieve semantically meaningful patches from a large database of manually annotated images. The corresponding segmentations of the retrieved patches are combined into a final segmentation of the whole scan. In contrast to previous patch-based approaches (Coupé et al., 2011; Rousseau et al., 2011) which rely on appearance-based distances only, the difference here is that patch similarity is defined employing a distance evaluated on the expert segmentation label maps.

3.2.1. Dataset

We use an in house database of 70 CT images with varying field-of-view. For each dataset a manual segmentation is available including the following structures: heart, liver, spleen, l. lung, r. lung, l. kidney, r.kidney, aorta, l. pelvis, r. pelvis, l. femur and r.femur. Fig. 11 shows some images from this database including their segmentations in the first two columns. For efficiency, we downsample all images to an isotropic voxel resolution of 4 mm. We would like to emphasize that the images do not need to be aligned for our method, and as can be seen in Fig. 11, the field-of-views of different images can be very different.

3.2.2. Application-specific NAF details

Distance $\rho(\cdot,\cdot)$: For the task of segmentation it is sensible to define a distance between patches based on the segmentations. In this experiment, the pairwise distance defining the neighbourhood structure in the database is $\rho(I,J) \triangleq \| seg(I) - seg(J)\|_0$, where $seg(I)$ denotes the label map for the image I and $\|\cdot\|_0$ is the L_0 norm. This $\rho(\cdot,\cdot)$ simply counts the number of voxels between I and J that have different labels. Fig. 9a displays nine image patches along with their manual segmentations. These images justify our choice in using the segmentations to define $\rho(\cdot,\cdot)$. The patches in the first row and the ones in the third row are very similar in terms of their appearance. However, they are very different in terms of their underlying anatomy as can be seen by the ground truth labels In

[1] http://homepage.tudelft.nl/19j49/Matlab_Toolbox_for_Dimensionality_Reduction.html.

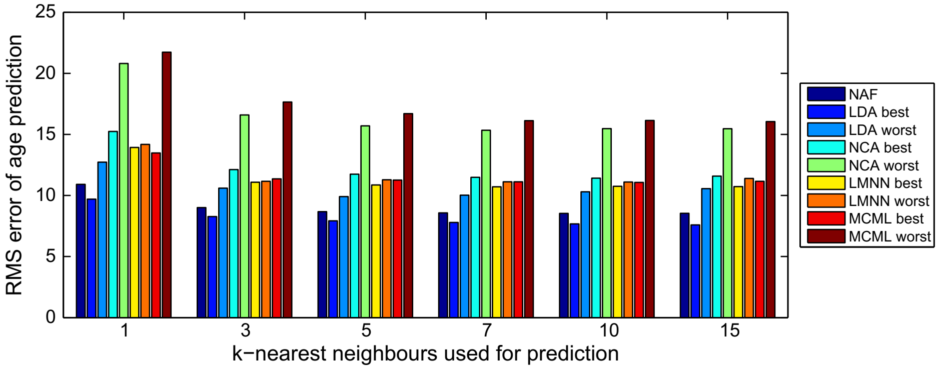


Fig. 8. Regression accuracies obtained by different neighbourhood-based approaches using NAF and four supervised metric learning algorithms. RMS scores are computed over 414 leave-one-out experiments for different neighbourhood sizes k. For each supervised metric learning method there are two bars showing results corresponding to the best and the worst parameter settings. We observe that NAF produces slightly better results than NCA, LMNN and MCML. LDA performs slightly better than NAF for the best parameter setting.

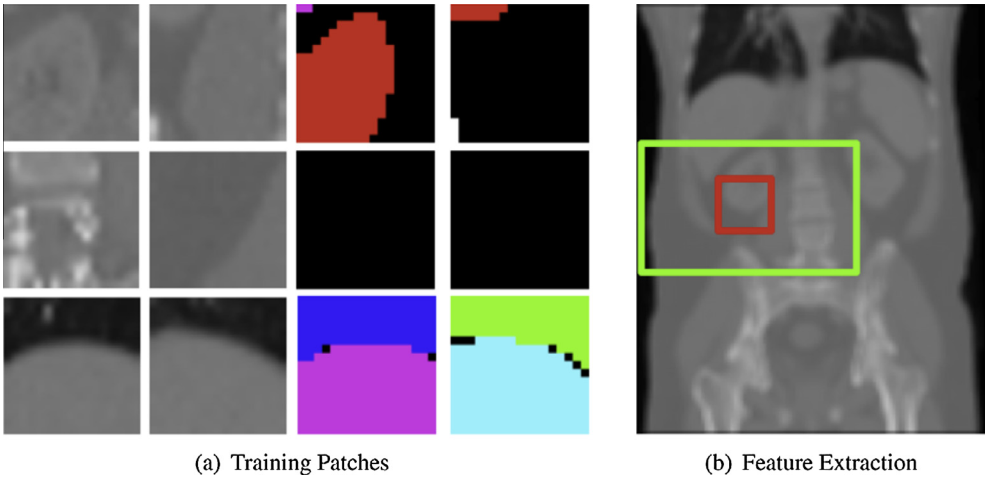


Fig. 9. (a) Training patches and the corresponding expert segmentations. We observe that some patches although similar in intensity profiles are very different in the associated label maps. (b) For each patch features are extracted which take into account the contextual information. For the patch denoted by a red square the features are intensity values at voxels randomly chosen within the green rectangle. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

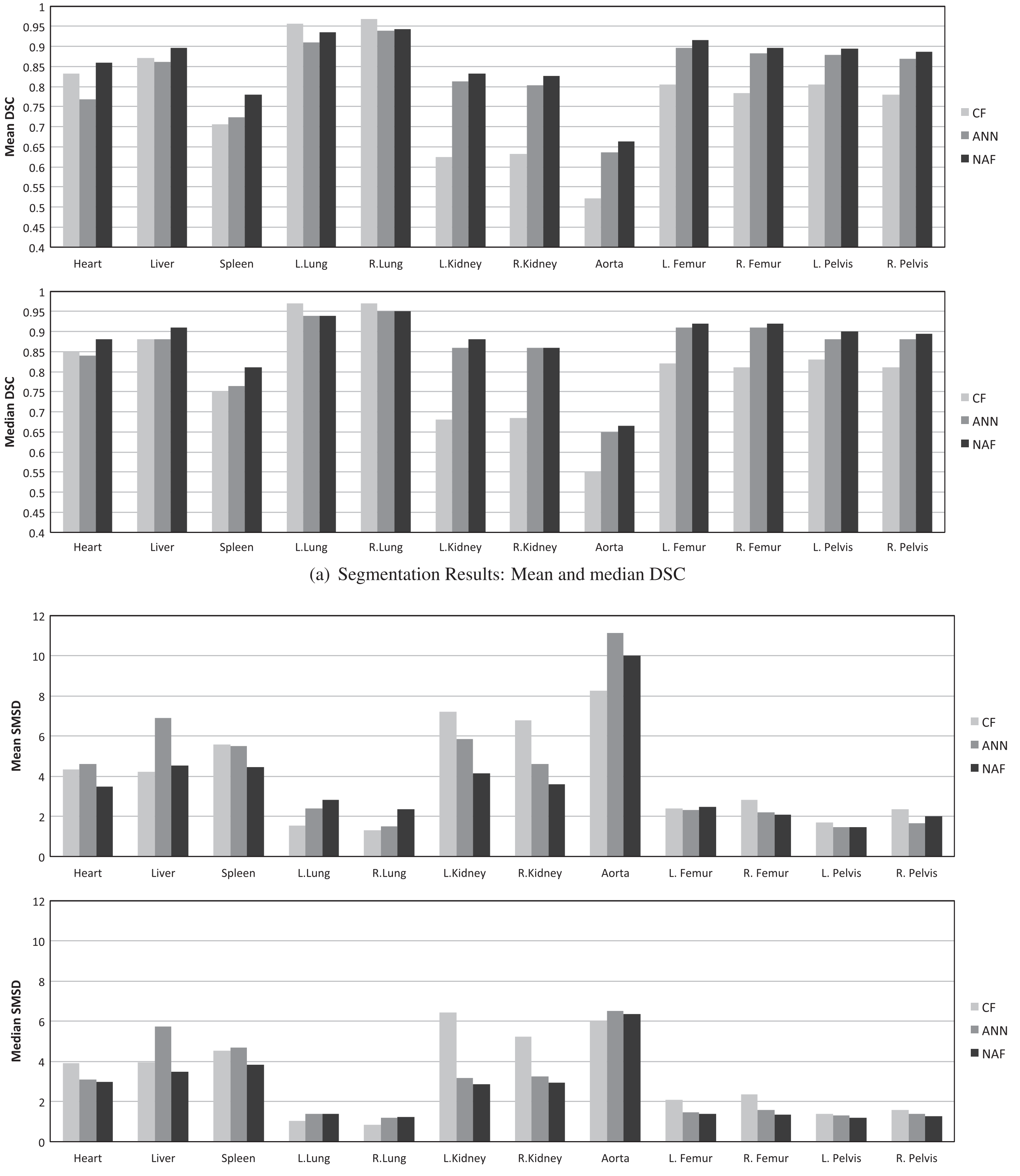
contrast, the patches in the second row are very different in appearance, but do not contain any structure of interest for this particular example. So in terms of segmentation purpose they are similar. The distance $\rho(\cdot,\cdot)$ will reflect the similarity (dissimilarity) of these cases during training.

Database **I**: We use 30 images for training and 40 for testing. We repeat the same experiment twice using different sets for training and testing. In each experiment the training database **I** is composed of 30,000 image patches along with their manual segmentations. These are obtained by dividing the training images into overlapping patches. The sampling interval for each patch centre is 32 mm and the size of the patch used for computing the pairwise distance is 68 × 68 × 68 mm^3. The distance $\rho(\cdot,\cdot)$ is computed between each pair of patches.

Feature space $\mathbf{f}(\cdot)$: The feature vector for each image patch consists of $Q = 1500$ intensity read-outs of voxels defined in an area of 200 × 200 × 80 mm^3 around the patch centre. These voxels are randomly chosen in a reference frame with respect to the patch centre, and the same set of local voxel coordinates are used for every patch. Fig. 9b illustrates the patch and the area for feature extraction. Intensities are obtained from smoothed version of the images. As before, we apply a mean filter with a physical size of 12 mm in each dimension.

NAF Parameters: We train NAF using the corresponding **I**, $\mathbf{f}(\cdot)$ and distance $\rho(I_n,I_m)$, $\forall I_n,I_m \in \mathbf{I}$ as explained above. The parameters are determined empirically and set to $q = 500$, F contains 150 trees, $\Delta = 20$ and the maximum depth for a tree is 17.

Patch-based segmentation details: The segmentation of a new image starts by dividing the image into overlapping patches. The sampling interval for patch centres is 12 mm in each dimension. The extracted patches used for generating the segmentations are of size 36 × 36 × 36 mm^3. For each test patch a feature vector is extracted and the set of 20 closest training patches within 30,000, i.e. $\mathbf{N}^{20}_{F(\rho)}(\cdot)$, is determined using NAF. For each of the 20 candidates, we perform a local refinement step using normalised cross correlation (NCC) within a search window of the same size as the actual patches centred at the patch location. The location that yields the highest NCC is assigned as the corresponding training patch for the test patch and its expert labels are used as a candidates for the test patch. As individual voxels are considered in several (overlapping) patches, their final label is obtained by majority voting. In addition to the overall segmentation, as a by-product, we obtain a reconstruction of the tested CT images based the retrieved training patches. Fig. 11 shows examples of the segmentation results using the patch-based NAF system, as well as the reconstructed CT images.



(a) Segmentation Results: Mean and median DSC

(b) Segmentation Results: Mean and median SMSD

Fig. 10. Segmentation scores for the three tested methods CF, ANN, and NAF. The bars correspond to per organ mean and median DSC in (a) and SMSD in (b) summarised over 80 subjects. On most organs, NAF achieves the highest DSC and lowest SMSD. The SMSD values are given in millimetres. Some visual examples are shown in Fig. 11.

In terms of computation times, the segmentation of a test image took between 7 and 30 min for volumetric images of sizes varying from 100 × 100 × 60 to 100 × 100 × 170 voxels. The training was done over night with most of the computation time spent on calculating the pairwise distances between 30,000 patches.

3.2.3. Evaluation and results

Fig. 10 shows the per-organ mean and median Dice's similarity coefficients (DSCs) and symmetric mean surface distances (SMSDs) between the expert annotations and the segmentations obtained with NAF and two other baselines. Note, higher DSC scores and

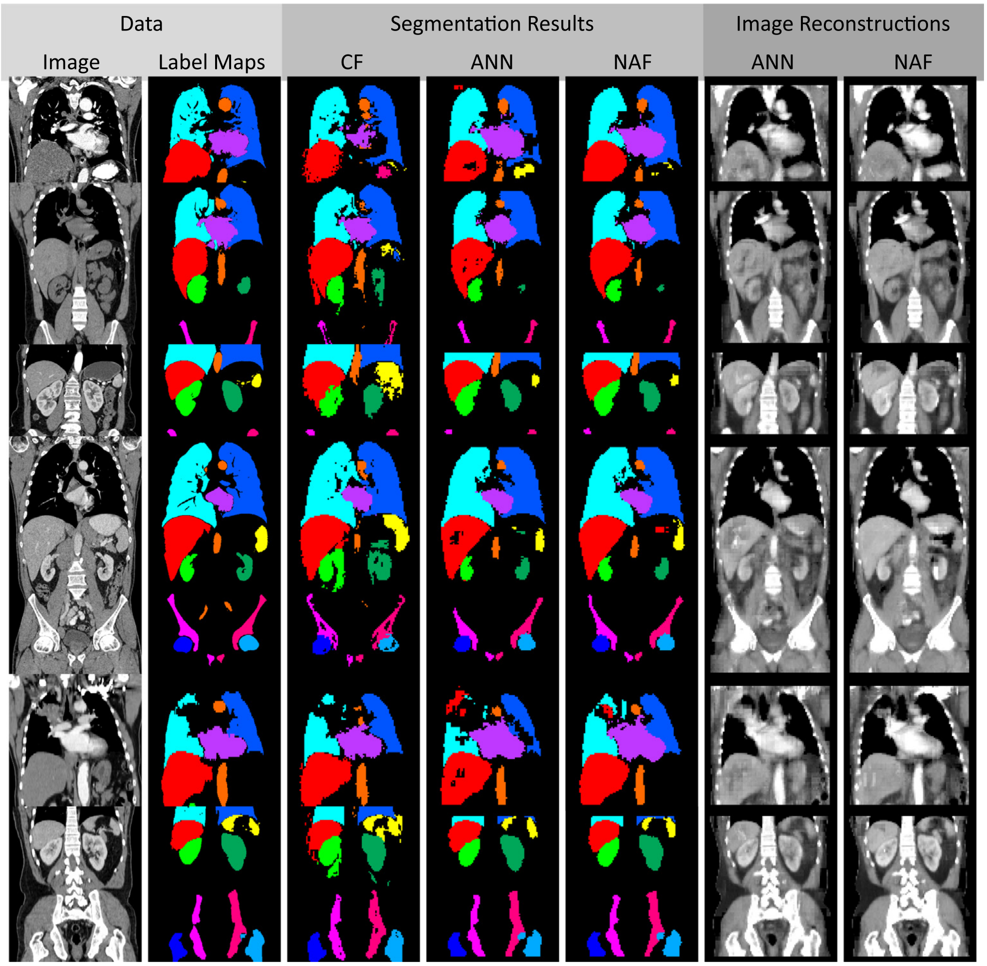


Fig. 11. Visual examples from segmentation experiments. Left: The test image and the manual segmentation label maps. Middle: Segmentations obtained by CF, ANN and NAF-based system. Right: The by-product reconstructions of the test image for ANN and NAF. Figures show improved accuracy and higher anatomical consistency for the NAF-based system. We also note that the reconstructed CT images are reasonably close to the original test images. This demonstrates that the closest patches found by NAF are not only consistent with respect to appearance but also consistent with respect to the underlying label maps.

lower SMSD correspond to higher segmentation accuracy. These values are obtained over 80 segmentations from the two runs with different training and test splits.

In order to assess the quality of these results, we compare them with results obtained with Classification Forests (CF) which operates as a voxel-based classification system (Lempitsky et al., 2009). The parameters of this baseline, such as tree depth, are optimised to obtain best possible results. The features used for CF are identical to the ones used for NAF.

Furthermore, in order to show the added benefit of the supervised learning for nearest neighbour retrieval, as performed by NAF, we also compare NAF's results to a patch-based system that uses unsupervised approximate nearest neighbour search (ANN) using kd-trees (Arya et al., 1998).[2] For the ANN approach, the nearest neighbours of a test image are determined based on the full appearance-based feature vectors, i.e. the distance between two images is $\rho_{ANN}(I_n, I_m) \triangleq \|\mathbf{f}(I_n) - \mathbf{f}(I_m)\|_2$. Here, kd-tree construction uses the same features NAF is exposed to during its training phase. The segmentation procedure for a test image is the same for both methods except the nearest neighbour search.

Quantitative evaluation confirms superior segmentation accuracy obtained by the patch-based approach using NAF when compared to the two baselines. For the ANN approach which also relies on patches, the difference is especially substantial for larger, homogeneous organs such as the liver or lungs appearance-based retrieval can fail by confusing ambiguous patches from different anatomical regions. In contrast, in these cases NAF is able to retrieve semantically meaningful patches with correct segmentation labels.

Focusing on smaller organs, such as kidneys or the elongated aortic arch, we observe the limitations of voxel-wise classification techniques such as CF. The better performance of patch-based systems is expected as spatial context is explicitly taken into account at test time, which yields spatially coherent label maps. Although CF is trained using features exploiting spatial context, at test time, each voxel is classified independently. The overall averaged mean DSC scores across all organs are 0.77 (±0.10) for CF, 0.83 (±0.10) for ANN, and 0.86 (±0.07) for NAF. The averaged mean SMSD scores are 4.04 (±2.72) mm for CF, 4.17 (±4.17) mm for ANN, and 3.62 (±3.87) mm for NAF.

We also observe that for the case of lungs CF outperforms both approaches, which we believe is due to lower spatial resolution of patch-based segmentation systems compared to voxel-based systems.

Fig. 11 displays some segmentation examples obtained with the three tested approaches along with the test images, expert segmentations, and the reconstructed CT images, both for ANN and

[2] Implementation available on http://www.cs.umd.edu/mount/ANN/.

NAF. The visual examples illustrate the behaviour of the different methods and confirm good quality segmentation results for NAF. It is interesting to note that both for ANN and NAF the image reconstructions are visually quite compelling, while the segmentation label maps reveal the better anatomical consistency for patches retrieved by NAF. We believe these results support the claim that NbA methods can benefit from supervision through the introduction of semantic information in the neighbourhood definition.

4. Conclusions

We propose an efficient algorithm for tackling one of the critical problems common to all neighbourhood-based approaches for image analysis: approximating the neighbourhood of an out-of-sample image within a (training) database. Our method is very flexible as it can handle neighbourhood structures induced by arbitrary distances, which can, for example, make use of semantic information. The algorithm does not assume pre-existing descriptive compact features. Instead, it automatically learns distance-specific features starting from high dimensional image representations. As a result, it offers a powerful tool that can be used in diverse applications.

NAF is generic and can be integrated in existing systems. Furthermore, it also enables novel applications. So far, the use of neighbourhood-based approaches in medical image analysis has been limited partly due to limitations of unsupervised clustering methods for nearest neighbour retrieval. NAF opens up new avenues in this respect by introducing supervised training for the retrieval procedure. We believe our experiments demonstrate the potential of NAF, and that our framework could become an important ingredient in other tasks, such as multi-atlas label fusion and manifold-based techniques.

One of the open questions that has not been addressed here is the requirement to keep all the training data at hand during testing. In the image-based age regression experiments presented here, this was obviously not an issue because only the age information needed to be stored. However, for the patch-based segmentation experiments, we assumed that the patch database was available during testing. In general, this can be a limitation for applications using very large databases or if there is a restricted memory budget. In fact, this is a persistent issue for all nearest neighbour methods. However, we believe that the framework of NAF would allow for possible application-specific solutions for this problem. Specifically, the leaf nodes, which currently store the indices to database images, could accommodate lower dimensional probabilistic models extracted from the database images falling into the same leaf. This could yield memory efficient systems while making available the variability present in the database for the subsequent analyses. Such schemes could be used to retrieve patient-specific a priori models.

Another possible area of investigation is the relation between various decision forest variants, such as classification (Shotton et al., 2008), regression (Glocker et al., 2012), visual codebook generation (Moosmann et al., 2007) and NAF. We already mentioned earlier that the use of decisions forests for clustering and manifold learning can be seen as special cases of NAF. Other variants can also be linked to NAF by defining the appropriate distances related to the task at hand. The important difference between NAF and the standard forest approaches is that the latter require task-specific objective functions, for example based on Shannon entropy in classification, or variance reduction in regression. NAF uses a generic objective function which is the same for all applications. This results: (1) in a forest-based framework that can be used for various applications without any modification on the algorithm itself and (2) a more general framework that can accommodate different applications such as patch-based segmentation, for which the objective function in the standard forest framework is not trivial to define.

References

Aljabar, P., Heckemann, R., Hammers, A., Hajnal, J., Rueckert, D., 2009. Multi-atlas based segmentation of brain images: atlas selection and its effect on accuracy. NeuroImage 46 (3), 726–738.

Aljabar, P., Wolz, R., Rueckert, D., 2012. Machine learning in computer-aided diagnosis: medical imaging intelligence and analysis. In: IGI Global, 2012, Ch. Manifold Learning for Medical Image Registration, Segmentation, and Classification, pp. 351–372.

Allassonnière, S., Amit, Y., Trouvé, A., 2007. Towards a coherent statistical framework for dense deformable template estimation. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 69 (1), 3–29.

Amit, Y., Geman, D., 1997. Shape quantization and recognition with randomized trees. Neural Computation 9 (7), 1545–1588.

Artaechevarria, X., Munoz-Barrutia, A., Ortiz-de Solorzano, C., 2009. Combination strategies in multi-atlas image segmentation: application to brain MR data. IEEE Transactions on Medical Imaging 28 (8), 1266–1277.

Arya, S., Mount, D., Netanyahu, N., Silverman, R., Wu, A., 1998. An optimal algorithm for approximate nearest neighbor searching fixed dimensions. Journal of the ACM (JACM) 45 (6), 891–923.

Atasoy, S., Mateus, D., Meining, A., Yang, G., Navab, N., 2012. Endoscopic video manifolds for targeted optical biopsy. IEEE Transactions on Medical Imaging 31 (3), 637–653.

Bengio, Y., Paiement, J., Vincent, P., Delalleau, O., Le Roux, N., Ouimet, M. 2004. Out-of-sample extensions for LLE, Isomap, MDS, Eigenmaps, and spectral clustering. In: Advances in Neural Information Processing Systems, vol. 16, pp. 177–184.

Boiman, O., Shechtman, E., Irani, M., 2008. In defense of nearest-neighbor based image classification. In: Computer Vision and Pattern Recognition, pp. 1–8.

Breiman, L., 2001. Random forests. Machine Learning 45 (1), 5–32.

Chechik, G., Sharma, V., Shalit, U., Bengio, S., 2010. Large scale online learning of image similarity through ranking. The Journal of Machine Learning Research 11, 1109–1135.

Cho, Y., Seong, J., Shin, S., Jeong, Y., Kim, J., Qiu, A., Im, K., Lee, J., Na, D., 2011. A multi-resolution scheme for distortion-minimizing mapping between human subcortical structures based on geodesic construction on Riemannian manifolds. NeuroImage 57 (4), 1376–1392.

Coupé, P., Manjón, J., Fonov, V., Pruessner, J., Robles, M., Collins, D., 2011. Patch-based segmentation using expert priors: application to hippocampus and ventricle segmentation. NeuroImage 54 (2), 940–954.

Criminisi, A., Shotton, J., Konukoglu, E., 2011. Decision forests: a unified framework. Foundations and Trends in Computer Graphics and Vision 7 (2–3).

Der, M., Saul, L.K. 2012. Latent coincidence analysis: a hidden variable model for distance metric learning. In: Advances in Neural Information Processing Systems, vol. 25, pp. 3239–3247.

Franke, K., Ziegler, G., Klöppel, S., Gaser, C., 2010. Estimating the age of healthy subjects from t_1-weighted mri scans using kernel methods: Exploring the influence of various parameters. Neuroimage 50 (3), 883–892.

Fu, H., Zhang, Q., Qiu, G., 2012. Random forest for image annotation. In: European Conference on Computer Vision, pp. 86–99.

Gaonkar, B., Davatzikos, C. 2012. Deriving statistical significance maps for SVM based image classification and group comparisons. In: Medical Image Computing and Computer Assisted Intervention, pp. 723–730.

Globerson, A., Roweis, S. 2006. Metric learning by collapsing classes. In: Advances in Neural Information Processing Systems, vol. 18, pp. 451–458.

Globerson, A., Roweis, S., 2006. Metric learning by collapsing classes. In: Advances in neural information processing systems, vol. 18.

Glocker, B., Feulner, J., Criminisi, A., Haynor, D., Konukoglu, E., 2012. Automatic localization and identification of vertebrae in arbitrary field-of-view ct scans. Medical Image Computing and Computer-Assisted Intervention – MICCAI 2012, 590–598.

Goldberger, J., Roweis, S., Hinton, G., Salakhutdinov, R., 2004. Neighbourhood components analysis. In: Advances in Neural Information Processing Systems 17.

Goldberger, J., Roweis, S., Hinton, G., Salakhutdinov, R. 2005. Neighbourhood Components Analysis. In: Advances in Neural Information Processing Systems, 17, pp. 513–520.

Gray, K., Aljabar, P., Heckemann, R., Hammers, A., Rueckert, D., 2011. Random forest-based manifold learning for classification of imaging data in dementia. In: Machine Learning in Medical Imaging (MICCAI Workshop), pp. 159–166.

Guerrero, R., Wolz, R., Rueckert, D. 2011. Laplacian eigenmaps manifold learning for landmark localization in brain MR images. In: Medical Image Computing and Computer Assisted Intervention, pp. 566–573.

Gunning-Dixon, F.M., Brickman, A.M., Cheng, J.C., Alexopoulos, G.S., 2008. Aging of cerebral white matter: a review of MRI findings. International Journal of Geriatric Psychiatry 24 (2), 109–117.

Hamm, J., Ye, D., Verma, R., Davatzikos, C., 2010. GRAM: a framework for geodesic registration on anatomical manifolds. Medical Image Analysis 14 (5), 633.

He, X., Niyogi, P., 2004. Locality preserving projections. In: Advances in Neural Information Processing Systems, vol. 16, pp. 153–160.

Isgum, I., Staring, M., Rutten, A., Prokop, M., Viergever, M., van Ginneken, B., 2009. Multi-atlas-based segmentation with local decision fusion – application to cardiac and aortic segmentation in CT scans. IEEE Transactions on Medical Imaging 28 (7), 1000–1010.

Jia, H., Yap, P., Shen, D., 2012. Iterative multi-atlas-based multi-image segmentation with tree-based registration. NeuroImage 59 (1), 422–430.

Konukoglu, E., Glocker, B., Zikic, D., Criminisi, A., 2012. Neighbourhood approximation forests. In: Medical Image Computing and Computer Assisted Intervention, pp. 75–82.

Lempitsky, V., Verhoek, M., Noble, J., Blake, A. 2009. Random forest classification for automatic delineation of myocardium in real-time 3D echocardiography. In: Functional Imaging and Modeling of the Heart, pp. 447–456.

Lian, N., Davatzikos, C., 2011. Morphological appearance manifolds for group-wise morphometric analysis. Medical Image Analysis 15 (6), 814–829.

Luft, A.R., Skalej, M., Schulz, J.B., Welte, D., Kolb, R., Bürk, K., Klockgether, T., Voigt, K., 1999. Patterns of age-related shrinkage in cerebellum and brainstem observed in vivo using three-dimensional MRI volumetry. Cerebral Cortex 9 (7), 712–721.

Marcus, D., Wang, T., Parker, J., Csernansky, J., Morris, J., Buckner, R., 2007. Open Access Series of Imaging Studies (OASIS): cross-sectional MRI data in young, middle aged, nondemented, and demented older adults. Journal of Cognitive Neuroscience 19 (9), 1498–1507.

Moosmann, F., Triggs, B., Jurie, F., et al. 2007. Fast discriminative visual codebooks using randomized clustering forests. In: Advances in Neural Information Processing Systems 19, pp. 985–992.

Muja, M., Lowe, D., 2009. Fast approximate nearest neighbors with automatic algorithm configuration. In: International Conference on Computer Vision Theory and Applications, pp. 331–340.

Nister, D., Stewenius, H., 2006. Scalable recognition with a vocabulary tree. In: Computer Vision and Pattern Recognition, pp. 2161–2168.

Norouzi, M., Fleet, D.J. 2011. Minimal loss hashing for compact binary codes. In: International Conference on Machine Learning, pp. 353–360.

Rohlfing, T., Russakoff, D., Maurer, C. 2003. Expectation maximization strategies for multi-atlas multi-label segmentation. In: Information Processing in Medical Imaging, pp. 210–221.

Rousseau, F., Habas, P., Studholme, C., 2011. A supervised patch-based approach for human brain labeling. IEEE Transactions on Medical Imaging 30 (10), 1852–1862.

Sabuncu, M.R., Van Leemput, K., 2011. The Relevance Voxel Machine (RVoxM): a Bayesian method for image-based prediction. In: Medical Image Computing and Computer Assisted Intervention, pp. 99–106.

Sabuncu, M., Balci, S., Shenton, M., Golland, P., 2009. Image-driven population analysis through mixture modeling. IEEE Transactions on Medical Imaging 28 (9), 1473–1487.

Shotton, J., Johnson, M., Cipolla, R. 2008. Semantic text on forests for image categorization and segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition, 2008. CVPR 2008, IEEE, pp. 1–8.

van der Maaten, L., Postma, E., van den Heric, H., 2009. Dimensionality Reduction: A Comparative Review. Tech. Rep. TiCC-TR 2009-005. Tilburg University.

van Rikxoort, E., Isgum, I., Arzhaeva, Y., Staring, M., Klein, S., Viergever, M., Pluim, J., van Ginneken, B., et al., 2010. Adaptive local multi-atlas segmentation: application to the heart and the caudate nucleus. Medical Image Analysis 14 (1), 39–49.

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

Weinberger, K.Q., Blitzer, J., Saul, L.K. 2006. Distance metric learning for large margin nearest neighbor classification. In: Advances in Neural Information Processing Systems, vol. 18.

Weiss, Y., Torralba, A., Fergus, R. 2009. Spectral Hashing. In: Advances in Neural Information Processing Systems, vol. 21, pp. 1753–1760.

Wolz, R., Aljabar, P., Hajnal, J., Hammers, A., Rueckert, D., 2010. LEAP: learning embeddings for atlas propagation. NeuroImage 49 (2), 1316–1325.

Xing, E., Ng, A., Jordan, M., Russell, S. 2002. Distance metric learning, with application to clustering with side-information. In: Advances in Neural Information Processing Systems, vol. 15, pp. 505–512.