Medical Image Analysis 58 (2019) 101550

Contents lists available at ScienceDirect

Medical Image Analysis

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

How many models/atlases are needed as priors for capturing anatomic population variations?

Ze Jin, Jayaram K. Udupa*, Drew A. Torigian

Medical Image Processing Group, Department of Radiology, University of Pennsylvania, Philadelphia, United States

ARTICLE INFO

Article history:
Received 29 October 2018
Revised 24 August 2019
Accepted 29 August 2019
Available online 3 September 2019

Keywords:
Model/atlas based methods
Image grouping
Thoracic CT
Head and neck CT
Hierarchical agglomerative clustering

ABSTRACT

Many medical image processing and analysis operations can benefit a great deal from prior information encoded in the form of models/atlases to capture variations over a population in form, shape, anatomic layout, and image appearance of objects. However, two fundamental questions have not been addressed in the literature: "*How many models/atlases are needed for optimally encoding prior information to address the differing body habitus factor in that population?*" and "*Images of how many subjects in the given population are needed to optimally harness prior information?*" We propose a method to seek answers to these questions.

We assume that there is a well-defined body region of interest and a subject population under consideration, and that we are given a set of representative images of the body region for the population. After images are trimmed to the exact body region, a hierarchical agglomerative clustering algorithm partitions the set of images into a specified number of groups by using pairwise image (dis)similarity as a cost function. Optionally the images may be pre-registered among themselves prior to this partitioning operation. We define a measure called *Residual Dissimilarity* (*RD*) to determine the goodness of each partition. We then ascertain how RD varies as a function of the number of elements in the partition for finding the optimum number(s) of groups. Breakpoints in this function are taken as the recommended number of groups/models/atlases.

Our results from analysis of sizeable CT data sets of adult patients from two body regions – thorax (346) and head and neck (298) – can be summarized as follows. (1) A minimum of 5 to 8 groups (or models/atlases) seems essential to properly capture information about differing anatomic forms and body habitus. (2) A minimum of 150 images from different subjects in a population seems essential to cover the anatomical variations for a given body region. (3) In grouping, body habitus variations seem to override differences due to other factors such as gender, with/without contrast enhancement in image acquisition, and presence of moderate pathology.

This method may be helpful for constructing high quality models/atlases from a sufficiently large population of images and in optimally selecting the training image sets needed in deep learning strategies.

1. Introduction

It is now well established that medical image processing and analysis operations, particularly segmentation, can benefit a great deal from prior information gathered over a population in form, shape, anatomic layout, and image appearance of objects. In this context, "object" here may refer to a well-defined body region, such as abdomen, an organ such as cervical esophagus, a sub-organ such as a lobe in the liver, an anatomic space occupied by a specific type of tissue such as visceral adipose tissue in a body region, or a well-defined lymph-node station/zone such as mediastinal zone 4R. For encoding (learning) population prior information, many methods have been developed. In image segmentation, for example, which is the main motivation for this work, known object shape, image appearance, and object relation information over a subject population are first codified (learned) and then utilized on a given image to bring constraints into the segmentation process. They evolved precisely to overcome failure of purely-image-based-approaches in situations such as lack of definable object boundaries in the image (an extreme example is lymph node zones which do not present with any identifiable intensity

* Corresponding author at: Medical Image Processing Group, 3710 Hamilton Walk, Room 602W, Goddard Building, 6th floor, Philadelphia, PA 19104, United States.

E-mail addresses: jay@pennmedicine.upenn.edu, jay@mail.med.upenn.edu (J.K. Udupa).

https://doi.org/10.1016/j.media.2019.101550
1361-8415/© 2019 Elsevier B.V. All rights reserved.

boundaries), variable object boundary characteristics, and image artifacts, and also simply to increase level of automation. Among such prior-knowledge-encoding methods, three distinct classes of methods can be identified – *model-based* (Staib and Duncan, 1992; Cootes et al., 1995; Bardinet et al., 1998; Cootes et al., 2001; Heimann and Meinzer 2009; Shen et al., 2011; Udupa et al., 2014), *atlas-based* (Gee et al., 1993; Christensen et al., 1994; Ashburner and Friston, 2009; Isgum et al., 2012; Wolz et al., 2012; Wang and Yushkevich, 2013; Yang et al., 2014; Wu et al., 2015; Bai et al., 2015; Xu et al., 2015; Shi et al., 2017; Nguyen et al., 2018), and *deep-learning-*(*DL*)*-based* (Turaga et al., 2010; Ciresan et al., 2013; Zhang et al., 2015; Moeskops et al., 2016; Oda et al., 2018; Drozdza et al., 2018). The division between model- and atlas-based groups is somewhat arbitrary and a matter of semantics. In fact, DL networks are also often referred to as "models".[1] Both modeling and atlas approaches assume the following: (i) A set of *representative* gray-level images (in a given imaging modality such as computed tomography (CT)) together with binary images depicting the objects of interest (as defined above) that are labeled (delineated) in the gray images are available. "Representative" here implies that the images *pertain to and properly represent* some subject population of interest. (ii) These images and the associated binarily-defined objects will act as a proxy for the same objects in any patient image of the same modality.

Model-based approaches rely on two key considerations: (M1) Building a *model* from the given gray and binary images to capture *explicitly* within the model variation over the population of interest in key attributes of objects including their shape, size, pose, anatomic/geographic layout (object relationships), and image appearance. (M2) Using this prior information to estimate as accurately as possible the values of these attributes for a given patient image to be analyzed so that the model can be "fitted" accurately to the patient image for each object that is included in the model. Methods differ in the way they represent objects, handle the attributes, and perform the fitting operation. It is conceivable that if multiple models are created and utilized to handle different body habitus (thin and tall, thin and short, etc.) within the population of interest, instead of employing a single model, then the model most suited to the body habitus of the patient under consideration can be chosen from the ensemble to improve the "fit". Interestingly, multi-model modeling approaches[2] (Rittner et al., 2014) are not as common in the literature as multi-atlas strategies (see below) to address the body habitus issue.

Atlas-based approaches operate differently from model-based methods but also rely on two considerations. (A1) *Atlas* building. The given images (gray and binary) are registered over the set so they are represented in the same space. The resulting set is called an *atlas* or a collection of *atlases* (or an *atlas ensemble*) depending on how it is subsequently employed. The idea is that, for any given patient image, we should be able to find in the atlas set some images that best match with the patient image. (A2) Combine the binary masks in the matched atlas images to determine the binary (or fuzzy or probabilistic) mask for each object in the patient image. As in the modeling approach, numerous possibilities exist for this approach. For example, the ensemble may contain just one gray image and the associated binary images, and objects are segmented by registering the atlas gray image to the patient image and accordingly propagating the binary masks to the patient image (Gee et al., 1993; Christensen et al., 1994). When the ensemble contains more than one image, the single image within the ensemble that best matches the patient image can be determined via registration, and then the binary object masks can be propagated (Yang et al., 2014). These are both *single-atlas* approaches. Alternatively, an *average atlas* can be created by averaging all (or selected) gray images in the ensemble and averaging all corresponding binary images and following the above single-atlas strategy to match the average atlas to the patient image (Ashburner and Friston, 2009). In *multi-atlas* strategies (Isgum et al., 2012; Wang and Yushkevich, 2013; Wu et al., 2015, Bai et al., 2015; Xu et al., 2015; Shi et al., 2017; Nguyen et al., 2018), a few top-matching images from the ensemble are found first, and subsequently their associated object masks are "fused" to yield segmentations of the objects. These methods have become popular in recent years owing to their improved accuracy and robustness, notwithstanding their higher computational burden. Notably, in all strategies, matching can also be done separately by individual objects and "patches" rather than whole images to reach a consensus.

There are two key requirements for building models or atlases from the given gray and binary images toward object segmentation: (1) Ability to capture explicitly within the model(s)/atlas(es) variations over the population of interest in attributes of objects including their size, shape, pose, anatomic/geographic layout, and appearance. (2) Ability to provide means as *precise as possible* to fit the model(s) or atlas(es) to the target image for each object. There is a contradiction of sorts with these two requirements. If a single model/atlas is employed for covering large populations with large variations, the boundary information of objects maybe blurred in atlas-based approaches beyond acceptable use and the object model may vary in unpredictable ways in model-based approaches.

On the other hand, if multiple models/atlases are employed, a fundamental question arises in both approaches[3]: *(Q1) How many models/atlases are needed for optimally encoding prior information to address the differing body habitus factor in that population?* We will assume that this number is the same for all objects,[4] whereby we may translate this question from object level to image/subject level as a related question: *(Q2) Images of how many subjects in the given population are needed to optimally harness prior information?* Optimality in Q2 implies that adding more images beyond the optimal number does not result in any significant gain in encoded prior information. It also implies that if the number is smaller, then there may be significant loss in the encoded prior information. Although some studies have examined the number of atlases needed for optimal segmentation accuracy (Sanroma et al., 2014; Van de Velde et al., 2016) in a given application, neither of these questions has been addressed in the literature. For most studies, the number chosen for the models/atlases has been quite arbitrary. It is understood in all previous works that modeling/atlas building pertains to a specific subject population considered for the application at hand. Although multi-atlas/multi-model methods have shown better accuracy for image segmentation, if the atlases/models do not cover representatively the distinct groups of subjects within the target population, the methods may not be generalizable to other samples of the same population irrespective of whether or not their total number is large enough.

Indeed, it may be difficult to seek satisfactory answers to the above questions independent of applications. This suggests that, perhaps, it is best to first define key biological characteristics most

[1] In this paper, we will not consider deep learning strategies any further. See Section 4 for comments on this topic.

[2] Note that *multi-model* modeling is different from *multi-object* modeling. In the latter case, multiple objects are treated as a single composite object and modeled. Multi-modeling strategies can be employed in both cases – single object and multiple objects.

[3] This issue is applicable to all approaches that harness and exploit prior information for image segmentation and analysis. See further comments in Section 4.

[4] This may not be true, but we believe that understanding grouping at the image level first is necessary. Similar strategies can then be devised at the object level either independently or as a second refined step following image-level grouping. See further comments in Section 4.

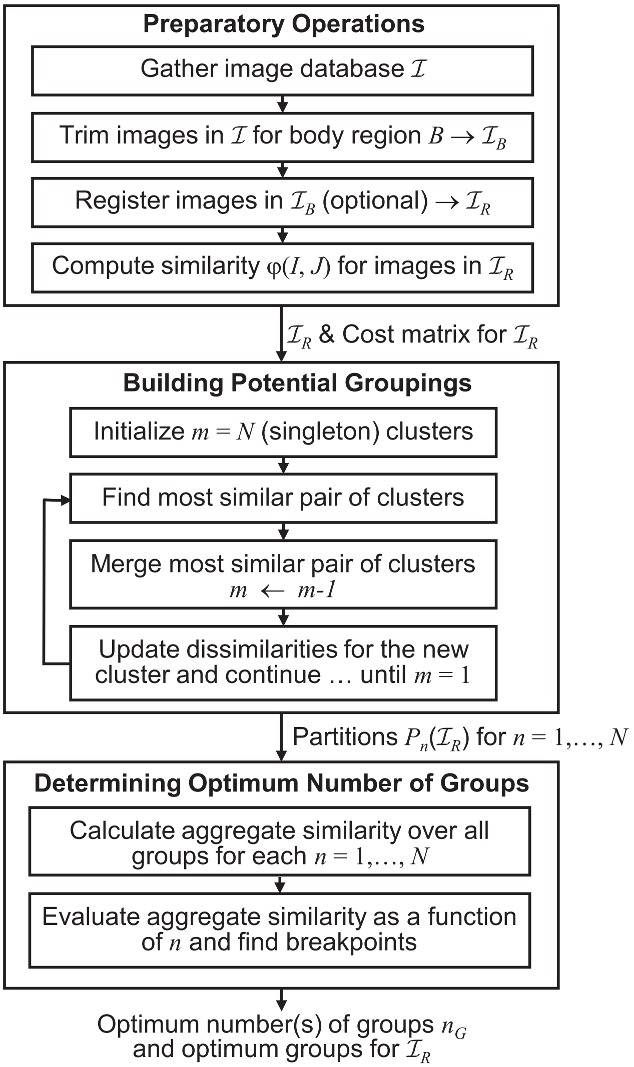
Preparatory Operations
Gather image database $\mathcal{I}$
Trim images in $\mathcal{I}$ for body region $B \to \mathcal{I}_B$
Register images in $\mathcal{I}_B$ (optional) $\to \mathcal{I}_R$
Compute similarity $\varphi(I, J)$ for images in $\mathcal{I}_R$
$\mathcal{I}_R$ & Cost matrix for $\mathcal{I}_R$
Building Potential Groupings
Initialize $m = N$ (singleton) clusters
Find most similar pair of clusters
Merge most similar pair of clusters $m \leftarrow m\text{-}1$
Update dissimilarities for the new cluster and continue … until $m = 1$
Partitions $P_n(\mathcal{I}_R)$ for $n = 1,\ldots, N$
Determining Optimum Number of Groups
Calculate aggregate similarity over all groups for each $n = 1,\ldots, N$
Evaluate aggregate similarity as a function of n and find breakpoints
Optimum number(s) of groups n_G and optimum groups for $\mathcal{I}_R$

Fig. 1. A schematic representation of the proposed approach. The three main steps are explained in Sections 2.1–2.3.

relevant for the application at hand and then set out to answer Q1/Q2 by using these characteristics as a guide. Defining these characteristics may be as hard as answering the questions themselves. Our application motivation for this work is segmentation of objects body-wide, where “objects”, as defined above, may be body regions (Bai et al., 2019), tissue regions (Liu et al., 2019), anatomic organs (Udupa et al., 2014; Wu et al., 2019), or lymph node zones (Xu et al., 2018), all handled within the same single system. In this context, we believe that the most relevant biological characteristic to guide grouping at the image level is anatomical body habitus – gross form, size, and geographic layout of objects in the body. The design of a proposed measure of the quality of grouping called *residual dissimilarity* is indeed for capturing exactly the body habitus characteristic.

In this paper, as described in Section 2, we present a method to seek an answer to the above two questions. We assume that there is a body region B of interest and a subject population $\mathcal{P}$ under consideration, and that we are given a set of images $\mathcal{I}$ of the body region B that is *representative* of the population $\mathcal{P}$. First, we perform several preparatory operations on the images in $\mathcal{I}$; see Fig. 1 for a schematic of our approach. They include trimming the images in $\mathcal{I}$ as per the definition of body region B; registration of the resulting images among themselves (this operation may be skipped); computing pairwise similarity of the (registered) images as per a given similarity metric; performing hierarchical agglomerative grouping of the images for a specified number, n, of groups by using pairwise (dis)similarity as a cost function; analyzing the aggregate similarity achieved over all groups for each n as a function of n; and identifying significant breakpoints in this functional behavior. Using sizeable CT data sets of adult subjects from two body regions – thorax and head and neck (H&N), we assess the breakpoints. We describe our experimental process and findings in Section 3. We summarize our concluding remarks in Section 4.

A very preliminary version of this work was reported at the SPIE 2019 Medical Imaging conference (Jin et al., 2019). The current paper differs from the conference paper in major ways: (i) A full description of the motivation, background, and literature which the conference paper did not include. (ii) A full and detailed description of the method; only an outline of the method was presented in the conference paper. (iii) A detailed presentation of the results and their analysis whereas a summary of specific results only for one population and for one body region was presented in the conference paper. (iv) A detailed discussion of the results, conclusions, and gaps remaining.

2. Methods

2.1. Notations

We assume that there is a well-defined subject population $\mathcal{P}$ of interest. $\mathcal{P}$ may be quite encompassing as in “the group of male or female adult subjects”, or more restrictive as in “male subjects in the age range of 40 to 59 years who are radiologically normal”. The answers we find for our question Q2 may accordingly vary and can thus be tailored to the application at hand. We also assume that there is a well-defined body region B of focus. We employ precise computationally motivated definitions of each body region (Udupa et al., 2014; Wu et al., 2019). For example, for the two body regions involved in our experimental study in this paper, we use the following definitions. *Thorax*: Extends cranio-caudally from 15 mm superiorly from the apex of the lung to 5 mm inferiorly from the base of the lung, with the full body region included within the field of view in the other two orthogonal directions. *Head-and-neck* (H&N): Extends cranio-caudally from the superior-most aspect of the mandible to the axial level at which the superior vena cava branches into left and right brachiocephalic veins, with the full body region included within the field of view in the other two orthogonal directions. Without such definitions, our aim of seeking commonality-based groupings within a population becomes meaningless. Following the principles underlying the previous automatic anatomy recognition (AAR) framework (Udupa et al., 2014), we additionally assume that we are given a set $\mathcal{I} = \{J_1, \ldots, J_N\}$ of images *covering fully* body region B of subjects from population $\mathcal{P}$. It is important that subjects whose images are included in $\mathcal{I}$ are representative of $\mathcal{P}$, meaning that $\mathcal{I}$ properly covers the body habitus variations for $\mathcal{P}$. “Covering fully” implies that each image in $\mathcal{I}$ properly includes the body region B.

Our approach, after images in $\mathcal{I}$ are trimmed and optionally registered, is to create a partition $P_n(\mathcal{I}) = \{G_1^n, G_2^n \ldots, G_n^n\}$ of $\mathcal{I}$ into a specified number n of groups by optimizing the collective similarity of images in each group. We then ascertain how the overall residual dissimilarity $\gamma(n)$ remaining in partition $P_n(\mathcal{I})$ varies as we change n from 1 to N. Subsequently, values of n at which there are significant changes in $\gamma(n)$ are determined. These “optimal” or significant values or *breakpoints* are taken as the recommended number n_G of groups/models/atlases.

Table 1
Summary of the number of CT scan data sets used in this study.

		Near-normal	Intermediately abnormal	Abnormal	Total
Thorax	With-contrast	73	4	5	82
Voxel size: $\sim 1 \times 1 \times 2.5$–$5\ \text{mm}^3$	No-contrast	172	35	57	264
	Total	245	39	62	346
Head & neck	With-contrast	75	36	116	227
Voxel size: $\sim 1 \times 1 \times 1.5$–$3\ \text{mm}^3$	No-contrast	36	10	25	71
	Total	111	46	141	298

2.2. *Preparatory operations*

2.2.1. *Gathering image data*

This retrospective study was conducted following approval from the Institutional Review Board at the Hospital of the University of Pennsylvania along with a Health Insurance Portability and Accountability Act waiver. We utilized CT image data sets from two body regions – thorax and H&N. Details pertaining to the data sets are summarized in Table 1.

For both body regions, the CT scan data included those acquired either with or without intravenous contrast material and were gathered from our hospital patient image database of adult *male and female*[5] patients with various types of cancers and other pathologies. The age range of the patients was 40 to 79 years. The scan resolutions were roughly $1 \times 1 \times 2.5$–$5\,\text{mm}^3$ for the thorax and $1 \times 1 \times 2\,\text{mm}^3$ for H&N. Both thoracic and H&N data sets also included multiple (2 to 3) scans from the same set of 30 patients, which corresponded to scans obtained at different time points during sequential fractionated radiation treatment. One of those multiple scans corresponded to the pre-treatment time point. The remaining (1 or 2) scans represented post-treatment scans and hence often included anatomic changes due to radiation treatment.

Following the basic premise for this grouping investigation of model/atlas building only from near-normal data sets, we labeled each of the 644 (=346+298) scans into three categories – near-normal, intermediately abnormal, and abnormal – following radiological image appearance factors under the guidance of a radiologist (co-author Torigian) with ~20 years of experience. Our protocol for this classification was as follows.

Thorax. *Near-normal*: Scans with small or mid-size isolated lung nodules, or slight intensity changes. *Intermediately abnormal*: Scans with mid-size lung nodules adjacent to chest wall or other mid-size lesions. *Abnormal*: Scans with lesions leading to lack of aerated lung parenchyma affecting more than 1/6 of the lungs (i.e., of approximately more than one lung lobe), large lung lesions adjacent to the chest wall, significantly abnormal curvature of the thoracic spine (scoliosis), or unusual patient set up.

H&N. *Near-normal*: Scans with imaging findings such as small nodules in the salivary glands without affecting the shape of the glands, lack of teeth, and presence of streak artifacts due to beam hardening. *Intermediately abnormal*: Scans with the absence of one or more salivary glands or presence of a small mass without invasion of surrounding tissues. *Abnormal*: Scans with a large mass in the pharynx, larynx, esophagus, or salivary glands, surgery or neck mass causing distortion of anatomical structures in H&N region, large lymph nodes, or any large focal lesion involving the osseous structures. Besides a plastic mask, other immobilization devices such as a mouth-piece/bite-block were often used during scanning. Images with such immobilization devices were very few in our data cohort, and we excluded them since some anatomical structures can be distorted compared with images acquired without presence of those devices.

For testing our methods, we created several populations ($\mathcal{P}$) of images from the categories listed in Table 1 for each of the two body regions. These are summarized in Table 2.

2.2.2. *Trimming for body region* B

Following the body region definition, we removed axial slices that are outside B from each image in $\mathcal{I}$ manually. We will denote the resulting set of images by $\mathcal{I}_B$. Note that trimming is done only in the cranio-caudal direction and not within axial slices of acquired scans so that each acquired slice that is selected for inclusion in B is represented as is.

2.2.3. *Image registration*

Recall that the goal of this investigation is to find how many distinct groups of similar body habitus may exist in the population. As such, the goal of registration of images in $\mathcal{I}_B$ is to remove irrelevant deviations that arise in subject images from processes such as different ways of positioning the subjects in the scanner, or different image resolutions etc., and not inherent body habitus variations among subjects. Therefore, we use a 6-parameter (3 translations, 3 rotations) registration operation to register images in $\mathcal{I}_B$ among themselves. The method is a simplified version of the minimum spanning tree (MST) algorithm described in (Grevera et al., 2016). Briefly, the method works as follows. It sets up a complete weighted directed graph with N nodes in which each node is connected to all other nodes. The weight (cost) assigned to each directed arc (I, J), where $I, J \in \mathcal{I}_B$, is the cost value $\alpha(I^r, J)$ defined by

$$\alpha(I^r, J) = \frac{\sum_{v \in I \cup J} |I^r(v) - J(v)|}{|I \cup J|}, \tag{1}$$

where I^r denotes image I after it is registered to image J via a 6-parameter transformation. $I \cup J$ denotes the set of voxels in the union of the body region within the skin outer boundary of (unregistered) I with the body region within the skin outer boundary of J. In other words, this entity constitutes the union of the foreground regions in the two images. $|I \cup J|$ represents the number of voxels in the union region. The cost value $\alpha(I^r, J)$ denotes the *Mean Absolute Difference* (MAD) between images I^r and J. After the graph is set up, a *Minimum Spanning Tree* (MST) – a tree that spans the graph and has the least total cost – is found for the graph, and each image in $\mathcal{I}_B$ is registered to the root image of the MST.[6]

If the images are acquired with adequate alignment in the scanner, then it may not be necessary to perform the above global MST registration among all images in $\mathcal{I}_B$. We have tested both options (global and no preliminary registration) to study the behavior of the grouping process. We will refer to these two cases throughout as “MST registration” and “no MST registration”, respectively. Even when no registration is employed, we will denote the resulting image set by $\mathcal{I}_R$ for the next step. In other words, for this case, $\mathcal{I}_R = \mathcal{I}_B$.

[5] Our analysis considered female and male subjects together. See results in Section 3 and comments in Section 4 on gender differences vis-à-vis grouping analysis.

[6] Any other suitable method of registration can be chosen instead. See Section 4 for further comments.

Table 2
Populations $\mathcal{P}$ considered for analysis for thoracic and head and neck body regions.

Abbreviation	Description	N	
		Thorax	H&N
NN	Near-normal; includes with-contrast and no-contrast cases.	245	111
NA	Non-abnormal; includes all categories except abnormal.	284	157
NNWC	Near-normal, with-contrast.	73	75
NNNC	Near-normal, no-contrast.	172	36
NAWC	Non-abnormal, with-contrast.	77	111
NANC	Non-abnormal, no-contrast.	207	46
AN	Abnormal; includes with-contrast and no-contrast cases.	62	141

2.2.4. Similarity (cost) function for grouping

For the case with MST registration, we define the similarity (rather dissimilarity or cost) function $\varphi(I, J)$ for grouping using the following function on the registered images in $\mathcal{I}_R$.

$$\varphi(I,J) = \frac{\sum_{v \in I \cup J} |I(v) - J(v)|}{|I \cup J|}, \; I \text{ and } J \in \mathcal{I}_R. \tag{2}$$

For the case of no-MST-registration, for grouping we will use the cost function $\varphi(I, J) = \alpha(I^r, J)$ of Eq. (1). That is, in this case, although images in $\mathcal{I}_R$ will not be pre-registered among themselves, the cost is estimated via MAD between images after they are pairwise registered. Note that in the first case with MST registration, $\varphi(I, J) = \varphi(J, I)$, since $I, J \in \mathcal{I}_R$ but not so in the second case of no-MST-registration since $I, J \in \mathcal{I}_B$.

The main computational burden comes from estimating $\alpha(I^r, J)$ in Eq. (1) for all N^2-N pairs of images, irrespective of whether or not pre-registration is employed among all images in $\mathcal{I}_B$. Once these values are estimated, the actual MST process incurs minimal additional computational cost. For fast calculation of MAD values, we used images that are down-sampled by doubling the pixel size but not changing the slice spacing. We observed only around 0.8% difference in the MAD values computed from the original versus down-sampled images.

As an example, we display in Fig. 2 the cost function $\varphi(I, J) = \alpha(I^r, J)$ as a heat map for the combined population NA ∪ AN involving all 346 scans for the thoracic body region (see Table 1). There is an interesting pattern in the lower right corner of the diagonal with larger low-valued patches. They correspond to patients with multiple scans obtained during their serial fractionated radiation treatment mentioned earlier.

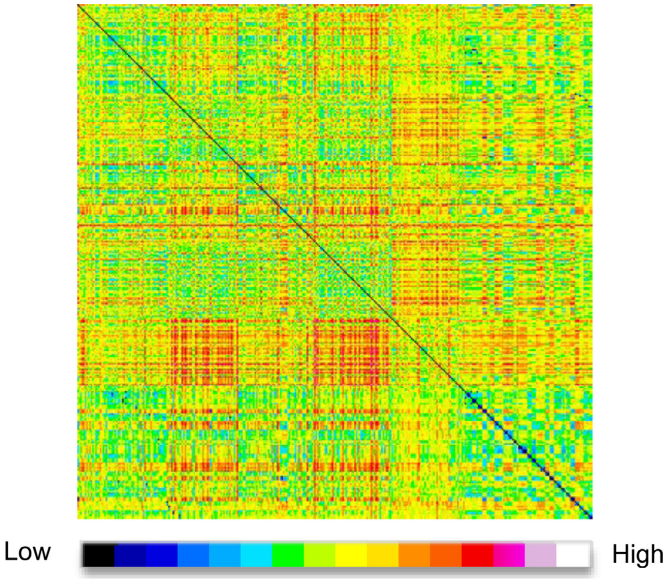

Fig. 2. Heat map display of the cost matrix $\alpha(I^r, J)$ for the combined population defined by NA ∪ AN, consisting of all 346 thoracic scans used in this paper. Note that the cost matrix is asymmetric around the all-zero principal diagonal.

2.3. Building potential groupings (finding partitions $\mathrm{P}_n(\mathcal{I}_R)$)

We employ a bottom-up strategy for grouping, called hierarchical agglomerative clustering (HAC) (Fernández and Gómez, 2008) with a complete-linkage strategy. This is a common tool utilized in data mining and statistics. Given the cost function $\varphi(I, J)$ defined in the previous section, the HAC strategy starts by defining N clusters (groups), each containing a single image of $\mathcal{I}_R$. It then iteratively merges pairs of clusters most eligible for merging based on "distance" $d(U, V)$ between two clusters U and V, until all images of $\mathcal{I}_R$ are merged into a single cluster/group. The procedure will produce a sequence of partitions $P_n(\mathcal{I}_R)$, $n = 1, \ldots, N$, we are seeking. We will explain the reason for selecting the complete-linkage method in comparison with another strategy in Section 3.1.

Our hierarchical grouping procedure, named HG, is presented below. It uses a distance function $d(U, V)$ on $P_n(\mathcal{I}_R) \times P_n(\mathcal{I}_R)$ to ascertain the closeness of clusters U and V of $P_n(\mathcal{I}_R)$ for determining if U and V should be merged during the iterative process. This distance function is defined as follows.

$$d(U,V) = max\,[\varphi(I,J) : I \in U \,\&\, J \in V]. \tag{3}$$

Procedure HG

```
Input: Image set $\mathcal{I}_{RR} = \{I_1, \ldots, I_N\}$ for body region $B$ and population $\mathcal{P}$; cost
        function $\varphi(I, J)$ on $\mathcal{I}_{RR} \times \mathcal{I}_R$.
Output: $P_n(\mathcal{I}_R)$, $n = 1, \ldots, N$.
Begin
    S1. Set $P_N(\mathcal{I}_R) = \{\{I_1\}, \ldots, \{I_N\}\}$; $n = N$; $d_M\,(U, V) = d(U, V)$;
    S2. Find the pair of "clusters" $(X, Y)$ among those in $P_n(\mathcal{I}_R)$ with the
        smallest distance
            $(X,Y) = \overset{argmin}{\scriptstyle (U,V)}[d_M(U,V) : U,V \in P_n(\mathcal{I}_R)]$;
    S3. Merge $X$ and $Y$ and update $P_n(\mathcal{I}_R)$:
        $n \leftarrow n-1$ and replace $X$ and $Y$ by $Z = X \cup Y$ in $P_n(\mathcal{I}_{\mathbf{R}})$;
    S4. If $n > 1$ then
    S5. Update distances $d_M\,(Z, W)$ of the new cluster $Z$ to other clusters in
        $P_n(\mathcal{I}_R)$ using complete linkage: $d_M(Z,W) = max\,[d(X,W), d(Y,W)]$,
        for all $W \in P_n(\mathcal{I}_R)$;
    S6. Go to S2;
    S7. Else output $P_n(\mathcal{I}_R)$, $n = 1, \ldots, N$;
End
```

In Step S1, HG is initialized with N clusters (groups). The cluster distances initially are just the cost values between pairs of individual images. In S2, the pair of clusters with the smallest distance among all pairs is selected and merged into a larger cluster in S3. The number of clusters now decreases by 1. The distances $d_M\,(Z, W)$ between the new cluster Z and all other clusters W are then updated in S5. This is accomplished by using the complete-linkage strategy where the distance between any two clusters is the largest pair-wise distance between images in the respective clusters, one image taken in one cluster and the other in the second cluster as depicted in Eq. (3).

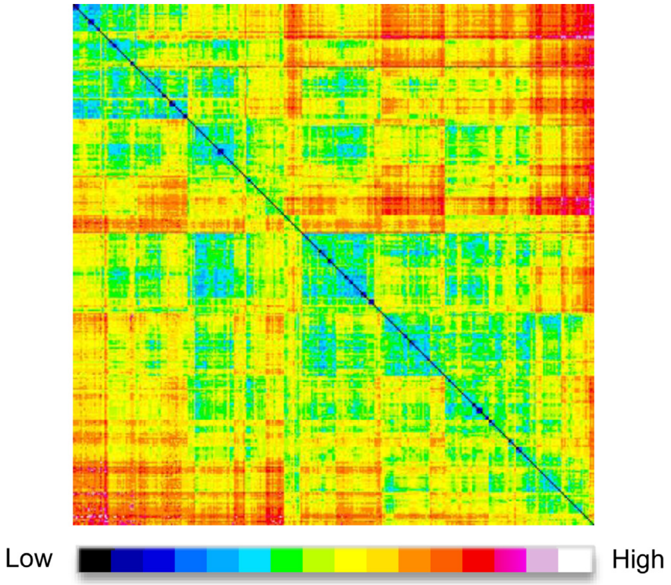

Fig. 3. Heat map display of the cost matrix for the same population as in Fig. 2 for all 346 thoracic scans but after grouping is performed and cost values were sorted out via HG. Groups (clusters) formed can be seen along the diagonal.

The heat map shown in Fig. 2 is displayed in Fig. 3 after grouping is performed via HG and the cost values are sorted as per the sequence of images in the final partition $P_1(\mathcal{I}_R)$ produced by HG. Groups formed can be visualized along the diagonal from very small groups consisting of just one image to larger groups consisting of many images. To understand how the cost-values become sorted along the horizontal and vertical axes, consider an example with $\mathcal{I}_R = \{I_1, \ldots, I_6\}$. Let the grouping formed by HG in successive iterations be as follows. $P_6(\mathcal{I}_R) = \{\{I_1\}, \ldots, \{I_6\}\}$, $P_5(\mathcal{I}_R) = \{\{I_1\}, \{I_2\}, \{I_3, I_6\}, \{I_4\}, \{I_5\}\}$, $P_4(\mathcal{I}_R) = \{\{I_1\}, \{I_2, I_4\}, \{I_3, I_6\}, \{I_5\}\}$, $P_3(\mathcal{I}_R) = \{\{I_1, I_5\}, \{I_2, I_4\}, \{I_3, I_6\}\}$, $P_2(\mathcal{I}_R) = \{\{I_1, I_5, I_2, I_4\}, \{I_3, I_6\}\}$, and $P_1(\mathcal{I}_R) = \{\{I_1, I_5, I_2, I_4, I_3, I_6\}\}$. In this case, the horizontal axis will have images from left to right in the order given by $P_1(\mathcal{I}_R)$, namely, I_1, I_5, I_2, I_4, I_3, and I_6. Similarly, the vertical axis will have the same order from top to bottom. This order determines the appearance of the heat map.

2.4. Determining optimum number of groups

For evaluating the quality of the resultant groupings and finding the optimum number(s) of groups, we define a measure called *Residual Dissimilarity* (*RD*), denoted $\gamma(n)$, to express the quality of groupings as a function of the number n of groups.

$$\gamma(n) = \frac{\sum_{i=1}^{n} \mu(P_n^i(\mathcal{I}_R))\sigma(P_n^i(\mathcal{I}_R))}{n}, \tag{4}$$

where $P_n^i(\mathcal{I}_R) \in P_n(\mathcal{I}_R) = \{P_n^1(\mathcal{I}_R), \ldots, P_n^n(\mathcal{I}_R)\}$, and $\mu(P_n^i(\mathcal{I}_R))$ and $\sigma(P_n^i(\mathcal{I}_R))$ respectively denote the mean and standard deviation of MAD values $\varphi(I, J)$ of all possible pairs of images (I, J), $I \neq J$, in$P_n^i(\mathcal{I}_R)$. In words, $\gamma(n)$ expresses (in units of [Hounsfield Unit]^2 or HU^2 in our case of CT images) our desire to have the dissimilarity remaining in each of the n groups in $P_n(\mathcal{I}_R)$ as small as possible by keeping both mean and standard deviation of MAD values for each group in $P_n(\mathcal{I}_R)$ as small as possible. The gain in similarity or reduction in cost by allowing an additional group each time within the iterative loop from S2 to S6 in HG can be evaluated by the derivative of $\gamma(n)$ with respect to n. We refer to values of n at which large changes take place in $\gamma(n)$ as *breakpoints*. From the manner in which RD-values change, these breakpoints offer a guidance for selecting the number of groups. See Figs. 4, 7, and 11 in the next section.

3. Experiments, results, and discussion

We have analyzed the grouping behavior for each of the populations listed in Table 2 separately for each body region and for the cases with and without MST registration. We will examine one (largest) population for each body region closely and summarize results for other populations. In addition, we will examine how groupings may be affected by the choice of a different clustering method/distance function for one body region, namely Thorax.

3.1. Thorax

Fig. 4 demonstrates the relationship between $\gamma(n)$ and the number of groups n for different populations listed in Table 2 for the thoracic populations. The figure contains considerable amount of information that we will further analyze below. We will also examine the grouping behavior vis-à-vis the clustering method/similarity function $\varphi(I, J)$.

(1) *No MST registration*: (a) Grouping results (the RD-value $\gamma(n)$) for populations NN, NA, NNNC, and NANC are very similar until $n = 50$ groups, which demonstrates that no additional distinctive anatomic forms are added to the population beyond the NNNC population (size = 172, see Table 2) by adding more non-abnormal images. Therefore, following our tenet of modeling only non-abnormal images (Udupa et al., 2014), ~150 images (subjects) seem to be sufficient to cover all important variations from the perspective of modeling.

This last point is further illustrated in Fig. 5 where we demonstrate how the mean value,

$$M(P_n(\mathcal{I}_R)) = \frac{1}{n}\sum_{i=1}^{i=n} \mu(P_n^i(\mathcal{I}_R)), \tag{5}$$

of $\mu(P_n^i(\mathcal{I}_R))$ over all n groups of $P_n(\mathcal{I}_R))$ varies as a function of n as we change the number of subjects N in the thoracic population NANC (see Table 2). When N is small, there are uncertainties at the beginning in forming small clusters. When large clusters are formed, similarities among samples within clusters become increasingly more stable. Starting with $N = \sim 150$ images, the clusters become stable as such there is no significant difference in the grouping result by adding more images; notice how close the curves are for $N = 150$, 175, 200, and 207, especially when $n < 50$.

(b) The grouping results for "with-contrast" (WC) and "no-contrast" (NC) cases seem to behave differently. See curves for NAWC and NNWC versus NANC and NNNC. (Curves for NAWC and NNWC are almost identical and overlap in Fig. 4 since NAWC included all 73 NNWC subjects with only 4 additional intermediately abnormal cases.) Note that although NAWC as a population on its own shows a different behavior from NANC, within the larger population NA, the difference between NC and WC does not seem to matter, as seen from our observations in (a). This suggests that we need to first define and fix the population $\mathcal{P}$ of our application for which we are seeking an answer to the main question addressed in this paper and then analyze the grouping behavior. (c) Population AN is the smallest ($N = 62$) among all $\mathcal{P}$s, and therefore it may not be large enough to make generalizable observations, especially in view of the above observations in (a) from Fig. 5. Its RD-vales are higher than those for the NN populations, notably, when $n \leq 16$.

(2) *With MST registration*: (a) For populations NN, NA, NNNC, and NANC, their RD-values $\gamma(n)$ are similar from 10 to 50 groups, analogous to the behavior we observed for the no-MST-registration case in (1). Their behavior now seems to be more similar among

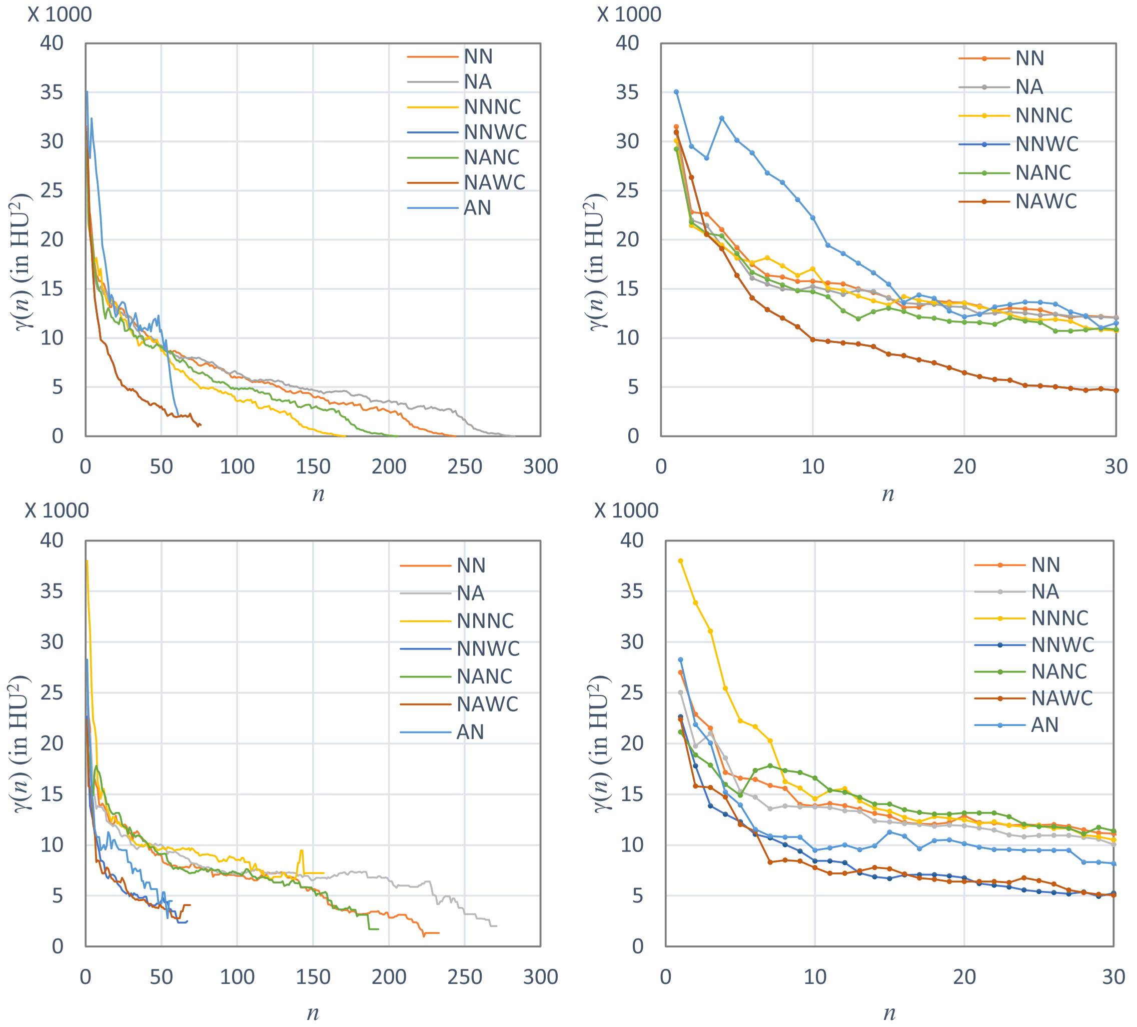

Fig. 4. Left: Plots of RD-value $\gamma(n)$ as a function of n for the 7 thoracic populations listed in Table 2. Right: A close up view of the plot around the rapidly decreasing portion. 1st row: No-MST-registration. 2nd row: With-MST-registration.

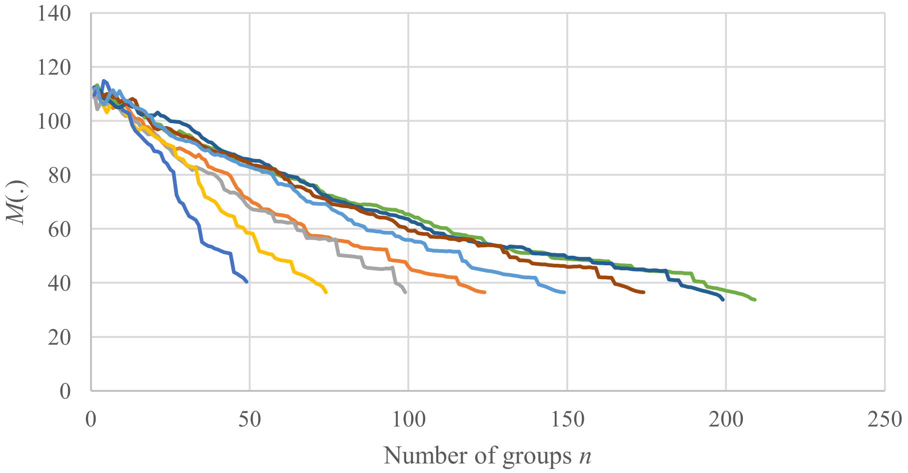

Fig. 5. Mean value $M(.)$ (Eq. (5)) denoting the mean, over all groups, of mean MAD values within groups for different number n of groups for different sizes $N = 50$, 75, …, 200, 207 of the thoracic population NANC.

themselves than in (1), perhaps owing to registration. (b) The RD-values for NNWC and NAWC, which were indistinguishable in (1), are now slightly separated. We suspect that this difference may have been caused by the different root subjects employed in MST registration for the two populations. (c) Interestingly the RD-values of AN seem to have improved due to registration, but we still maintain that N is too small to draw any conclusions on this population.

(3) *With or no-MST-registration*: The overlapping curves for the different populations indicate that NC/WC and NN/IA do not cause significant differences in grouping. By examining images for population NA in different groups, we observed that with-

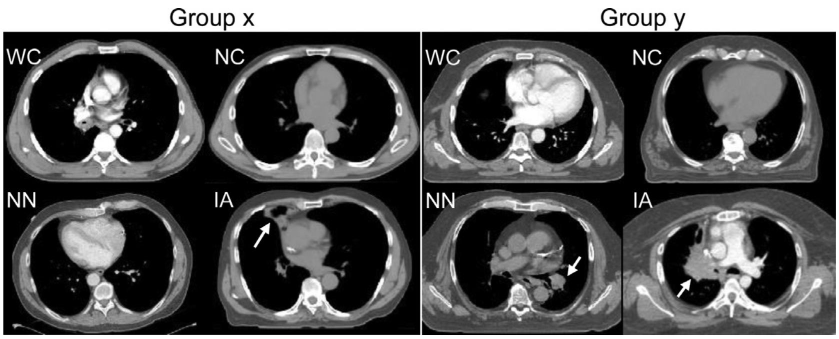

Fig. 6. Exemplar slices from images in population NA belonging to different groups illustrating our observation that grouping occurs by similarity of anatomic form rather than with-contrast/no-contrast or degree of abnormality (near-normal/intermediately abnormal). Images in the left and right blocks of columns come from different groups. Other image characteristics are indicated by labels: WC = with-contrast, NC = no-contrast, NN = near-normal, IA = intermediately abnormal. Abnormal factors are indicated by arrows in images.

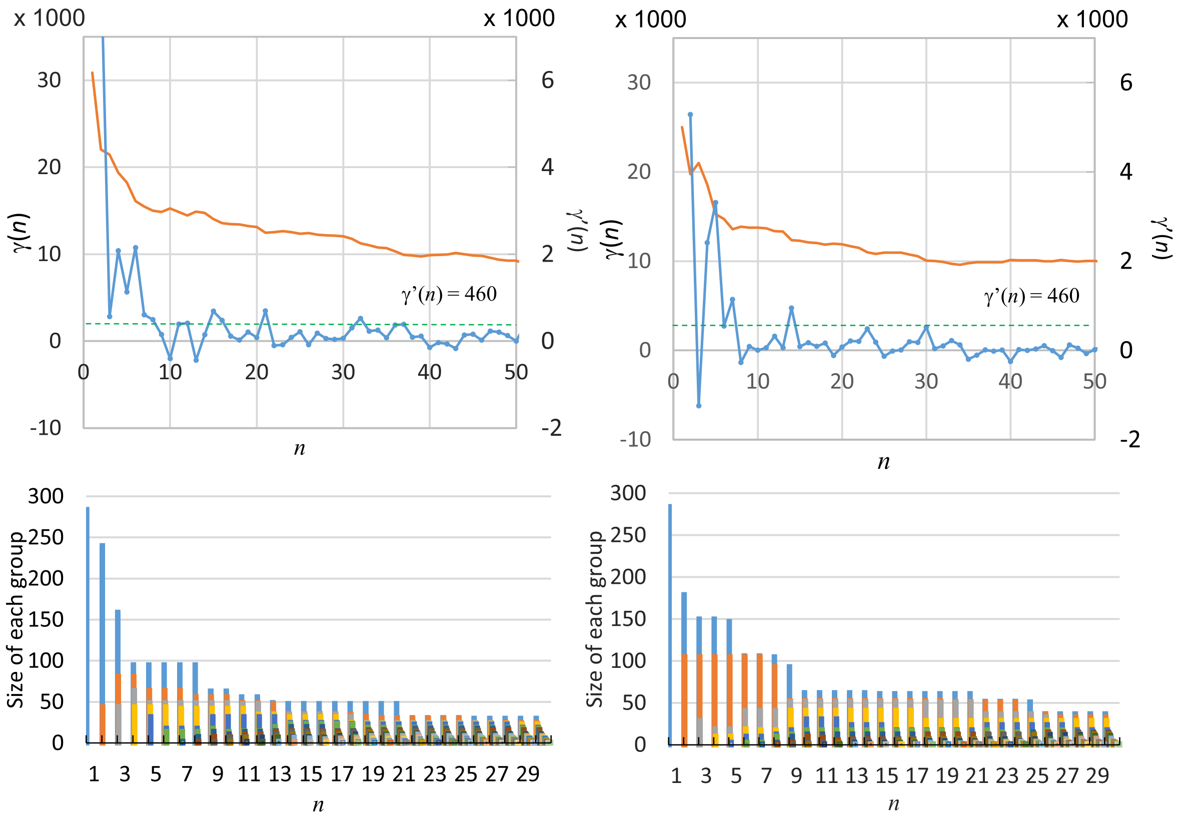

Fig. 7. Top: $\gamma(n)$ curve for the thoracic population NA. Bottom: Histogram of MAD values for each n for population NA. Left: Without MST registration. Right: With MST registration.

contrast images and images with intermediate abnormalities are distributed in different groups with other near-normal no-contrast images with substantially similar anatomical form, as illustrated in Fig. 6.

To study the grouping behavior more closely, we examined $\gamma(n)$ and its breakpoints for population NA. In Fig. 7 (top), we display $\gamma(n)$ and its derivative $\gamma'(n)$ (expressed as $\gamma(n\text{-}1)$ - $\gamma(n)$) for NA, where the *breakpoints* are defined as values of n at which $\gamma'(n) \geq$ 460 HU^2/group; the threshold is indicated by a dashed threshold line.[7] For the case of no-MST-registration, breakpoints in $\gamma(n)$ are observed for NA at $n = 2$, 3, 4, 5, 6, 7, 8, 15, 21, and 32. At these points, the corresponding $\gamma'(n)$ values are 8856.7, 564.2, 2076.1, 1130.3, 2150.3, 606.1, 494.3, 690.6, 695.6, and 521.8 HU^2/group, respectively. The corresponding values for the case with MST registration are: $n = 2$, 4, 5, 6, 7, 14, 22, and 30, and $\gamma'(n) = 5286.9$, 2413.6 3314.1 546.5, 1141.8, 949.8, 483.3, and 519.0, HU^2/group, respectively.

At the bottom of Fig. 7, we demonstrate qualitatively how the groups split as n increases and how the histogram of MAD values changes with n. A rapid drop in $\gamma(n)$ can be obtained at the point when the largest group splits into two groups (as at n from 1 to 2, 2 to 3, 3 to 4, etc., for the no-MST registration case and from 1 to 2, 2 to 3, 5 to 6, etc., for the case of MST-registration). Conversely, describing in the manner of progression of procedure HG, $\gamma(n)$ would increase significantly when two groups merge into the largest cluster. The continuing benefit in similarity among images within groups at $n = 8$ for no-MST-registration and $n = 7$ for MST-registration seems essential. Points at 15, 21, 32 for no-MST

[7] This threshold of 460 HU^2/group is set up based on our observation of the behavior of different populations in the two body regions. If this threshold is increased, the number of breakpoints will decrease and vice versa if it is decreased.

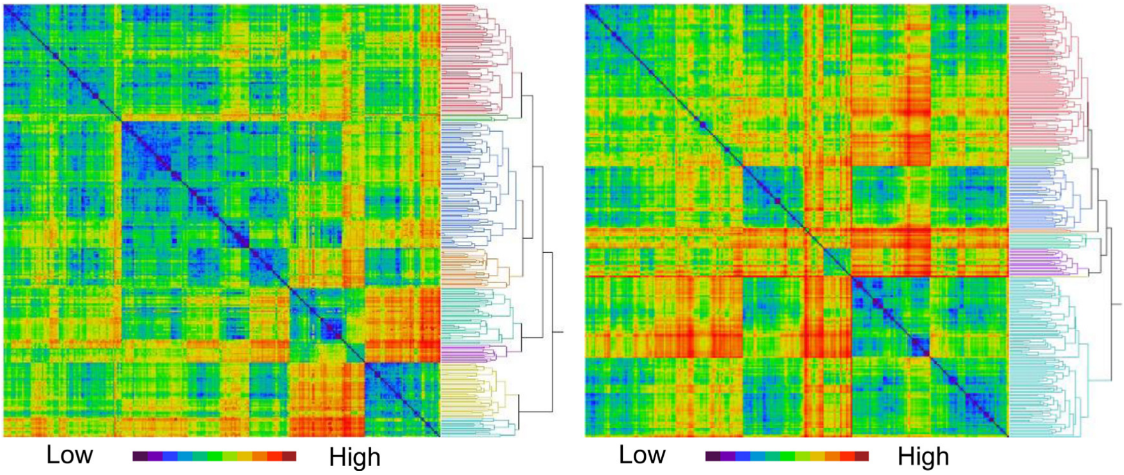

Fig. 8. Heat map and dendrogram of sorted MAD ($\varphi(I, J)$) matrix for the thoracic population NA at $n = 8$ for no-MST-registration (left) and MST-registration (right). Groups can be identified on the dendrogram sections with different colors and corresponding blocks appearing on the diagonal of the matrix.

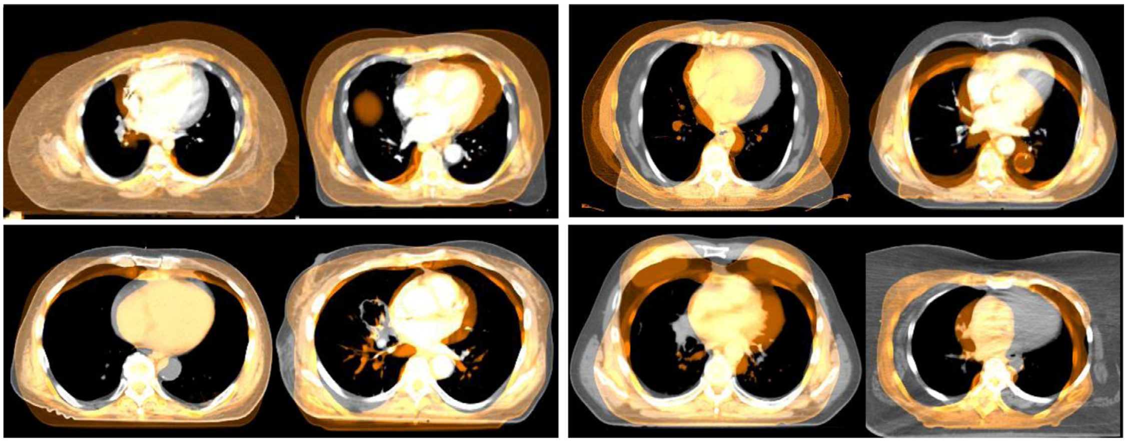

Fig. 9. Randomly selected pairs of images at an axial level passing through the heart – within the same group (left) and from different groups (right), at $n = 8$ for the thoracic population NA. Top: No-MST-registration. Bottom: MST-registration. Images in a pair are overlaid, one of them in color and the other as gray image.

registration and 14, 22, 30 for MST-registration cases show potential optimum number of groups if a finer grouping is needed for both cases. Notably, these grouping results in terms of the optimum number of groups to consider are very similar for the two cases of pre-registration of images.

The partitions $P_n(\mathcal{I}_R)$ can be arranged into a dendrogram, which can be utilized to learn how the elements of the partitions in the sequence are related in terms of images shared. The parcellation of images by their similarity is illustrated from this angle in Fig. 8 via a heat map and a dendrogram. The figure illustrates grouping at $n = 8$ for no-MST-registration and MST-registration. Groups can be identified by their different colors in the dendrogram display. The corresponding groups can also be recognized as blocks along the diagonal. Note some small groups, such as the group represented in green (no-MST-registration, in the upper part) and orange (MST-registration, in the middle), in the dendrogram section.

To illustrate how well images may match within each group, we randomly selected and overlaid pairs of images within and across groups as portrayed in Fig. 9 for no-MST registration (top row) and MST-registration (bottom row). In both cases, the overlaid images in the pair are registered, in the first case pair-wise for the computation of $\varphi(I, J)$, and in the second case by the MST method. Notice that all images seem to align properly (as best as possible) anatomically because of pairwise registration irrespective of whether they are from the same group (left column) or different groups (right column). However, when image pairs from the same group are overlaid, they seem to also match in other aspects of body habitus while images from different groups seem not to satisfy this property although they seem properly aligned because of registration. To study the different anatomic forms by which the images may have grouped, we examined the images in each group for the $n = 8$ case for NA. We randomly selected four members from each group, two at an axial level passing through the heart and two at the level in the vicinity of the carina. These images are displayed in Fig. 10 in an 8×4 matrix for the case of no-MST-registration. Some common anatomical forms (or aspects of body habitus) within each group can be recognized. For example, Group 4 is composed of *only* female subjects with large breasts. The grouping behavior was similar for the case of MST-registration.

Finally, in Table 3, we summarize the breakpoints obtained with $\gamma'(n) \geq 460$ and the corresponding $\gamma'(n)$ values for all thoracic populations for the two cases of pre-registration. Although the breakpoints are understandably not identical among populations, there is a clear pattern that can be observed in these groupings. There are generally four stages for the optimum number of groups: (i) $n = 5$ to 10 seem to be the essential number of groups; (ii) 11 to 15 constitute the second finer stage; (iii) 16 to 24 represent the third stage; and (iv) 28 to 35 denote the finest grouping stage. A similar phenomenon can also be observed in Fig. 7 during the initial turbulent behavior of $\gamma(n)$ appearing as sharp oscillations.

(4) *With a different grouping method/similarity (cost) function:* To study how and if the grouping behavior would change com-

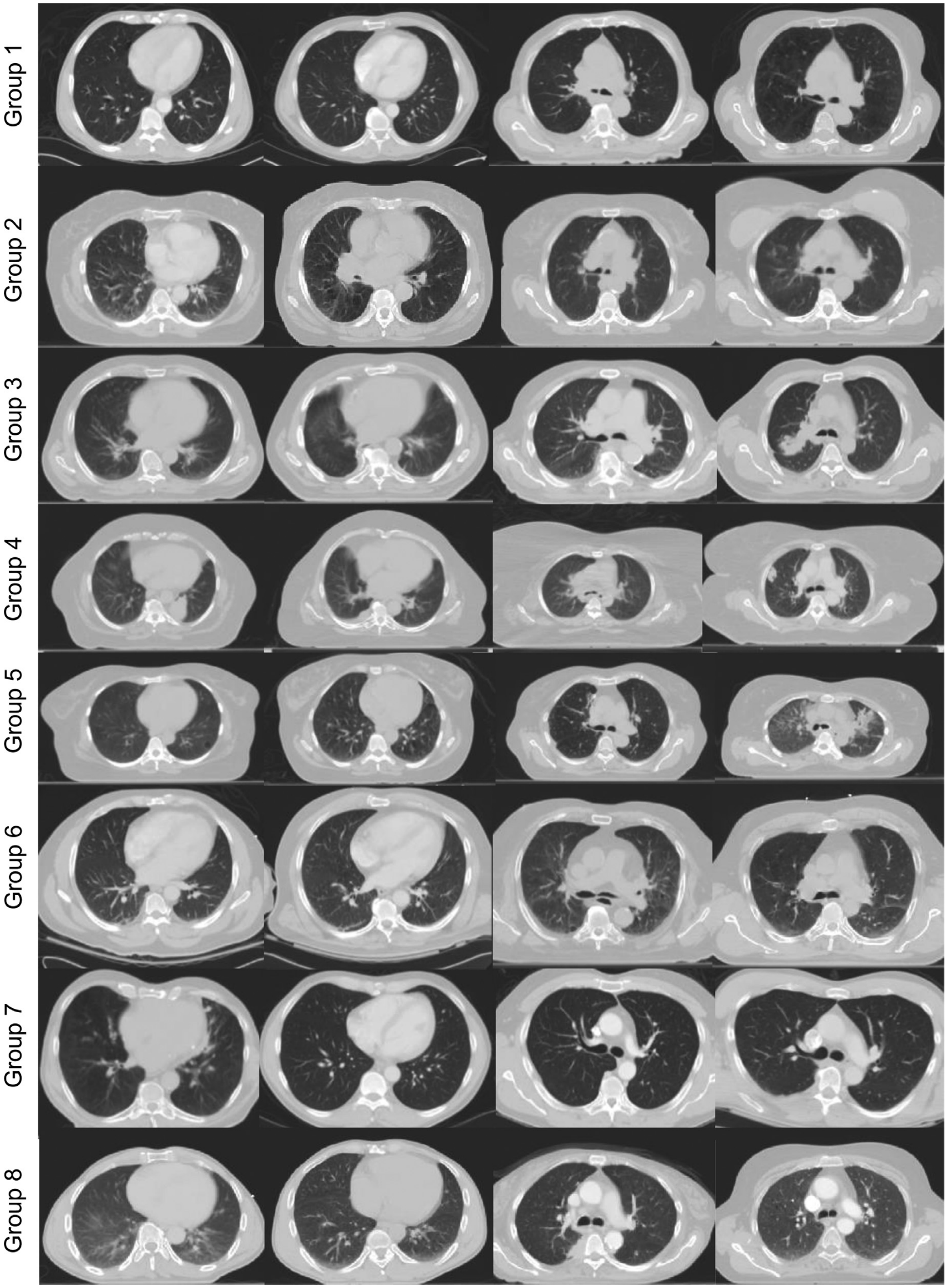

Fig. 10. Randomly selected images, two at an axial level passing through the heart (left two columns) and two at the level of the carina (right two columns), for each of the 8 groups for the case of $n = 8$ for the thoracic population NA. Grouping is done with no-MST-registration.

Table 3
Breakpoints found for the different thoracic populations and their corresponding $\gamma'(n)$ values for the two cases of pre-registration.

	No-MST-registration		MST-registration	
	Breakpoints n	$\gamma'(n)$	Breakpoints n	$\gamma'(n)$
NA	8, 15, 21, 32	494.3, 690.6, 695.6, 521.8	7, 14, 22, 30	1141.8, 949.8, 483.3, 519.0
NN	7, 13, 16, 31	1103.9, 499.1, 1004.9, 477.1	5, 7, 9, 16, 21, 31	564.0, 597.0, 1581.1, 631.7, 657.6, 610.5
NNWC	10, 15, 20, 24	1315.1, 777.5, 514.0, 528.5	6, 10, 13, 21	1230.5, 973.6, 1024.4, 556.8
NNNC	6, 9, 11, 14, 23, 28, 35	480.4, 979.6, 1946.3, 485.4, 478.8, 685.7, 657.8	10, 14, 16, 28	1059.0, 717.8, 605.2, 683.9
NAWC	10, 15, 20, 24	1315.1, 777.5, 514.0, 528.5	7, 11, 16, 27	3042.6, 556.7, 504.3, 593.1
NANC	9, 13, 17, 26, 32, 39	593.0, 818.4, 571.7, 865.5, 617.1, 686.2	5, 8, 11, 14, 16, 24, 28, 31	1063.7, 478.4, 1187.3, 669.0, 556.9, 736.5, 536.6,532.2
AN	3, 16, 20, 27, 29, 35, 37	1195.8, 1866.9, 591.0,792.4, 1218.8, 1074.7,598.7	7, 10, 13, 29, 32, 38	652.8, 1287.7, 467.1, 1170.0, 986.6, 1096.2

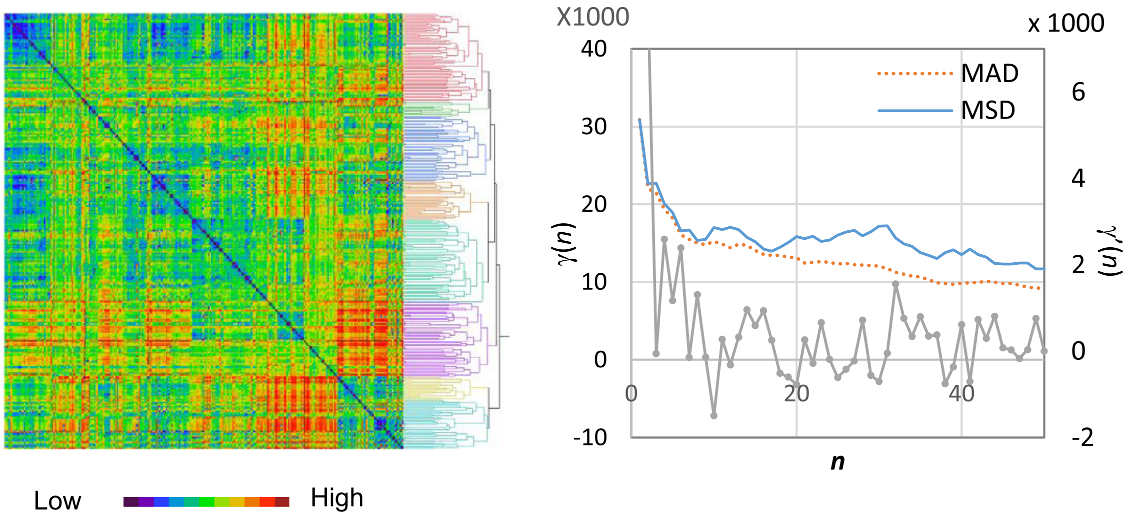

Fig. 11. Left: Heat map and dendrogram of sorted MSD ($\varphi_1(I, J)$) matrix for the thoracic population NA at $n = 8$ for no-MST-registration. Groups can be identified on the dendrogram sections with different colors and corresponding blocks appearing on the diagonal of the matrix. Compare this heat map with that in Fig. 8 on the left. Right: $\gamma(n)$ curve for the thoracic population NA without MST registration and its derivative $\gamma'(n)$ curve. The curve obtained with MAD is also shown for reference. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

pared to the above results from Procedure HG utilizing the similarity function $\varphi(I, J)$ formulated in Eq. (2), we implemented another similarity function $\varphi_1(I, J)$ defined below based on mean squared differences (MSD).

$$\varphi_1(I,J) = \frac{\sum_{v\in I\cup J}[I(v)-J(v)]^2}{|I\cup J|},\ I \text{ and } J \in \mathcal{I}_R. \tag{6}$$

We show in Fig. 11 results from the MSD method for the NA population using no-MST registration. Both the heat map with the dendrogram and the $\gamma(n)$ curve are displayed. In order to compare the grouping results, we scaled the $\gamma(n)$ curve of MAD method by the ratio of $\gamma(1)$ values to the $\gamma(1)$ value of our proposed method. Obviously, when $n = 1$, there is only one cluster, and the measure of similarity will be the same. Comparing Fig. 11 to the left side of Fig. 8, we notice that the heat map of the MSD method is more noisy, but otherwise very similar to that in Fig. 8. Meanwhile the grouping results are similar as per dendrograms. Compared to Fig. 7, the RD curve of the MSD method, although rougher, also behaves similarly to the curve of the MAD method. The fast drop in RD value occurs at the beginning of the curve. Breakpoints are obtained at $n = 2$, 4, 6, 8, 15, 23, 28, and 32, which are close to the results from the MAD method.

MSD has the property of penalizing large differences more. Large MSD values may be caused by noise or real local differences. It may be argued that this cost/similarity function is less appropriate for our purpose of body-habitus-based grouping. However, the two cost functions produce similar grouping results.

We also implemented another grouping method, the Ward's method (Ward, 1963), for comparison. Ward's method is designed to minimize variance within groups. Similar to our HG approach, it starts with all singletons, each containing a single image of $\mathcal{I}_R$. The initial cluster distances in this method are defined by $\varphi(I, J)$. At each step, *a pair of clusters that leads to minimum increase in total within-cluster variance* is merged into a new cluster for the next recursive step until only one cluster remains. The results for the thoracic population NA are shown in Fig. 12.

We have chosen the complete-linkage method for HG since it guarantees that, with any number of groups, pairs of images within a group will always have higher similarities than any pairs of images selected between two different groups, implication being that

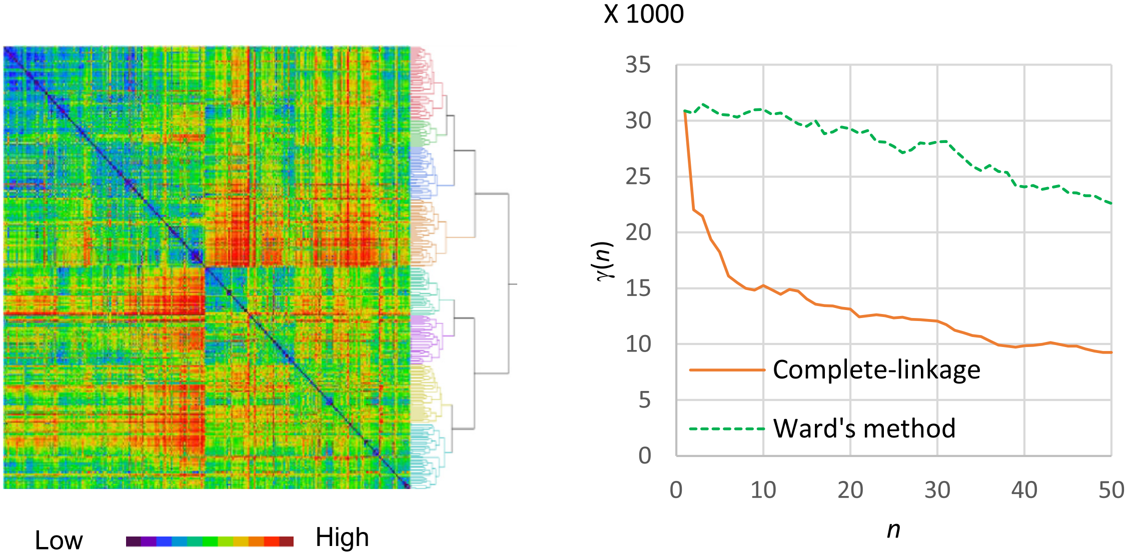

Fig. 12. Left: Heat map and dendrogram of sorted MAD ($\varphi(I, J)$) matrix for Ward's method for the thoracic population NA at $n=8$ for no-MST-registration. Groups can be identified on the dendrogram sections with different colors and corresponding blocks appearing on the diagonal of the matrix. Compare this heat map with that in Fig. 8 on the left. Right: $\gamma(n)$ curve for the thoracic population NA without MST registration for Ward's method. The curve for HG is also shown for reference. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

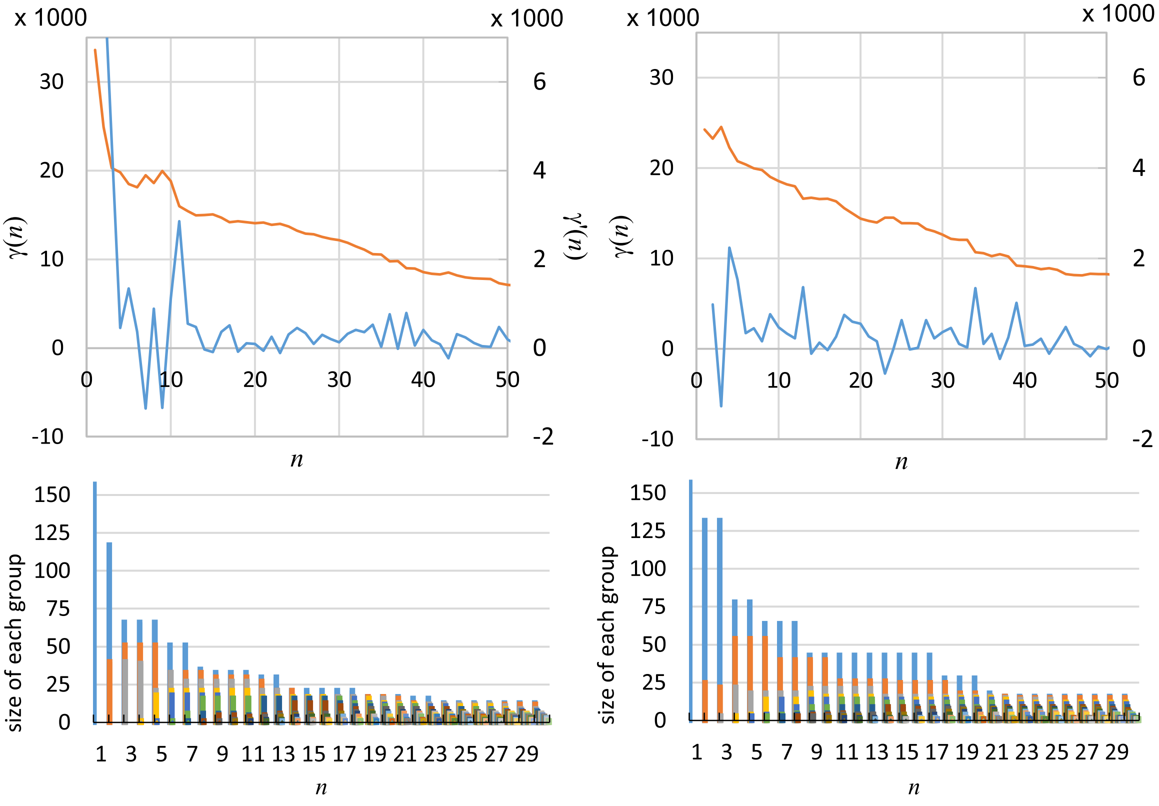

Fig. 13. Top: $\gamma(n)$ curve for the head and neck population NA. Bottom: Histogram of MAD values for each n for population NA. Left: No-MST-registration. Right: MST-registration.

the resulting clusters will be as compact as possible. We make the following observations from Fig. 12 in comparison to Figs. 7 and 8 showing results for HG. Both Ward's and the HG method are able to provide relatively reasonable clusters for the population (see the dendrograms). Comparing with the Ward's method, HG gives clusters that are more compact, meaning that the similarities within the clusters are higher (lower cost). Ward's method, while minimizing the variance, tends to equalize cluster sizes, causing hardship in throwing out outlier clusters such as the second cluster in the HG method. A fast drop in cost could be obtained in HG method compared to Ward's method at $n < 10$. Note that we designed $\gamma(n)$ for showing dissimilarity remaining keeping both mean and standard deviation of MAD/MSD values for each group in $P_n(\mathcal{I}_R)$. Considering that the complete-linkage method maximizes similarity within clusters, while Ward's method minimizes variance within clusters, the comparison shows that maximizing similarity is superior to minimizing variance in clustering for the goals of this study.

3.2. Head and neck

As seen from Table 2, the largest H&N population we have is NA with $N = 157$, which is much smaller than the 4 largest populations for thorax. Since other H&N populations are even smaller, we conducted our analysis only for NA.

Analogous to Fig. 7 for Thorax, we present H&N grouping results for NA in Fig. 13. We make the following observations from these results and other analyses we carried out similar to those for thoracic populations. (i) As in thoracic populations, grouping seems to be directed more by anatomic form than by other factors such as with-contrast/no-contrast and abnormalities. (ii) Grouping results for $n \geq 20$ seem identical for the two cases of pre-registration. For $n \in [2, 20]$, large groups “collapse” faster for the no-MST registration method. (iii) With a threshold value $\gamma'(n) \geq 456$ breakpoints are observed at $n = 5$, 8, 13, 17, 25, 34, 36, and 38 with the corresponding $\gamma'(n) = 1342.4$, 883.8, 477.4, 515.1, 456.3, 527.8, 758.3, and 787.7 for no-MST-registration and at $n = 2$, 5, 10, 13, 20, 25, 28, and 34 with $\gamma'(n) = 983.4$, 1533.1, 477.6, 1362.7, 556.2, 639.7, 638.8, and 1354.2 for the case of MST-registration. (iv) As in Thorax, the optimum number of groups can be divided into 4 stages depending on the requirements of the application: 5 to 9, 10 to 17, 18 to 20, and 22 to 35. Improvement beyond $n = 39$ appears to be negligible.

3.3. Computational considerations

Program execution times are estimated on a ThinkCenter computer with the following specifications: Quad-core AMD A10-6700 3.7 GHz CPU with 16GB RAM and running a Linux system. A point to note is that the computational time here pertains to the atlas-building step in an image analysis pipeline, so the task for which this computation needs to be carried out is called for very infrequently, and hence the computational cost really does not matter as long it is within practical limits. In the proposed approach of determining the appropriate number of atlases to build for the image analysis task at hand, computational time for image registration becomes the main bottleneck irrespective of whether grouping is performed with or without MST registration. The computational expenses of all other operations (estimating cost function values, finding MST, performing grouping, etc.) are insignificant. Both strategies require pair-wise image registration for all N^2-N possible pairs of images in a given population $\mathcal{P}$. This operation takes roughly 40 s per pair, and for the largest population we considered with $N \approx 300$, it takes ~1000 hrs in a purely sequential implementation. Obviously, the required operations can be highly parallelized, easily at the stage of running on multiple computers and multiple cores and, with more effort, on GPUs.

4. Concluding remarks

In this paper, we sought answers to two basic questions that arise in multi-model/multi-atlas-based approaches to image analysis as to how many models/atlases and how many images are needed for properly representing prior information. Although these questions arise in several image analysis applications, our main application motivation is body-wide object recognition and delineation, and the empirical results shown have this focus. Starting by clearly defining the body regions and the subject populations under consideration, we translated the problem into one of finding optimal groupings in a given, sufficiently large, and representative set of images for each body region and population of interest. The optimal groupings were arrived at by using a greedy clustering strategy to first create different numbers of groups, then by defining a quality metric for the groups to examine how this metric varied as a function of the number of groups, and finally by detecting breakpoints in this function. Our main conclusions from studying several populations of images for thorax and head and neck body regions via CT can be summarized as follows.

(1) A minimum of 150 images from different subjects in a population seems essential to cover the anatomical variations for a given body region. Additional subjects will not affect the grouping significantly.

(2) A minimum of 5 to 8 groups (or models/atlases) seems essential to properly capture information about differing anatomic forms and body habitus. A larger number may be desirable if finer discrimination needs to be made as required by the application at hand. In any case, a number larger than ~35 seems unnecessary.

(3) Images seem to group predominantly by anatomic form and body habitus rather than by image variations due to presence/absence of induced contrast or the extent of abnormality as long as the pathology is not extensive.

(4) Drastically different or outlier groups are sorted out first in the grouping process and the largest group then splits into multiple groups depending on the grouping quality (residual dissimilarity) sought after.

(5) The performance of grouping for the largest population NA considered was similar for the two body regions studied in this paper with similar threshold on residual dissimilarity values (~460) and similar resulting breakpoints. The 4 granularity levels of the groupings found for the two body regions are also analogous with minor differences.

(6) Image registration-related issues: In place of our MST-registration method, any other global method to register images among themselves within a population $\mathcal{P}$ can be used. For example, 7-parameter registration (3 translations, 3 rotations, 1 isotropic scaling) could provide a scale-invariant registration result, leading to scale-invariant grouping. This is a matter of choice as to whether different sizes should/should not be considered within the realm of different body habitus conditions. However, more flexible deformable registration methods will certainly affect the body habitus itself, will therefore encroach into the precept of grouping itself, and hence should be avoided. Since the results for the two cases of no-MST and MST registration were similar, our study suggests that global registration in the spirit mentioned above may not be needed. Even for the no-MST-registration case, the HG process performs pair-wise image registration to define similarity function $\varphi(I, J)$. An interesting question is what if we eliminate even this pair-wise registration. For images that are acquired with adequate alignment in the scanner, we may define the similarity function without incurring registration via $\varphi(I, J) = \alpha(I, J)$ for $I, J \in \mathcal{I}_B$. This will lead to a huge gain in computational savings. However, we suspect that this may unnecessarily increase the optimum number of groups influenced by the vagaries of placing subjects in the scanner.

As to question Q2, our results from the largest studied thoracic populations (NA, NN, NANC, and NNNC) suggest that $N \approx 150$ is sufficient to observe the most important variations within a population. As pointed out earlier, the given image set $\mathcal{I}$ should contain images that are *representative* of the population $\mathcal{P}$ being studied. If $\mathcal{P}$'s scope is broad, for example, involving all races, ethnicities, and age groups, then a much larger set may be needed to cover the scope of $\mathcal{P}$.

Obviously, the method can be extended readily to other body regions via CT imagery and other application foci involving different biological characteristics. Extending to magnetic resonance (MR) images may be more challenging due to the fact that MR image intensities do not possess tissue-specific numeric meaning unlike CT. This issue of non-standardness of intensities can be addressed by using intensity standardization techniques (Nyúl and Udupa, 1999). In this manner, the grouping behavior can be studied for each MR imaging protocol (T1-weighted, T2-weighted, etc.) independently. With appropriate design of similarity/cost functions, it may be feasible to deal with multiple protocols simultaneously to study grouping, although we do not see any additional advantages of such an approach compared to grouping based on one protocol that is best for portraying anatomy, at least from the perspective of the main application of this paper.

It is also worth studying similarity/cost functions other than MAD for grouping and their influence on the actual partitions that result in comparison to MAD. Our experiments with a different grouping method (Ward's method) suggest that a proper choice of the algorithm is crucial for achieving groupings that are appropriate for the biological characteristics considered for the application. For the same grouping algorithm, but by using different similarity functions (MAD and MSD), we found the grouping behavior to be similar. MAD or similar intensity-based dissimilarity functions offer broad and general characteristics for grouping which encompass within them form, shape, and appearance characteristics of objects as well as their geographic layout (which we broadly referred to as *body habitus*). We recommend using MAD as the cost function because of its superior robustness compared to MSD. As we demonstrated quite amply, such functions are appropriate for our application of body-wide object recognition and delineation and we do not claim them to be a panacea for all or most applications.

There are several gaps in this investigation that need further closer scrutiny. We did not analyze grouping separately for female and male populations due to size restrictions in our data sets. Interestingly, however, our grouping method HG seems to be able to handle this situation automatically by creating a separate group for female subjects or "subjects with large breasts", as observed in our thoracic population NA. We examined closely how the female and male subjects were distributed in the groupings for NA for both thorax and H&N. For thorax, as illustrated in Fig. 10, Group 4 selected only females and Group 5 consisted of all females plus 2 males with excessive fat tissues. Other large groups for thorax were mixtures of males and females. For H&N, there is no significant influence of gender on groupings; all large groups included male and female subjects. Our results suggest that, for certain body regions, there is no need to perform grouping analysis separately for males and females (unless the application calls for such separation); our method will automatically handle differences in body habitus. In body regions such as abdomen and pelvis, where there are major gender-specific differences in organs and their layout, clearly populations should be defined and analyzed separately for males and females.

Another gap in the paper is that we focused only on image level grouping and it may be argued that, for model/atlas construction, object-level analysis and insight is more appropriate. While we are currently examining grouping based on segmented objects, this issue is quite complex. Our earlier studies indicated the non-linear nature of the variation of the properties of and the relationships among objects (Matsumoto et al., 2016). For example, some objects vary much less than others do in their size *and* their spatial relationship to other objects over a population. The size and form aspect of individual objects can be handled by analyzing grouping at the object level. However, the geographical layout of objects where object relationships matter cannot be studied by object-level grouping of individual objects separately. It may be argued that this latter aspect can be handled quite well by image-level grouping. Perhaps image- and object-level grouping processes may have to be combined in some manner to handle intra- and inter-object variations. This is clearly an open area for investigation.

We are also studying the utility of the grouping approach in multi-model/atlas-based strategies for object segmentation, which is perhaps the strongest arbiter of the effectiveness/usefulness of grouping. It may be argued that image-level grouping may be useful also in deep-learning strategies for image segmentation and analysis since this can allow networks as models to be trained more effectively and tailored to each group individually.

Declaration of Competing Interest

None.

CRediT authorship contribution statement

Ze Jin: Conceptualization, Investigation, Writing - original draft, Writing - review & editing. **Jayaram K. Udupa:** Conceptualization, Investigation, Writing - original draft, Writing - review & editing. **Drew A. Torigian:** Conceptualization, Investigation, Writing - original draft, Writing - review & editing.

Acknowledgment

This work is supported by a DHHS grant R42 CA199735-02 from the National Cancer Institute.

References

Ashburner, J., Friston, K.J., 2009. Computing average shaped tissue probability templates. Neuroimage 45, 333–341. https://doi.org/10.1016/j.neuroimage.2008.12.008.

Bai, P.R., Udupa, J.K., Tong, Y., Xie, S.P., Torigian, D.A., 2019. Body region localization in whole-body PET/CT scans using virtual landmarks. Med. Phys. 46 (3), 1286–1299. https://doi.org/10.1002/mp.13376.

Bai, W., Shi, W., Ledig, C., Rueckert, D., 2015. Multi-atlas segmentation with augmented features for cardiac MR images. Med. Image. Anal. 19 (1), 98–109. https://doi.org/10.1016/j.media.2014.09.005.

Bardinet, E., Cohen, L.D., Ayache, N., 1998. A parametric deformable model to fit unstructured 3D data. Comput. Vis. Image. Underst. 71, 39–54. https://doi.org/10.1006/cviu.1997.0595.

Christensen, G., Rabbitt, R., Miller, M., 1994. 3-D brain mapping using a deformable neuroanatomy. Phys. Med. Biol. 39 (3), 609–618.

Cireşan, D.C., Giusti, A., Gambardella, L.M., Schmidhuber, J., 2013. Mitosis detection in breast cancer histology images with deep neural networks. Med. Image. Comput. Comput. Assist. Interv. 16 (2), 411–418. https://doi.org/10.1007/978-3-642-40763-5_51.

Cootes, T.F., Taylor, C.J., Cooper, D.H., Graham, J., 1995. Active shape models-their training and application. Comput. Vis. Image. Und. 61 (1), 38–59. https://doi.org/10.1006/cviu.1995.1004.

Cootes, T.F., Edwards, G.J., Taylor, C.J., 2001. Active appearance models. IEEE Trans. Pattern Anal. Mach. Intell 23, 681–685.

Drozdza, lM., Chartrand, G., Vorontsov, E., Shakeri, M., Di Jorio, L., Tang, A., Romero, A., Bengio, Y., Pal, C., Kadoury, S., 2018. Learning normalized inputs for iterative estimation in medical image segmentation. Med. Image. Anal. 44, 1–13. https://doi.org/10.1016/j.media.2017.11.005.

Fernández, A., Gómez, S., 2008. Solving non-uniqueness in agglomerative hierarchical clustering using multidendrograms. J. Classification 25 (1), 43–65. https://doi.org/10.1007/s00357-008-9004-x.

Gee, J.C., Reivich, M., Bajcsy, R., 1993. Elastically deforming 3D atlas to match anatomical brain images. J. Comput. Assist. Tomogr. 17 (2), 225–236.

Grevera, G.J., Udupa, J.K., Odhner, D., Torigian, D.A., 2016. Optimal atlas construction through hierarchical image registration. In: Medical Imaging 2016: Image-Guided Procedures, Robotic Interventions, and Modeling, 9786. International Society for Optics and Photonics, p. 97862C.

Heimann, T., Meinzer, H.P., 2009. Statistical shape models for 3D medical image segmentation: a review. Med. Image. Anal. 13 (4), 543–563. https://doi.org/10.1016/j.media.2009.05.004.

Isgum, I., Prokop, M., Niemeijer, M., Viergever, M.A., Van Ginneken, B., 2012. Automatic coronary calcium scoring in low-dose chest computed tomography. IEEE Trans. Med. Image 31 (12), 2322–2334. https://doi.org/10.1109/TMI.2012.2216889.

Jin, Z., Udupa, J.K., Torigian, D.A., 2019. Obtaining the potential number of models/atlases needed for capturing anatomic variations in population images. In: Proceedings of SPIE, Medical Imaging 2019, 10949. 109493G-1– 109493G-8. https://doi.org/10.1117/12.2513073.

Liu, T., Udupa, J.K., Miao, Q., Torigian, D.A., 2019. Quantification of body-torso-wide tissue composition on low-dose CT images via automatic anatomy recognition. Med. Phys. 46 (3), 1272–1285. https://doi.org/10.1002/mp.13373.

Matsumoto, M.M., Udupa, J.K., Tong, Y., Saboury, B., Torigian, D.A., 2016. Quantitative normal thoracic anatomy at CT. Comput. Med. Imaging Graph. 51, 1–10. https://doi.org/10.1016/j.compmedimag.2016.03.005.

Moeskops, P., Viergever, M.A., Mendrik, A.M., de Vries, L.S., Benders, M.J., Išgum, I., 2016. Automatic segmentation of brain MR images with a convolutional neural network. IEEE Trans. Med. Imaging 35 (5), 1252–1261. https://doi.org/10.1109/TMI.2016.2548501.

Nguyen, D.C.T., Benameur, S., Mignotte, M., Lavoie, F., 2018. Superpixel and multi-atlas based fusion entropic model for the segmentation of X-ray images. Med. Image. Anal. 48, 58–74. https://doi.org/10.1016/j.media.2018.05.006.

Nyúl, L.G., Udupa, J.K., 1999. On standardizing the MR image intensity scale. Magnetic resonance in medicine. Magn. Reson. Med. 42 (6), 1072–1081. https://doi.org/10.1002/(SICI)1522-2594(199912)42:6%3C1072::AID-MRM11%3E3.0.CO;2-M.

Oda, H., Roth, H.R., Bhatia, K.K., Oda, M., Kitasaka, T., Iwano, S., Homma, H., Takabatake, H., Mori, M., Natori, H., Schnabel, J.A., Mori, K., 2018, February. Dense volumetric detection and segmentation of mediastinal lymph nodes in chest CT images. Medical Imaging 2018: Computer-Aided Diagnosis, 10575. International Society for Optics and Photonics.

Rittner, L., Udupa, J.K., Torigian, D.A., 2014. Multiple fuzzy object modeling improves sensitivity in automatic anatomy recognition. In: Medical Imaging 2014: Image Processing, 9034. International Society for Optics and Photonics, p. 90342U.

Sanroma, G., Wu, G., Gao, Y., Shen, D., 2014. Learning to rank atlases for multiple-atlas segmentation. IEEE Trans. Med. Imaging 33 (10), 1939–1953. https://doi.org/10.1109/TMI.2014.2327516.

Shen, T., Li, H., Huang, X., 2011. Active volume models for medical image segmentation. IEEE Trans. Med. Imaging. 30 (3), 774–791. https://doi.org/10.1109/TMI.2010.2094623.

Shi, C., Cheng, Y., Wang, J., Wang, Y., Mori, K., Tamura, S., 2017. Low-rank and sparse decomposition based shape model and probabilistic atlas for automatic pathological organ segmentation. Med. Image Anal. 38, 30–49. https://doi.org/10.1016/j.media.2017.02.008.

Staib, L.H., Duncan, J.S., 1992. Boundary finding with parametrically deformable models. IEEE Trans. Pattern Anal. Mach. Intell. 11, 1061–1075. http://doi.ieeecomputersociety.org/10.1109/34.166621.

Turaga, S.C., Murray, J.F., Jain, V., Roth, F., Helmstaedter, M., Briggman, K., Denk, W., Seung, H.S., 2010. Convolutional networks can learn to generate affinity graphs for image segmentation. Neural Comput. 22 (2), 511–538. https://doi.org/10.1162/neco.2009.10-08-881.

Udupa, J.K., Odhner, D., Zhao, L., Tong, Y., Matsumoto, M.M.S., Ciesielski, K.C., Falcao, A.X., Vaideeswaran, P., Ciesielski, V., Saboury, B., Mohamadanrasanani, Sin, S., Arens, R., Torigian, D.A., 2014. Body-wide hierarchical fuzzy modeling, recognition, and delineation of anatomy in medical images. Med. Image. Anal. 18, 752–771. https://doi.org/10.1016/j.media.2014.04.003.

Van de Velde, J., Wouters, J., Vercauteren, T., De Gersem, W., Achten, E., De Neve, W., Van Hoof, T., 2016. Optimal number of atlases and label fusion for automatic multi-atlas-based brachial plexus contouring in radiotherapy treatment planning. Radiat. Oncol. 11 (1), 1. https://doi.org/10.1186/s13014-015-0579-1.

Wang, H., Yushkevich, P.A., 2013. Multi-atlas segmentation without registration: a supervoxel-based approach. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, Berlin, Heidelberg, pp. 535–542.

Ward Jr, J.H., 1963. Hierarchical grouping to optimize an objective function. J. Am. Stat. Assoc. 58, 236–244.

Wu, G., Kim, M., Sanroma, G., Wang, Q., Munsell, B.C., Shen, D., Initiative, Alzheimer's Disease Neuroimaging, 2015. Hierarchical multi-atlas label fusion with multi-scale feature representation and label-specific patch partition. Neuroimage 106, 34–46. https://doi.org/10.1016/j.neuroimage.2014.11.025.

Wolz, R., Chu, C., Misawa, K., Mori, K., Rueckert, D., 2012. Multi-organ abdominal CT segmentation using hierarchically weighted subject-specific atlases. Med. Image. Comput. Comput. Assist. Interv. 15 (1), 10–17. https://doi.org/10.1007/978-3-642-33415-3_2.

Wu, X., Udupa, J.K., Tong, Y., Odhner, D., Pednekar, G.V., Simone II, C.B., McLaughlin, D., Apinorasethkul, C., Lukens, J., Mihailidis, D., Shammo, G., James, P., Camaratta, J., Torigian, D.A., 2019. AAR-RT - A system for auto-contouring organs at risk on CT images for radiation therapy planning: principles, design, and large-scale evaluation on head-and-neck and thoracic cancer cases. Med. Image. Anal. 54, 45–62. https://doi.org/10.1016/j.media.2019.01.008.

Xu, G., Udupa, J.K., Tong, Y., Cao, H., Odhner, D., Torigian, D.A., Wu, X., 2018. Thoracic lymph node station recognition on CT images based on automatic anatomy recognition with an optimal parent strategy. In: Proceedings of SPIE, Medical Imaging 2018, 10574. 10574F-1 – 10574F7. https://doi.org/10.1117/12.2293258.

Xu, Z., Burke, R.P., Lee, C.P., Baucom, R.B., Poulose, B.K., Abramson, R.G., Landman, B.A., 2015. Efficient multi-atlas abdominal segmentation on clinically acquired CT with SIMPLE context learning. Med. Image. Anal. 24 (1), 18–27. https://doi.org/10.1016/j.media.2015.05.009.

Yang, J., Beadle, B.M., Garden, A.S., Gunn, B., Rosenthal, D., Ang, K., Frank, S., Williamson, R., Balter, P., Court, L., Dong, L., 2014. Auto-segmentation of low-risk clinical target volume for head and neck radiation therapy. Pract. Radiat. Oncol. 4 (1), e31–e37. https://doi.org/10.1016/j.prro.2013.03.003.

Zhang, W., Li, R., Deng, H., Wang, L., Lin, W., Ji, S., Shen, D., 2015. Deep convolutional neural networks for multi-modality isointense infant brain image segmentation. Neuroimage 108, 214–224. https://doi.org/10.1016/j.neuroimage.2014.12.061.