Automatic anatomy recognition using neural network learning of object relationships via virtual landmarks

Fengxia Yan, Jayaram K. Udupa, Yubing Tong, Guoping Xu, Dewey Odhner, et al.

Fengxia Yan, Jayaram K. Udupa, Yubing Tong, Guoping Xu, Dewey Odhner, Drew A. Torigian, "Automatic anatomy recognition using neural network learning of object relationships via virtual landmarks," Proc. SPIE 10574, Medical Imaging 2018: Image Processing, 105742O (2 March 2018); doi: 10.1117/12.2293700

Event: SPIE Medical Imaging, 2018, Houston, Texas, United States

Automatic Anatomy Recognition using Neural Network Learning of Object Relationships via Virtual Landmarks

Fengxia Yan[a,b], Jayaram K. Udupa[b], Yubing Tong[b], Guoping Xu[b], Dewey Odhner[b], Drew A. Torigian[b]

a. College of Science, National University of Defense Technology, Changsha 410073, P. R. China;

b. Medical Image Processing Group, 602 Goddard building, 3710 Hamilton Walk, Department of Radiology, University of Pennsylvania, Philadelphia, PA 19104, United States

ABSTRACT

The recently developed body-wide Automatic Anatomy Recognition (AAR) methodology depends on fuzzy modeling of individual objects, hierarchically arranging objects, constructing an anatomy ensemble of these models, and a dichotomous object recognition–delineation process. The parent-to-offspring spatial relationship in the object hierarchy is crucial in the AAR method. We have found this relationship to be quite complex, and as such any improvement in capturing this relationship information in the anatomy model will improve the process of recognition itself. Currently, the method encodes this relationship based on the layout of the geometric centers of the objects. Motivated by the concept of virtual landmarks (VLs), this paper presents a new one-shot AAR recognition method that utilizes the VLs to learn object relationships by training a neural network to predict the pose and the VLs of an offspring object given the VLs of the parent object in the hierarchy. We set up two neural networks for each parent-offspring object pair in a body region, one for predicting the VLs and another for predicting the pose parameters. The VL-based learning/prediction method is evaluated on two object hierarchies involving 14 objects. We utilize 54 computed tomography (CT) image data sets of head and neck cancer patients and the associated object contours drawn by dosimetrists for routine radiation therapy treatment planning. The VL neural network method is found to yield more accurate object localization than the currently used simple AAR method.

Keywords: Virtual landmarks, neural network learning, object recognition, head and neck, computed tomography (CT).

1. INTRODUCTION

The practice of Radiology has largely been qualitative ever since the discipline was established with the discovery of x-rays in 1895. However, it is moving towards Quantitative Radiology (QR) rapidly. To make QR a reality in radiological practice, the problem of image segmentation must be solved so as to offer adequate levels of automation and accuracy for any body region. We believe that computerized Automatic Anatomy Recognition (AAR) during radiological image interpretation becomes essential for this purpose. The recently developed AAR methodology based on fuzzy modeling [1] demonstrated its ability for automatically recognizing and delineating body-wide anatomy in given patient images on over 100 organs and conceptual anatomic regions such as lymph node zones. There are three main steps in AAR: model building, object recognition, and object delineation. In the model building part of AAR, there is a very important element called parent-to-offspring relationship ρ which explicitly encodes the parent-to-offspring spatial relationship information into a hierarchy H in which objects are organized. This information is subsequently exploited in the following object recognition and delineation steps. We think that the more accurate the relationship is modeled and described, the better the recognition performance will be. In the current AAR method, the simple geometric relationships between object centroids and their

Medical Imaging 2018: Image Processing, edited by Elsa D. Angelini, Bennett A. Landman, Proc. of SPIE
Vol. 10574, 105742O · © 2018 SPIE · CCC code: 1605-7422/18/$18 · doi: 10.1117/12.2293700

statistics were used to express this relationship [1]. However, these relationships between parent and offspring objects can be quite complex. As such, more sophisticated ways of capturing these relationships may be beneficial to the AAR process. Motivated by another novel concept introduced recently, called virtual landmarks (VLs) [2], this paper presents a new one-shot AAR-recognition method based on expressing and neural network learning of object relationships via VLs. The neural network is used to learn not only the relationship between two sets of VLs from parent and offspring objects, but also the relationship between the VLs of parent object and transformation parameters of the child object. The method is evaluated on 54 computed tomography (CT) studies used for planning radiation therapy of head and neck cancer patients.

2. METHODS

In this paper, we will focus on one body region for initial demonstration, namely head & neck (H&N). Following published guidelines [1, 3] for H&N anatomic object definitions, we formulated detailed and precise definitions for specifying each object and for delineating its boundaries on axial CT slices. The objects considered in this study are: Skin outer Boundary (SB), Left and Right Parotid Glands and their union (LPG, RPG, PG), Left and Right Submandibular Glands and their union (LSG, RSG, SG), Esophagus (ES), Larynx (LX), Spinal Canal (SC), Mandible (MD), and Orohypopharynx constrictor muscle (OHP). We further subdivided object SB into an inferior portion below the neck (SBi) and a superior portion (SBs) within the neck.

The proposed method investigates how the VLs of objects can be used to improve object recognition in AAR by learning via neural networks the relationship between parent and offspring objects. The method works overall as follows. For each object (from 14 objects in total) for all its samples, the corresponding VLs are computed first. For each pair of objects, we design two neural networks – one for learning the relationship between VLs of the parent object and the transformation parameters to express the relationship between the parent and the offspring, and the other for learning the relationship between VLs of the parent and child objects. Subsequently, the trained networks are used to predict the transformation parameters and VLs of child object for any test image. The idea is that once the parent object is known, we can use its VLs to predict through the neural network the pose of the child object, as well as its VLs. In our situation, the recursive process starts at the root object which is always SB. Finally, the one-shot recognition is performed to the object based on the predicted transformation information. These steps are further described below. Note that the results of this one-shot recognition can be further refined by using a variety of other strategies as described in [3]. The one-shot method can be thought of as an initializer of later refined strategies. If this process can be made more accurate, then the entire AAR process will benefit from it.

(a) **Image data**: We have retrospectively collected CT image and object contour (ground truth segmentation) data sets from the Department of Radiation Oncology, University of Pennsylvania, for 216 H&N cancer patients following approval from the Institutional Review Board at the Hospital of the University of Pennsylvania along with a Health Insurance Portability and Accountability Act (HIPAA) waiver. The voxel size in these data sets ranges from 0.97×0.97×1.5 mm^3 to 1.46×1.46×3 mm^3. These were routine clinical scans for which the contour data were previously created for clinical purposes by the dosimetrists in the process of routine radiation therapy (RT) planning. The data sets constitute 54 cases in each of four groups: 40-59-year-old males and females, and 60-79-year-old males and females. In the results reported here, we have used the 54 cases in group 60-79-year-old males, and we are in the process of performing similar analysis on all data sets. Among all cases, 36 of them have all 14 objects.

(b) **Computing virtual landmarks**: A previous publication described the concept and techniques underlying the idea of VLs [1, 3]. Briefly, VLs associated with an anatomic object are reference points which are tethered to the object. The points may reside inside, on the boundary of (although rarely), or outside the object, and are tethered to the object in the sense of being homologous. They can be defined on the binary image representing the object or using both object shape and object gray value appearance. The approach of obtaining VLs is straightforward, simple, and recursive in nature, proceeding from global features initially to local features in later levels to detect landmarks. The landmarks are obtained through a process of recursive subdivision of the object guided by principal component analysis (PCA). At the highest level, the geometric center of the object is the only landmark produced. The eigenvectors associated with the object define a principal axis system which divides the object into 8 octants. The part of the object in each of these octants is again subjected to PCA which yields a geometric center and a principal axis system. Thus, at the second level, 8 landmarks are generated. In the third level, continuing this process of subdivision, 64 landmarks are generated, and so on. It is clear that the method allows selection of any desired virtual landmarks and any number of them since each landmark has a unique identifier associated with it in the process of subdivision. In this work, we used different numbers of VLs as described below. In the 3D case, at level x we will have 8^{x-1} octants, and there will be in total $\sum_{n=1}^{x} 8^{n-1}$ points for x levels. For x =2, 3, and 4 levels, the number of virtual landmarks generated is 9, 73, and 585, respectively.

(c) **Learning parent-to-child relationship**. Our goal is to learn the relationship between the VLs of parent object and parameters of the geometric transformation needed to predict the pose of the offspring object in relation to its parent for each parent-child pair (O_1, O_2) in the hierarchy via one neural network. When we use a hierarchy that has more than 2 levels, some offspring objects O_2 will recursively become the parent object of the next level offspring objects. As such, for each such an object O_2, we will need to predict not only the transformation relative to its parent but also its own VLs since these VLs will be needed to predict the transformation parameters of O_2's children. So, for each such pair, we design a second neural network to learn the relationship between the VLs of the parent and child. The first network is configured as a regressor by feeding VLs and scaling factor of the parent object as input and the transformation parameters (confined to translation and isotropic scaling in this paper, but easily generalized to more complex transformations) as output data. The second network is configured by feeding VLs of the parent as input and VLs of the child as target output data. Thus, the number of networks set up will be *A*+*B* where *A* denotes the number of parent-child pairs in the chosen hierarchy, and *B* denotes the number of parent-child pairs where the child itself is a parent in the hierarchy.

Here we adopt a simple architecture of a multiple-layer neural network with one hidden layer. The number of neurons in the input layer is the same as the dimension of VLs of the parent object, and the number of neurons in the output layer equals the size of the target transformation vector or the dimension of VLs of the child object. In addition, the numbers of neurons in the hidden layer are determined by choosing optimal numbers to yield minimum error. The details of the neural network configuration and training will be presented in the next section.

(d) **One-shot Recognition**. The idea of one-shot recognition described in [1, 3] is to determine the pose of the child object directly from knowledge of the parent object (in our case, fuzzy model) from prior knowledge. This gives an initial pose which is further refined by using different techniques that make use of the intensity information in the particular given image [1, 3]. If the one-shot result is accurate, the subsequent refinement techniques can achieve better object localization and delineation. Therefore, the role sought here for pose prediction via VLs is to learn the complex parent-to-child relationship, codify that as prior knowledge, and harness this knowledge to locate the child object with high pose accuracy. Assume that we know the VLs of the parent object. This is true for the root object which in our approach is usually SB (which can be localized quite accurately and delineated [1, 3] and so its VLs can be computed), and by recursion, it is true for other offspring objects. We use parent VLs and scaling factor to predict the transformation parameters and VLs of each offspring object and proceed recursively down the object hierarchy employed in our AAR process. Once the transformation needed for the child is estimated, we perform this new one-shot recognition based on VLs to locate the child object. We then compare the simple one-shot method currently used in [1, 3] with this new approach.

3. EXPERIMENTS AND RESULTS

To compare the proposed VL method with the current one-shot method, we choose two different hierarchies to perform one-shot recognition (Figure 1). Hierarchy1 is the simplest one-level tree structure, and Hierarchy2 is the optimal tree structure derived from our current AAR method. For each object (total 14 objects) for all its samples, a set of 9 virtual landmarks are derived from two levels (x=2). For Hierarchy1, we need only the first type of neural network to learn the relationship between VLs of object SB and the transformation parameters of each of the 13 child objects since no offspring object has its own offspring. Here, the VL and scaling factor data of object SB with dimension as $(N \times D+1) \times S$ are set as input data to the neural network, where N represents the number of VLs (in our case $N = 9$), D represents the dimensionality of the spatial coordinates of the VLs (in our case $D = 3$), and S represents number of subjects. Thus, if all data were to be utilized, the dimensionality of input data will be $28 \times S$. The dimensionality of the target output data is $L \times S$, where L represents the dimensionality of the transformation parameters (in our case $L = 4$). For Hierarchy2, we need both types of neural networks; one to learn the relationship between VLs of parent object and transformation parameters for each parent-child pair in different levels: 1) (SB, SC), (SB, SBs), (SB, LX), 2) (SBs, SBi), (SBs, RPG), (SBs, LPG), (SBs, MD), (SBs, SBi), and 3) (SBi, ES), (MD, OHP), (MD, PG), (MD, RSG), (MD, SG), (MD, LSG). At the stage of recognition, SBs, SBi, and MD will be the parent objects for objects in the next level. Therefore, we design the second type of neural network in which the VL data of parent object are set as input, and the VLs of child object are set as output for each of the parent-child pairs (SB, SBs), (Sbs, SBi), and (Sbs, MD). Using these neural networks, we can predict the VLs of object SBs, SBi, and MD. Obviously, the dimensionality of the target output data for these networks should be $28 \times S$. Thus, the total number of networks trained for the two hierarchies is: $A+B=13$ ($A=13$, $B=0$) for Hierarchy1, and $A+B=16$ ($A=13$, $B=3$) for Hierarchy2.

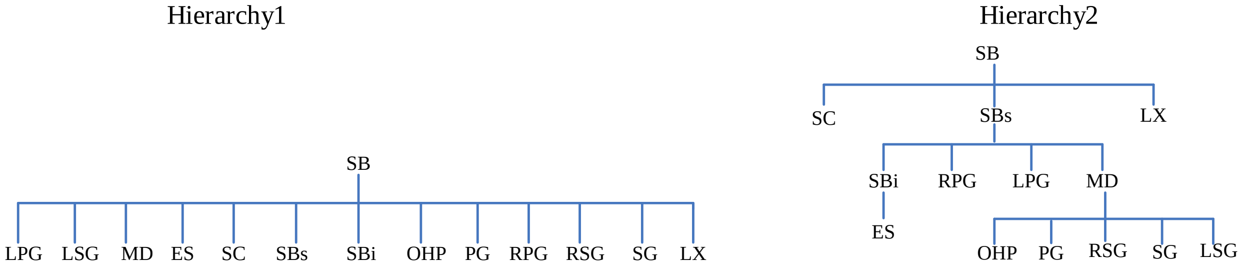

Figure 1. Object hierarchies used in this work.

After configuring the input data and target data of the neural network regressor, we still need to set data division, network architecture, and the training algorithm. Here we use the neural network toolbox of MATLAB, which is powerful and convenient to implement these operations [4]. As in our previous work [5], for both neural networks, we choose the Bayesian Regularization algorithm to implement the training process because it can prevent overfitting and provide better performance than the Levenberg-Marquardt algorithm. A multilayer neural network with a single hidden layer is chosen as the architecture, and different numbers of neurons in the hidden layer are selected in terms of search for minimum testing error by doing rotation training. This means that we choose 5 subjects for test, 6 subjects for validation, and the remaining subjects for training each time, and then do a rotation until each subject has been used for test. We also use this rotation training to obtain the optimal number of neurons in the hidden layer. After obtaining the optimal neural network for each parent-child pair in Hierarchy1 and Hierarchy2, for each test subject, we predict the corresponding transformation parameters for each offspring object and then perform the one-shot recognition. Here we used the known ground truth location of the root object SB to initialize the recognition process for both experiments.

The recognition accuracy is expressed in terms of position error and size error. The position error is defined as the distance between the geometric centers of the known true objects and the predicted center of the recognized objects. The size error is expressed as a ratio of the estimated size of the object at recognition and true size. Values of 0 and 1 for the two measures, respectively, indicate perfect recognition.

We utilized 20 subjects to build the model and 12 subjects to perform the recognition test. Fig.2 displays the ground truth of a sample object mandible (MD) and the one-shot recognition result from the proposed method. Results for recognition through Hierarchy1 are summarized in Tables 1 and 2 for the proposed and the current AAR one-shot method, respectively. Tables 3 and 4 similarly present the results for recognition through Hierarchy2. We can see that for Hierarchy1, for each object, the location error of the proposed method is much better than that of the current one-shot method [1]. For Hierarchy2, the location error of the proposed method is also better than that of the current one-shot method [1] except for objects SBi and ES. The size error of the proposed method is always close to 1 for all objects for Hierarchy1 and Hierarchy2.

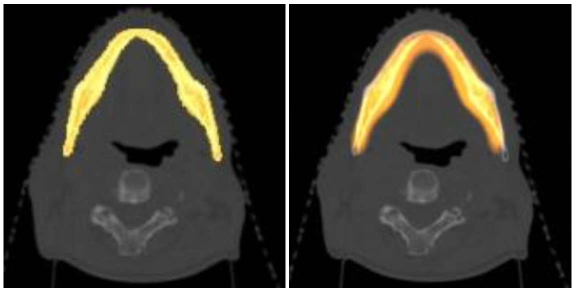

Fig 2. Sample displays of ground truth and recognition result for mandible.

Table 1. One-shot recognition results (mean, standard deviation) for H&N objects using VLs based on Hierarchy1

	LPG	LSG	MD	ES	SC	SBs	SBi	OHP	PG	RPG	RSG	SG	LX
Location error (mm)	10.99	15.66	12.42	10.13	8.32	10.23	5.85	13.91	12.51	13.58	14.94	13.1	14.28
	9.85	11.6	8.29	5.09	3.39	9.86	5.18	8.57	8.02	10.0	9.62	8.49	7.30
Size error	1.01	1.02	1.00	1.00	0.99	1.00	1.00	0.98	1.02	1.03	1.01	0.99	1.00
	0.14	0.09	0.04	0.12	0.05	0.05	0.03	0.06	0.05	0.13	0.1	0.05	0.09

Table 2. One-shot recognition results (mean, standard deviation) for H&N objects using strategy of [1] based on Hierarchy1

	LPG	LSG	MD	ES	SC	SBs	SBi	OHP	PG	RPG	RSG	SG	LX
Location error (mm)	15.97	20.2	17.23	11.06	12.29	16.66	10.69	17.07	17.56	16.82	16.43	16.64	16.08
	11.03	8.32	13.94	5.48	3.94	12.87	7.22	10.76	13.24	9.98	9.52	12.07	8.48
Size error	1.06	1.05	1.03	1.01	0.96	1.02	1.02	0.99	1.05	1.09	1.04	1.02	1
	0.13	0.1	0.07	0.15	0.08	0.05	0.06	0.08	0.1	0.11	0.09	0.08	0.11

Table 3. One-shot recognition results (mean, standard deviation) for H&N objects using VLs based on Hierarchy2

	LPG	LSG	MD	ES	SC	SBs	SBi	OHP	PG	RPG	RSG	SG	LX
Location error (mm)	12.9	15.31	14.92	12.44	8.32	10.23	12.27	15.16	13.58	14.09	15.79	15.54	14.28
	10.05	12.15	13.5	10.2	3.29	9.86	11.12	12.13	10.58	10.19	13.42	10.03	7.30
Size error	1.02	1.04	1.10	1.01	0.99	1.00	1.00	1.00	1.03	1.05	1.12	0.11	1.00
	0.14	0.09	0.04	0.12	0.05	0.05	0.03	0.06	0.05	0.13	0.1	0.05	0.09

Table 4. One-shot recognition results (mean, standard deviation) for H&N objects using strategy of [1] based on Hierarchy2

	LPG	LSG	MD	ES	SC	SBs	SBi	OHP	PG	RPG	RSG	SG	LX
Location error (mm)	16.19	20.43	17.38	11.36	12.29	16.66	10.55	17.18	17.69	17	17.01	17.12	16.08
	11.16	8.72	14.42	5.75	3.94	12.87	6.85	10.94	13.48	10.18	9.51	11.88	8.48
Size error	1.06	1.05	1.03	1.01	0.96	1.02	1.02	0.99	1.05	1.09	1.04	1.02	1
	0.13	0.1	0.07	0.15	0.08	0.05	0.06	0.08	0.1	0.11	0.09	0.08	0.11

4. CONCLUSION

This paper introduces a novel VLs-based one-shot recognition approach in which VLs of objects are utilized to capture the parent-offspring relationships. The method is based on designing two types of neural networks, one to learn the relationships between the VLs of parent and child objects, and another to learn the relationship between the VLs of parent object and the transformation parameters needed to express the pose relationship between the parent and child, respectively. These networks were then used to predict the transformation parameters of the child object. Finally, the obtained transformation parameters were employed to perform the one-shot recognition.

The initial results seem to indicate that the new pose prediction method is better than the current simple method. In this work, we have used a very simple 4-parameter homothetic transformation. More sophisticated transformations including 6-, 7-, and 9-parameter cases may yield better approximations of the object relationships. We may also use optimal hierarchical registration [6] and thereby find parent-to-child relationships in an optimal manner over an ensemble of data sets for training the network. Currently the VLs derived from binary objects are employed. VLs derived from gray-valued objects may further refine VL definition and recognition accuracy.

Acknowledgement

This work is supported by a grant IIP-1549509 from the US National Science Foundation and a grant from the National Institutes of Health R41 CA199735-01A1. The work of Dr. Fengxia Yan was partially supported by the National Natural Science Foundation of China (under Grant No. 61471369) and her visit and tenure in the Medical Image Processing Group, Department of Radiology, University of Pennsylvania, Philadelphia, for the duration of one year was supported by China Scholarship Council.

REFERENCE

[1] Udupa J K, Odhner D, Liming Z, *et al*. "Body-wide hierarchical fuzzy modeling, recognition, and delineation of anatomy in medical images". Medical Image Analysis, 18(5),752-771 (2014).

[2] Yubing Tong, Jayaram K Udupa, Dewey Odhner, Peirui Bai, Drew A. Torigian. "Virtual landmarks," Proceedings of SPIE Medical Imaging Conference, 10135: 1013521-1 - 1013521-6, doi: 10.1117/12.2254855 (2017).

[3] Wang H, Udupa J K, Odhner D, *et al*. "Automatic anatomy recognition in whole-body PET/CT images". Medical Physics, 43(1), 613-629 (2016).

[4] Peirui Bai, Tong, Jayaram K Udupa, Yubing Tong, ShiPeng Xie, Drew A. Torigian. "Automatic thoracic body region localization", Proceedings of SPIE Medical Imaging Conference, 10134: 101343X-1 - 101343X-6, doi: 10.1117/12.2254862 (2017).

[5] Martin T. Hagan, Howard B. Demuth, Mark Hudson Beale, *et al*. [Neural Network Design], 2nd Edition, eBook. (2014).

[6] Grevera. G.J., Udupa, J.K., Odhner, D., Torigian, D.A.: "Optimal atlas construction through hierarchical image registration, Proceedings of SPIE Medical Imaging Conference, 9786: 97862C-1 – 97862C-6, 2016.