Anthony P. Reeves ; Shuang Liu ; Yiting Xie; Image segmentation evaluation for very-large datasets. Proc. SPIE 9785, Medical Imaging 2016: Computer-Aided Diagnosis, 97853J (March 24, 2016);

doi:10.1117/12.2217331.

Image segmentation evaluation for very-large datasets

Anthony P. Reeves, Shuang Liu and Yiting Xie

School of Electrical and Computer Engineering, Cornell University, Ithaca, NY 14853

ABSTRACT

With the advent of modern machine learning methods and fully automated image analysis there is a need for very large image datasets having documented segmentations for both computer algorithm training and evaluation. Current approaches of visual inspection and manual markings do not scale well to big data. We present a new approach that depends on fully automated algorithm outcomes for segmentation documentation, requires no manual marking, and provides quantitative evaluation for computer algorithms. The documentation of new image segmentations and new algorithm outcomes are achieved by visual inspection. The burden of visual inspection on large datasets is minimized by (a) customized visualizations for rapid review and (b) reducing the number of cases to be reviewed through analysis of quantitative segmentation evaluation. This method has been applied to a dataset of 7,440 whole-lung CT images for 6 different segmentation algorithms designed to fully automatically facilitate the measurement of a number of very important quantitative image biomarkers. The results indicate that we could achieve 93% to 99% successful segmentation for these algorithms on this relatively large image database. The presented evaluation method may be scaled to much larger image databases.

Keywords: large-scale evaluation, large datasets, image segmentation

1. INTRODUCTION

Fully automated analysis of medical images will provide a set of quantitative measurements that will be able to guide and improve physician decision-making. Medical images, especially those that are obtained periodically in the context of screening, provide a rich source of information for disease detection and health monitoring. Fully automated algorithms must be validated on a very large number of test cases before they can be approved for general clinical use. Evaluation cases must well represent the spectrum of clinical presentations. These requirements are in contrast to most image biomarker studies reported in the research literature. Those studies typically involve very small selected datasets, and further, they employ semi-automated computer methods that require physician interaction.

Key to the successful evaluation of most quantitative image biomarkers is the correct segmentation of the region of interest. However, the precise truth of a segmentation is typically not known in medical imaging applications[1]. For many current image biomarkers the measurement value is simple to obtain once the correct segmentation has been established.

1.1 Fully Automated 3D CT Image Segmentation Evaluation

There have been a number of fully automated image segmentation systems reported in the literature. In this paper we focus specially on CT images of the human body. Segmentation performance is usually reported by one of two methods: subjective post hoc visual evaluation (VE) or by quantitative evaluation (QE) comparison to a small number of pre-established expert manual markings. For QE, two popular variations that reduce the effort required are automated QE (aQE) in which a semi-automated method with manual corrections is used for image markings and sampled QE (sQE) in which only a subset of 2D slices in a 3D image are marked and evaluated.

Validation by VE is applicable to large datasets, but must be repeated from scratch for each algorithm development; further, repeated use of VE is subject to human bias and variation and, for algorithm development, is very time consuming and may lack repeatability. QE is often considered to be a higher quality evaluation since the correct outcome is established in advance; further this "correct" outcome once established may be used repeatedly for quantitative evaluation for algorithm development. However, QE suffers from the following serious shortcomings:

1. Manual marking is very time consuming and for large image regions the marking burden may be impractical. For example, marking the boundary region of the lungs in a typical chest CT scan may require specifying the

Medical Imaging 2016: Computer-Aided Diagnosis, edited by Georgia D. Tourassi, Samuel G. Armato III,
Proc. of SPIE Vol. 9785, 97853J · © 2016 SPIE · CCC code: 1605-7422/16/$18 · doi: 10.1117/12.2217331

location of over 400,000 boundary pixels. Even using aQE the time required to mark the outline of the breast in a CT scan was on average 18.6 minutes (range: 8.9-45.2)[2] using a commercial software tool.
2. Marker bias and variation may be very large and is permanently encoded as segmentation "truth", for example, experienced in the LIDC study[3].
3. Since the method is based on human behavior it is not repeatable even on exactly the same image dataset.

The aQE method has the further disadvantage that the algorithm assistance provides its own bias to the marking process. The sQE method is really only assessing area regions in 2D image slices and does not provide a volumetric assessment.

From a review of 37 studies in the literature (not including our own work) on fully automated segmentation from chest and abdomen CT scans, all except for two studies reported on less than 102 cases; these two studies used VE (302[4] and 1000[5] cases). The QE method was used in 26 studies having a median of 29 cases (min 7, max 101[6]). At least 16 of these studies employed an aQE or sQE variant of QE.

Studies that involve more than a thousand cases achieve this goal by validating with respect to some biomarker measurement outcome (such as a coronary artery calcium score) rather than the related segmentation issue (segmentation of coronary artery regions containing detectable calcium)[7]. While the outcome of such algorithms may be clinically useful; they are not validated for their stated design.

Very-large image datasets for the validation of fully automated algorithms have additional requirements. First they need to be extensible such that they can efficiently accommodate new image studies as they become available (for example to test against new imaging technology changes). Second they need to accommodate cases for which the automated algorithms typically fail; a desirable quality property of a fully automated algorithm is that it is able to detect a segmentation failure.

To address requirements for very-large documented image databases that are not met by traditional VE and QE methods, we have developed a Visual Evaluation Quantitative Revision (VEQR) method that scales to big data and also allows revisions and additions to a very-large documented image database. The key components of this method are: (a) customized visualization for rapid VE, (b) graded assessments that allow for some cases not to have acceptable segmentation documentation, (c) no modification of algorithm segmentation by manual marking, (d) revision to documentation by VE comparison of different algorithm outcomes and (e) automated quantitative evaluation of new algorithms.

1.2 Segmentation Error Types

Segmentation errors may be considered to fall into two general classes: minor errors (Me) and catastrophic errors (Ce). Minor errors occur due to differences in details between algorithms or between algorithms and the variation of manual image annotations: for this error type the difference between methods is typically small (dice coefficient > 0.9). Catastrophic errors occur when an algorithm incorrectly identifies or includes a significantly different region with the target region (for example, includes a nearby vessel as part of a lesion); the size of these errors may be very large. In most studies on segmentation algorithms, the dataset is usually small (< 100 cases) and of carefully selected images such that the majority of the errors are Me. However, when larger datasets with a wider range of imaging parameters and presentations are considered the likelihood of Ce errors is significantly increased. In a semi-automated environment where the primary target is image region characterization, Ce errors are rarely an issue since the operator manually corrects them. However, in fully automated systems the objective frequently has a focus on abnormality detection rather than characterization and Ce errors are a major consideration since they may correspond to an unacceptable large number of false positive abnormality detections. The evaluation criterion we use for image segmentations in this case primarily relates to the Ce error type.

In the fully automated context of this work we visually categorize the segmentation into three categories of "good", "acceptable" and "unacceptable"; the unacceptable category is caused by Ce error. When we have competing algorithms that provide good or acceptable outcomes we visually compare outcomes to select the best segmentation; typical this evaluation is with regards to Me error.

1.3 Segmentation Application

The evaluation method was initially designed for and has been tested on our research application of determining a range of quantitative image biomarkers for major diseases in the context of low-dose CT images resulting from lung cancer screening (LCS) or lung health monitoring. With the recent approval of LCS reimbursement for high-risk population[8], several million people will undergo LCS every year. The primary task for image analysis is to detect the very small pulmonary nodules that may indicate early stage lung cancer. However, subjects at risk for lung cancer are typically at risk for other major diseases of the chest including COPD and heart disease. The annual screening for lung cancer provides an opportunity for periodic monitoring patient health for many other diseases. Key to evaluating many quantitative image biomarkers is good image segmentation of the organ or region of interest. Once the correct regions identified then evaluating the biomarker is relatively simple; e.g., computing the Agatson score for coronary calcium once the heart region is identified. In this paper we use illustrate the image evaluation method with six fully automated algorithms that segment major organs and bones in the chest: the major airways, the lungs, the ribs, the vertebra, the skin surface plus fat regions, and the heart and major arteries. The low-dose screening scans have more image noise than typical clinical scans, which makes the segmentation task more challenging.

2. METHODS

2.1 Validation by Visual Evaluation and Quantitative Revision

The VEQR system addresses the need to evaluate fully automated image segmentation algorithms on very large image databases. Key components for the VEQR method are the VEQR database and custom 3D visualizations for rapid Ce segmentation quality review.

2.1.1 The VEQR Database

The validation database D comprises of image set I, label image set L, label assessment set A, and reference algorithm set R, i.e., D = {I, L, A, R}, where each of the four sets is defined as follows.

1. Image set I

$$I = \{ i \mid \text{ i is an image to be segmented and analyzed}\}$$

2. Label image set L

$$L = \{ l(i) \mid \forall i \in I\}$$

where l(i) is a label image of the same size as image i, and the value of each voxel in l(i) represents the label value assigned by segmentation algorithms to the corresponding voxel in image i. For instance, a label value of LungR indicates the respective voxels in image i belong to the right lung region, and a label value of LungL indicates the respective voxels belong to the left lung region, where both of the labels are assigned by the lung segmentation algorithm; a label value of Heart indicates the respective voxels in image i belong to the heart region, which is assigned by the Cardiac region segmentation algorithm.

3. Label assessment set A

$$A = \{a(i, s) \mid \forall i \in I , \forall s \in S\}$$

Where a(i, s) is the quality grade determined by visual assessment for a specific segmented region s in image i; S is the set of regions that can be segmented, for instance, S= {Lung, Airway, Rib, Vertebra, Skin, Cardiac region}. The quality grades usually take categorical values, for instance, a(i, s) ∈ {Good, Acceptable, Unacceptable}.

Note that the set of segmented regions, S, is not equivalent to the set of segmentation labels in the label image. In general, several sub-regions with distinct label values in the label image correspond to a segmented region. For example, the segmented region Rib is composed of voxels assigned with one of the 24 labels (ribL1, ribL2, …, ribL12, ribR1, ribR2, …, ribR12) in the label image; the segmented Cardiac region consists of voxels with label values of Heart, Aorta or Pulmonary trunk. The grade a(i, s) is assigned according to the overall quality of the segmented region s, and is penalized if there is mislabeling of sub-regions or confusion between sub-regions.

4. Reference algorithm set R

$$R = \{r(s) \mid \forall\, s \in S\}$$

Where r(s) is an algorithm that segments target region s from the image; S is the set of segmented regions as described above. For instance, r(Lung) is the algorithm that segments lung region from the image and assigns labels of LungR and LungL to the corresponding voxels in the label image.

The validation database supports three main functions: algorithm evaluation, new image addition and database revision.

1. **Algorithm evaluation** is a fully automated operation that provides quantitative comparison between the reference segmentation of the target region s and the outcome of a new segmentation algorithm n(s). The aggregate performance score for the new algorithm n(s) is determined on the evaluation on a subset I_{EV} of the image set I, where I_{EV} = {i | ∀ i ∈ I , and a(i, s) = Good or Acceptable}, i.e., only images with good or acceptable quality grades are used to evaluate a new algorithm. For each image i ∈ I_{EV}, if we let l_S (i) denote the segmented region s, which is recorded in label image l(i) in the database, and let l_N(i) denote the segmented region by the new algorithm n(s), the dice coefficient (DC) can be computed as follows to serve as the comparison score of the two segmented regions:

$$DC(l_S(i),\ l_N(i)) = \frac{2 \mid I_s\,(i) \cap\ I_N\,(i) \mid}{|I_s\,(i)| + |\,I_N\,(i)\,|}$$

The set of DC values associated with all images in I_{EV} in the database is then used to provide an aggregate performance score for the new algorithm n(s).

2. **Database revision** is an update to the database documentation based on the outcomes of a new algorithm n(s) that may for some cases provide superior segmentation outcomes to the current reference segmentation. For any image i ∈ I and a target segmentation region s, the database revision is accomplished by first computing the DC(l_S(i), l_N(i)) for the new algorithm segmentation l_N(i) with respect to the reference segmentation l_S(i) recorded in the database. Then the update is made as follows:
    a. If i ∈ I_{EV}, and the DC is less than a preset level T_{DClow}, then the new segmentation for that image is considered to be inferior to the reference segmentation, thereby no update is made.
    b. If i ∈ I_{EV}, and the DC is greater than a preset level T_{DChigh}, then the new segmentation is not considered to be a significant improvement over the reference segmentation, thereby no update is made.
    c. For the remainder of the cases, visual inspection is used to compare the reference segmentation and the new segmentation. If the new segmentation is considered to be superior to the current segmentation the label image is updated by replacing the respective segmented region recorded in the database with the outcomes of the new algorithm.
3. **New image addition**: The reference segmentation algorithms for all segmentation types R are applied to the new image i_n and the outcome of each segmented region s is visually evaluated to a quality grade a(i_n, s). The database in then updated correspondingly by adding the new image i_n, its label image l(i_n) and the quality grade a(i_n, s) for each segmented region s to the image set I, the label image set L and the label assessment set A respectively

For the most precise results for a new algorithm the database should first be revised by that algorithm before it is evaluated; however, that involves a cycle of visual inspection and other algorithms would need to be evaluated on the new database for performance comparisons. For algorithm development it may be useful to select a subset of the database for evaluation; typically this subset is made of cases for which the segmentation is rated as acceptable or unacceptable. It is possible to sequester a partition of the database for blind algorithm evaluations if necessary.

2.1.2 Customized Visualizations

Algorithm outcomes are graded for a score *a*(i,s) by a two-stage VE process. An image is first evaluated by a 3D customized visualization and, for additional review when necessary, a more traditional 2D image slice viewing is provided. An example for whole lung segmentation is show in Fig. 1.

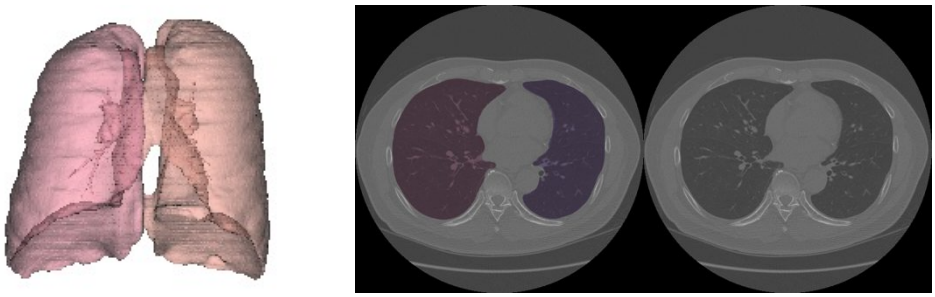


Fig. 1 Two lung 3D visualization (left) and a single image slice of the traditional 2D slice-based visualization with the original image (right).

In most cases a quick review of the 3D visualization is adequate. If needed a pixel-by-pixel traditional 2D image review is available. The 3D review is carefully customized for the segmentation of interest. For the database revision additional comparative reviews of the two image segmentations are available.

In the current VEQR method, a region segmented by an algorithm is evaluated by visual evaluation into three categories: good (G), acceptable (A) and unacceptable (U). The criterion for acceptable is that although there is some visible defect in the segmentation, the quality is sufficient for evaluation of related biomarker measurements. An example of an unacceptable segmentation is shown in Fig. 2. This is clearly visible in the 3D visualization alone. (The problem for this case is visible in the 2D review but this is not needed for database documentation.)

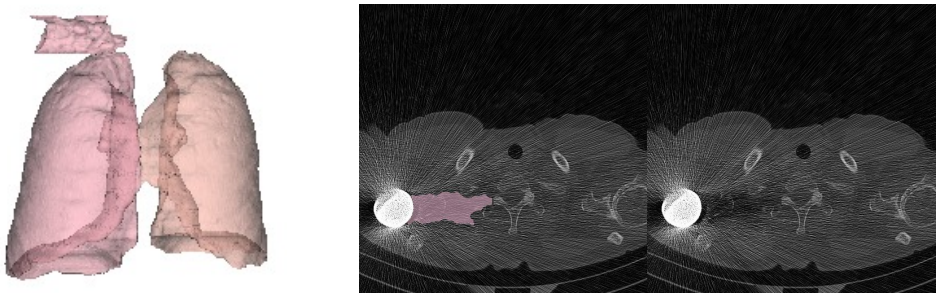


Fig. 2 Example unacceptable lung segmentation in 3D (left) and 2D (right).

2.2 Segmentation Evaluation of Low-dose CT Chest Images

The VEQR method was tested on a database of 7440 whole lung thin-slice non-contrast CT images that were used in our ongoing research. Algorithms for six different image segmentation tasks were evaluated on all images to establish their performance a on a large image dataset. Prior to testing with the very large database these algorithms were first evaluated on a development dataset with traditional expert image annotations for QE and a focus on minimizing Me type segmentation errors.

In our segmentation system a top-down strategy is employed where simple robust segmentations are determined first then subsequent algorithms take advantage of previously segmented regions. The precedence dependency relationships for these algorithms are in Fig. 3. The six algorithms are summarized below and the VEQR criterion for segmentation quality category is also specified.

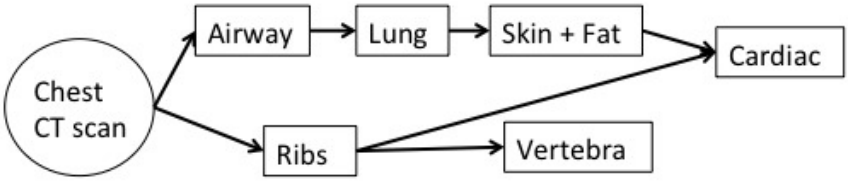


Fig. 3 Dependency relationship of the segmentation algorithms: the Cardiac algorithm requires good outcomes from both the Ribs and Skin algorithms while all other algorithms have a single dependency.

1. Airway

A fully automated algorithm is performed on chest CT scans to segment the airway tree using a cylinder tracking and 3D region growing based method, which allows for accurate detection of leakage by growing regions within a locally-defined envelope[9]. A coronal view of the airway segmentation as shown in in Fig. 4 is used for the VEQR, where the trachea and the remainder of the airway tree are shown in different colors, thereby enabling the validation of the location of carina at the same time. Airway biomarkers are the airway diameter and wall thickness evaluated for each detectable segment. The grading criteria for the airway algorithm are: G: no visible errors; A: correct trachea and two main bronchi but unable to segment sufficient peripheral branches; U: visible leakage.

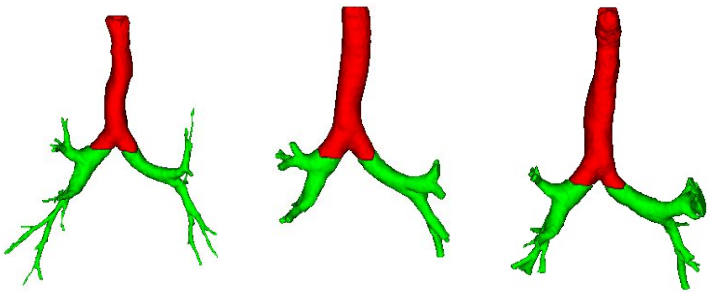


Fig. 4 Customized visualization for airway tree segmentation: trachea to carina, red; bronchi, green.

2. Lungs

The left and right lungs are segmented using image filtering, intensity thresholding and morphological operations[10]. The left and right lungs are partitioned with a minimum distance path-cutting algorithm. Coronal visualization of the two lungs in different colors is shown in Fig. 5. Lung region biomarker evaluations include the detection of pulmonary nodule and the measurement of lung health indicators such as the emphysema index. The grading criteria for the lung algorithm are: G: no visible errors; U: major errors, e.g. other region mistaken as lung.

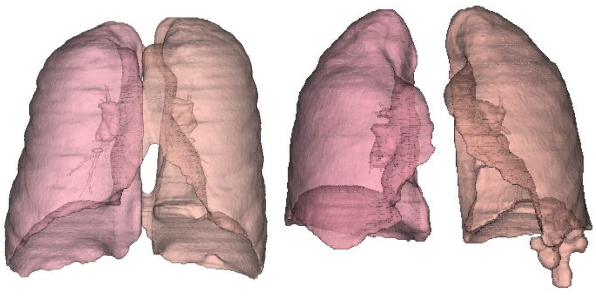
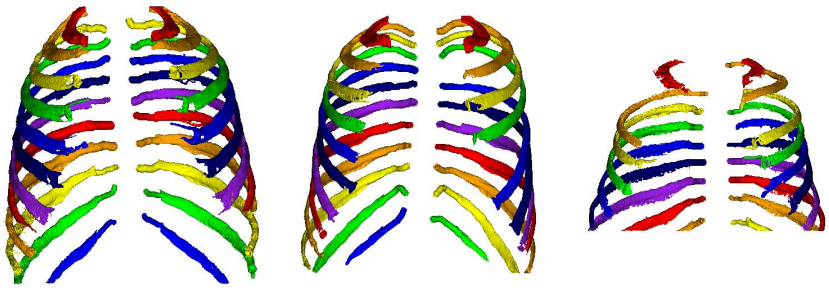


Fig. 5 Customized visualization for segmentation of separate lungs: left and right lung in different shades of pink.

3. Individual ribs

Individual ribs are segmented from low-dose chest CT scans by employing an algorithm based on region growing and cylinder tracking[11]. 3D visualizations of rib segmentation in coronal views are adopted for the VEQR as shown in Fig. 6, where the individual ribs are coded by different colors for the validation of individual rib labeling. There are no biomarkers associated with the ribs; their segmentation is currently used to facilitate the segmentation of other organs. The grading criteria for the rib algorithm are: G: no visible errors; A: minor over- or under-segmentation; U: missing multiple ribs or other major errors.

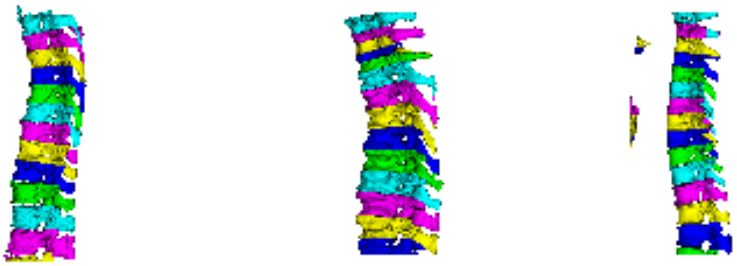


Good Acceptable Unacceptable

Fig. 6 Customized visualization for segmentation and labeling of ribs: individual ribs colored differently.

4. Individual vertebra

The whole spine is first segmented by thresholding and connected component analysis, and then further divided into individual vertebra by fitting separating planes in the 3D space[12]. VEQR of the individual vertebra segmentation is performed on the sagittal view of the 3D segmentation as shown in Fig. 7, where different colors are used to indicate individual vertebra for the purpose of validating the vertebra labeling. Vertebra biomarker evaluations include the measurement of bone mineral density and the prediction of compression fracture. The grading criteria for the vertebra algorithm are: G: no visible errors: A: minor over-segmentation; U: major errors including mislabeling or failure to separate vertebra.

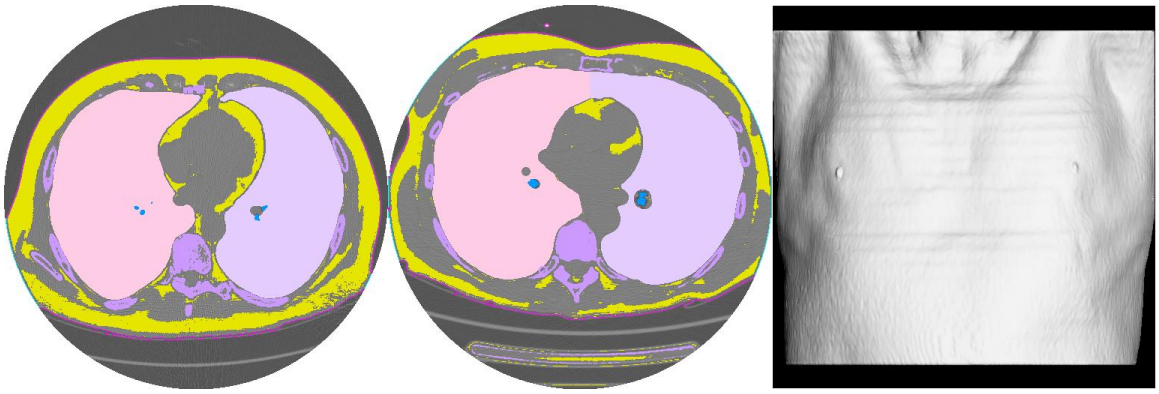


Good Acceptable Unacceptable

Fig. 7 Customized visualization for segmentation of individual vertebra: individual vertebra colored differently.

5. Skin surface

Skin surface includes the segmentation of skin surface and fat tissues. The skin and image boundary is segmented by differentiating body from outside air; and the fat is segmented using a local noise-aware algorithm[13]. VEQR of skin and fat is performed in an axial slice across the center of the scan as shown in Fig. 8. A coronal visualization of the reconstructed skin surface is also available (see Fig. 8 right image). There are no biomarkers directly associated with skin surface or fat content. They are used to facilitate the segmentation of other organs. The grading criteria for the skin and fat algorithm are: G: no visible errors in skin surface or fat; U: visible errors, e.g. table labeled as skin or fat.

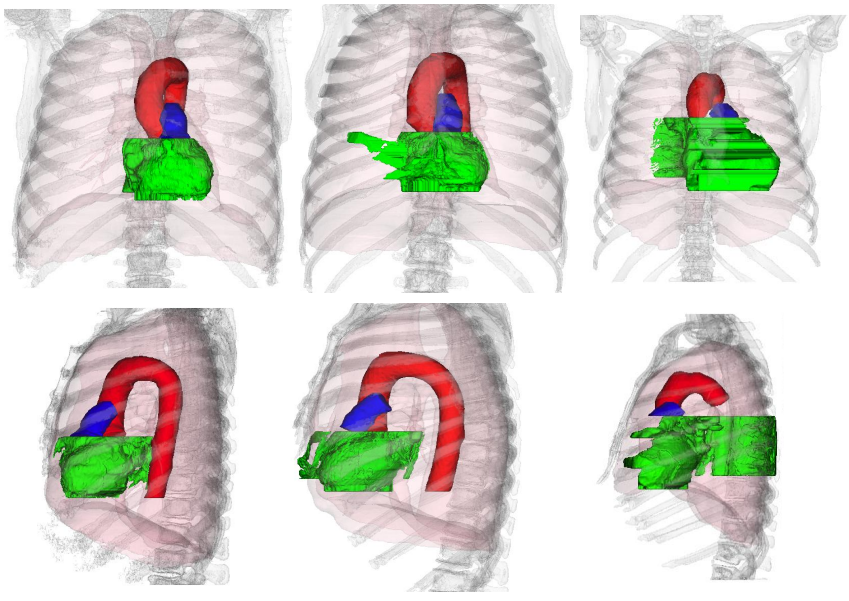


Good Unacceptable Skin surface

Fig. 8 Customized visualization for skin and fat segmentation: skin surface, purple; fat, yellow. Skin surface is also shown in a coronal view (right image).

6. Cardiac region

Cardiac region segmentation includes the segmentation of aorta, heart region, and pulmonary trunk. A cylinder tracking algorithm and triangular mesh model is used to segment the aorta[14] and the pulmonary trunk[15]. The heart region is mainly determined from constraints of adjacent segmented organs[16]. VEQR of cardiac region consists of the evaluation of coronal and sagittal views of the segmented aorta, heart region and pulmonary trunk in different colors as shown in Fig. 9. Cardiac region primary biomarkers are coronary artery calcification and aorta diameter profile. Other cardiac biomarkers include pulmonary trunk to aorta diameter ratio, aortic calcification and cardiac visceral fat. The grading criteria for the cardiac algorithm are: G: no visible errors in any of the 3 regions; A: visible errors but usable for measurement of biomarkers; U: major errors and not usable for biomarker measurements.



Good Acceptable Unacceptable

Fig. 9 Customized visualization for cardiac region segmentation: aorta, red; heart, green; pulmonary trunk, blue.

2.3 Algorithm Evaluation

We started with a set of six algorithms that had provided good outcomes on all cases of a small evaluation (< 100 cases) documented image dataset with no Ce errors. Evaluation on the large image dataset was performed in two stages: first development was refined on a dataset formed from two public image datasets: the VIA-ELCAP[17] and LIDC[18] with a

total of 364 CT scans. In detail, the VIA-ELCAP dataset contains 50 low-dose chest CT scans with similar image acquisition parameters, of which 4 scans were excluded due to severe image artifacts caused by metal implants. The LIDC dataset contains more than 1,000 CT scans with a great variation of image acquisition parameters. A subset of LIDC scans were selected as acceptable for the thin-slice protocol used in lung cancer screening using the following criteria: whole chest, non-contrast, slice thickness<=1.25mm and without severe image artifacts, resulting in a total of 318 CT scans. While all the VIA-ELCAP scans have a similar CT protocol, the LIDC dataset is highly heterogeneous with scans originating from several different locations.

Algorithms were then refined based on VEQR of segmentation outcome and re-evaluated on these scans. The process was repeated until all algorithms had more than 95% Good or Acceptable outcomes (agreement between two visual reviewers). From a review of the pilot outcomes the values for T_{DClow}, and T_{DChigh} may be determined, typical values are 0.75 and 0.95 respectively.

Following initial development on the public dataset the algorithms were then evaluated on the full datasets that contains an additional 7,076 3D CT images from 4,076 different subjects. All scans are whole chest, non-contrast, thin-slice (slice thickness<=2mm) and without severe image artifacts. Database revisions in this case were conducted with a single reviewer. There have been several cycles of algorithm revision over the last three years; in that time the database has increased by several thousand scans each year. Algorithm refinement has been aided by small testing with cohorts of cases with unsuccessful segmentations.

3. RESULTS

The VEQR results for all six segmentation algorithms are given in Table 1-2. Table 1 shows the VEQR results for the two public datasets VIA-ELCAP and LIDC (364 CT images, agreement between two reviewers). Table 2 shows the combined VEQR results for the full dataset of 7440 cases (the additional 7,076 CT images were reviewed by a single reviewer). Initially there were between 13 and 24 unsuccessful segmentations across algorithms for the public datasets. As can be seen from Table 1, this was reduced to 0-8 cases after algorithm development.

Table 1. VEQR results for VIA-ELCAP and LIDC.

Images		Airway	Lung	Ribs	Vertebra	Skin Surface	Cardiac Region
364	G	364 (100%)	364 (98%)	347 (95%)	300 (82%)	363 (99.7%)	333 (91%)
	A	0 (0%)	0 (0%)	10 (3%)	56 (15%)	0 (0%)	26 (7%)
	U	0 (0%)	0 (0%)	7 (2%)	8 (2%)	1 (0.3%)	5 (1%)

Table 2. VEQR results for all datasets.

Images		Airway	Lung	Ribs	Vertebra	Skin Surface	Cardiac Region
7440	G	7069 (95%)	7298 (98%)	6443 (87%)	6025 (81%)	7436 (99.9%)	6007 (81%)
	A	88 (1%)	0 (0%)	489 (6%)	893 (12%)	0 (0%)	967 (13%)
	U	283 (4%)	142 (2%)	508 (7%)	522 (7%)	4 (0.1%)	466 (6%)

4. DISCUSSION

The algorithms obtained a result of over 90% good or acceptable for a first round evaluation. The cases in which the algorithm failed or rated acceptable provide a rich source for algorithm improvements. The most mature algorithms for lung and skin surface, achieved good performance on over 98% of the cases; these algorithms have now had the benefit of several update cycles.

The effort to maintain the database will only be manageable if VEQR is efficient and the number of evaluations required for revisions are minimized. The careful definition of the 3D visualization is important for the efficiency of the review process. Examples of visualization for segmentations are shown in Fig. 4-9. In general these visualizations make possible the grading of several region segmentations per minute. Further, care must be taken to minimize the number of revision reviews. To accomplish this the revision triggering parameters need to be carefully determined; in addition, new algorithms must report an a performance of over 90% good on the pilot test of VIA-ELCAP and LIDC cases before they are considered of high enough quality for database revisions.

We have been using the described system for about three years; to date we have not had the opportunity to use the system on algorithms other than our own so the full benefits of the system have not yet been demonstrated. Each year we have been able to add several thousand images to our database; for next year we plan to increases the database to well over 10,000 images. For our research program we have improved all algorithms and have made an update to the reference algorithms on an approximately annual basis. We have found that the large number of images have facilitated algorithm modifications with respect to robustness, (i.e., Ce type errors). For example, one new image cohort had scans with a 2 mm slice thickness that was larger than the 0.5 to 1.5 mm range that we typically use. The algorithm failures were addressed by adjusting an algorithm parameter, which incurred more execution time. We have now linked this parameter to image slice thickness so that the algorithm only makes this adjustment when needed. This has greatly improved the algorithm robustness for these images without requiring any modifications (or increase in execution time) for the thinner slice images. A second example is a recent cohort of images that had significantly more image noise than we had previously encountered. We have now addressed that situation by using additional image filtering when high noise images are detected again without additional time or complexity for the legacy images. In this way we have been able to incrementally make the algorithms achieve better outcomes and adapt to new and changing image properties as we augment the image database with new images.

5. CONCLUSION

Very large documented image databases require different strategies than the traditional VE and QE evaluation methods. The VEQR method presented here meets the need for efficient large-scale evaluation and provides automated quantitative assessment of algorithm performance. This method avoids the variation errors introduced by manual marking by involving experts in a visual review capacity only. It also provides an efficient mechanism for updating target segmentations and for the addition of new image data.

To address big data, the VEQR paradigm involves quantitative performance thresholds and customized visualizations to dramatically reduce the burden of visual inspection for database revisions and updates. This method has been demonstrated with an example application of low-dose chest CT images by evaluating 6 different segmentation algorithms on 7,440 whole-lung CT images achieving a successful segmentation rate of 93-99%.

Acknowledgements

The authors would like thank C. Henschke, D. Yankelevitz and R. de la Hoz of Mount Sinai medical School, A. Miller of Queens College, and the Lung Image Database Consortium (LIDC) for making the CT images available for this study.

References

[1] Sullivan, D. C., Obuchowski, N. A., Kessler, L. G., Raunig, D. L., Gatsonis, C., Huang, E. P., Kondratovich, M., McShane, L. M., Reeves, A. P., Barboriak, D. P., Guimaraes, A. R., Wahl, R. L. and RSNA-QIBA Metrology Working Group, "Metrology Standards for Quantitative Imaging Biomarkers," Radiology, 277(3), 813-825 (2015).
[2] Reed, V.K., Woodward, W. A., Zhang, L., Strom, E. A., Perkins, G. H., Tereffe, W., Oh, J. L., Yu, T. K., Bedrosian, I., Whitman, G. J., Buchholz, T. A. and Dong, L., "Automatic segmentation of whole breast using atlas approach and deformable image registration," Inl J Radiation Oncology Biology Physics, 73(5), 1493-1500 (2009).
[3] Reeves, A. P., Biancardi, A. M., Apanasovich, T. V., Meyer, C. R. MacMahon, H., van Beek, E. J., Kazerooni, E. A., Yankelevitz, D., McNitt-Gray, M. F., McLennan, G., Armato, S. G. 3rd, Henschke, C. I., Aberle, D. R., Croft, B. Y. and Clarke, L. P., "The lung image database consortium (LIDC): A comparison of different size metrics for pulmonary nodule measurements," *Academic Radiology*, 14(12), 1475-1485 (2007).
[4] Haas, B., Coradi, T., Scholz, M., Kunz, P., Huber, M., Oppitz, U., Andre, L., Lengkeek, V., Huyskens, D., van Esch, A. and Reddick, R., "Automatic segmentation of thoracic and pelvic CT images for radiotherapy planning using implicit anatomic knowledge and organ-specific segmentation strategies," Phys Med Biol., 53(6), 1751-1771 (2008).
[5] Zhou, X., Ito, T., Zhou, X., Chen, H., Hara, T., Yokoyama, R., Kanematsu, M., Hoshi, H. and Fujita, H., "A universal approach for automatic organ segmentations on 3D CT images based on organ localization and 3D GrabCut," Proc. SPIE Medical Imaging 9035, 90352V (2014).
[6] Leader, J. K., Zheng, B., Rogers, R. M., Sciurba, F. C., Perez, A., Chapman, B. E., Patel, S., Fuhrman, C. R. and Gur, D., "Automated Lung Segmentation in X-Ray Computed Tomography: Development and Evaluation of a Heuristic Threshold-Based Scheme," Acad Radiol., 10 (11), 1224-1236 (2003).
[7] Takx, R. A. P., de Jong, P. A., Leiner, T., Oudkerk, M., de Koning, H. J., Mol, C. P., Viergever, M. A. and Isgum, I., "Automated Coronary Artery Calcification Scoring in Non-Gated Chest CT: Agreement and Reliability," PLOS ONE, 9(3), e91239 (2014).
[8] CMS, "Decision Memo for Screening for Lung Cancer with Low Dose Computed Tomography (LDCT)," https://www.cms.gov/medicare-coverage-database/details/nca-decision-memo.aspx?NCAId=274, Feb. 5, 2015 (accessed on Jan. 30, 2016).
[9] Lee, J. and Reeves, A. P., "Segmentation of the airway tree from chest CT using local volume of interest," In Proc. of Second International Workshop on Pulmonary Image Analysis, pp. 333-340 (2009).
[10] Enquobahrie, A., Reeves, A. P., Yankelevitz, D. F. and Henschke, C. I., "Automated Detection of Small Solid Pulmonary Nodules in Whole Lung CT Scans from a Lung Cancer Screening Study," Academic Radiology 14(5), 579-593 (2007).
[11] Lee, J. and Reeves, A. P., "Segmentation of individual ribs from low-dose chest CT," Proc. SPIE Medical Imaging 7624, 76243J (2010).
[12] Lee, J., "Automated analysis of anatomical structures from low-dose chest computed tomography scans," Dissertation, Cornell University (2011).
[13] Padgett, J., Biancardi, A. M., Henschke, C. I., Yankelevitz, D. F. and Reeves, A. P., "Local noise estimation in low-dose chest CT images," Int J CARS 9(2), 221-229 (2013).
[14] Xie, Y., Padgett, J., Biancardi, A. M. and Reeves, A. P., "Automated aorta segmentation in low-dose CT images," Int J CARS 9(2), 211-219 (2013).
[15] Xie, Y., Liang, M., Yankelevitz, D. F., Henschke, C. I. and Reeves, A. P., "Automated measurement of pulmonary artery in low-dose non-contrast chest CT images," Proc. SPIE Medical Imaging 9414, 94141G (2015).
[16] Xie, Y., Cham, M. D., Henschke, C., Yankelevitz, D. and Reeves, A. P., "Automated coronary artery calcification detection on low-dose chest CT images," Proc. SPIE Medical Imaging 9035, 90350F (2014).
[17] ELCAP Public Lung Image Database, www.via.cornell.edu/databases/lungdb.html, 2003 (accessed Jan. 30 2016).
[18] Armato III, S. G., McLennan, G. et al, "The Lung Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI): A Completed Reference Database of Lung Nodules on CT Scans," Medical Physics 38(2), 915-931 (2011).