HHS Public Access

Author manuscript

Proc SPIE Int Soc Opt Eng. Author manuscript; available in PMC 2018 August 13.

Published in final edited form as:

Proc SPIE Int Soc Opt Eng. 2018 February ; 10576: . doi:10.1117/12.2293622.

Image Quality and Segmentation

Gargi V. Pednekar[1], **Jayaram K. Udupa**[2], **David J. McLaughlin**[1], **Xingyu Wu**[2], **Yubing Tong**[2], **Charles B. Simone II**[3], **Joseph Camaratta**[1], and **Drew A. Torigian**[2]

[1]Quantitative Radiology Solutions, 3624 Market St., Suite 5E, Philadelphia PA 19104 United States

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

[3]Department of Radiation Oncology, Maryland Proton Treatment Center, University of Maryland School of Medicine, 850 W. Baltimore St., Baltimore, MD 21201 United States

Abstract

Algorithms for image segmentation (including object recognition and delineation) are influenced by the quality of object appearance in the image and overall image quality. However, the issue of how to perform segmentation evaluation as a function of these quality factors has not been addressed in the literature. In this paper, we present a solution to this problem. We devised a set of key quality criteria that influence segmentation (global and regional): posture deviations, image noise, beam hardening artifacts (streak artifacts), shape distortion, presence of pathology, object intensity deviation, and object contrast. A trained reader assigned a grade to each object for each criterion in each study. We developed algorithms based on logical predicates for determining a 1 to 10 numeric quality score for each object and each image from reader-assigned quality grades. We analyzed these object and image quality scores (OQS and IQS, respectively) in our data cohort by gender and age. We performed recognition and delineation of all objects using recent adaptations [8, 9] of our Automatic Anatomy Recognition (AAR) framework [6] and analyzed the accuracy of recognition and delineation of each object. We illustrate our method on 216 head & neck and 211 thoracic cancer computed tomography (CT) studies.

Keywords

Image quality; image segmentation; segmentation evaluation

1. Introduction

Many publicly available data sets, performance metrics, methods, and associated software under the name “Segmentation Challenges” exist for evaluating medical image segmentation algorithms. However, it is currently not possible to obtain a quantitative understanding of segmentation performance as a function of input image quality. Consequently, it is impossible to present a holistic picture of segmentation performance independent of input-image-specific vagaries due to unknown quality. We present a novel methodology to overcome this hurdle. For a holistic evaluation, it is important to define object and image quality metrics and segmentation evaluation metrics as a function of these quality metrics.

No such efforts seem to have been undertaken to date in segmentation challenges and other quantitative medical imaging application efforts.

We describe our method of segmentation assessment as a function of image and object quality in Section 2. We illustrate our method on 216 head & neck cancer CT studies and present results on these quality measures in Section 3. We summarize our conclusions in Section 4.

2. Methods

Data sets

We retrospectively created a database of a mix of contrast-enhanced and un-enhanced CT images and dosimetrist-drawn contours in 216 cancer studies in head-and-neck (H&N) from the Department of Radiation Oncology, University of Pennsylvania on this IRB-approved study. Image and contour data pertained to patients in four groups (54 studies in each group) – male and female in the age range 40–59 and 60–79. Voxel size in these data sets ranged from $0.93 \times 0.93 \times 1.5\ mm^3$ to $1.6 \times 1.6 \times .\ 3\ mm^3$.

We developed precise definitions of 11 key organs at risk (OARs) [5], by extending object definitions from recent guidelines [3,4], and modified contour data to fulfill these definitions. This turned out to a very arduous task since adherence to definitions is loose and the guidelines had many gaps which prevented them from being used directly for computational modeling of objects which require precise definitions.

Quality criteria

We devised a set of key quality criteria that influence segmentation (global and regional):

- body posture deviations
- image noise
- beam hardening artifacts (streak artifacts)
- shape distortion
- presence of pathology
- object intensity deviation
- object contrast

Some of these criteria are illustrated in Figure 1 for data sets from our cohort.

Quality metrics

A trained reader then assigned a grade to each object for each criterion in each study. We developed algorithms based on logical predicates for determining a 1 to 10 numeric quality score for each object and each image from reader-assigned quality grades. We analyzed these object and image quality scores (OQS and IQS) in our cohort by gender and age. We then described the performance of a segmentation method for any given metric over the entire quality score scale as a distribution of that metric. We performed recognition and

delineation of all objects using recent adaptations [8, 9] of our AAR framework [6] and analyzed the accuracy of recognition and delineation of each object as a function of OQS for each object considered in the body region and as a function of IQS at the study level.

Below, we present examples of the image and object quality criteria we developed as well as the basis of assigning grades to them. The quality criteria variables run from x_1 through x_9. These consist of four image-wise/ global (x_1 - x_4) and five object-specific/ local ($x_5 - x_9$) variables. To illustrate the level of detail involved, two examples are presented – two for global (x_1, x_4) and two for local (x_7, x_8) variables.

Criterion number/ logical variable	Criterion	Quality Grade
IQC1 (x_1)	Neck posture deviation	Neck normally positioned ($x_1 = 0$): Neck is in neutral position and properly aligned with the body. Neck not normally positioned ($x_1 = 1$): Note the ways in which the neck can deviate (flexion, extension, left/right rotation, left/right tilt). Threshold the image roughly for skin and visualize in 3D rendering to determine posture.
IQC4 (x_4)	Image noise	Not Present ($x_4 = 0$): Significant image noise is not visible in the body region. Present ($x_4 = 1$): Significant image noise is visible in the body region. Use soft-tissue window. Examine at body-region.
IQC7 (x_7)	Presence of pathology	Not Present ($x_7 = 0$): No pathology is visible inside the object. Minimal ($x_7 = 1$): Visible pathology occupies less than 25% of the object by volume. Severe ($x_7 = 2$): Visible pathology occupies greater than or equal to 25% of the object by volume.
IQC8 (x_8)	Object intensity deviation	Contrast enhanced: This should be treated as a different modality from non-contrast enhanced. Independent of the above, use criteria below. Glands: Lean (closer in attenuation to muscle than to fat, $x_8 = 0$) vs. Fatty (closer in attenuation to fat than to muscle, $x_8 = 1$). Mandible: Normal ($x_8 = 0$) vs. Either Diffusely lucent or Diffusely sclerotic ($x_8 = 1$). All other organs: Normal ($x_8 = 0$) vs. Abnormal ($x_8 = 1$). The vast majority of the organs will be normal.

Let $\Delta = (I, I_b)$ denote an image data ensemble for a body region B, where I is a set of acquired images of B and I_b is a set of binary images constituting a set of objects O in B in the images in I. That is, I_b contains a binary image corresponding to each object (OAR) O in O in each image I in I. Let $\Theta(q, O, I)$ denote the image quality grade determined for object O in image I for image quality criterion q (one of IQC1, …, IQC9 in the table). Given Δ and its grade assignment $\Theta(q, O, I)$ for the objects in O, we devised an Algorithm αO which generates object quality score $OQS(O, I)$ that reflects how well O is portrayed in I. Algorithm αI presented below subsequently generates image quality score $IQS(I)$ that

reflects the quality of image I considering the quality of portrayal of all objects in I. Note: $x_1 \ldots, x_9$ are all logical variables.

Algorithm αI presented below estimates $IQS(I)$ as the median of the object quality scores $OQS(O, I)$ over all objects in I.

Algorithm αI

Input: Object quality scores $OQS(O, I)$ for all O in O and I in I.

Output: Image quality scores $IQS(I)$ for all I in I.

Begin

S1. *For* each I in I, *do*

S2. Set $IQS(I)$ to be the median of the set of values $\{OQS(O, I): O \in O\}$;

S3. *End for*;

S4. Output $IQS(I)$ for all I in I;

End

3. Results

The number of scans in our cohort that were completely free of deviations with respect to the 9 image-quality criteria was 1 for H&N. The mean object quality score over all objects in the H&N is 3.9, with scores for 3 objects in the upper quartile and 8 in the lower quartile. A similar evaluation on 210 thoracic data sets showed a mean object quality score over all 12 objects in the thorax to be 5.7, with scores for 7 objects in the upper quartile and 5 in the lower quartile. Overall, H&N objects had a lower quality score than thoracic objects.

In Figure 2 we show sample OQS distributions as well as IQS distribution. Figure 2 shows that OQS tends to cluster around the lower and upper ends of the scale. Also, objects for younger patients seem to have better quality than older subjects except for the orohypopharynx constrictor muscle where the opposite seems to be true. This object consistently showed the poorest score among all objects. IQS for male and younger subjects seems to be better than that for females and older subjects.

As an example, in Table 1 we list the recognition and delineation results obtained on H&N data sets [8, 9] as a function of object quality. Because of the clustering of OQS at the lower and higher end of the scale, we roughly divided objects into high-quality and low-quality groups based on OQS. High-quality in this instance means scores greater than 7. This roughly translated to objects with streak artifacts or other deviations in not more than 3 slices through the object. The table shows that for objects with OQS in the upper end of the scale, recognition accuracy for the very challenging H&N region is about 1.5 voxels, Dice coefficient (DC) for delineation is about 0.8, and Hausdorff distance (HD) is about 1.5 mm. There is no statistically significant difference in accuracy between the two gender groups. DC is known to be very sensitive to errors in small objects, HD being a more robust measure. There is considerable variation in dosimetrists' contouring (not shown here), as determined by our separate experiment where two dosimetrists outlined 5 H&N OARs twice. The above accuracy from our system is well within the range of this variability.

The majority of the H&N objects are affected by strong streak artifacts as shown in Figures 1 and 2. The percentage of object samples with major streak artifacts in the H&N were 12.08%, whereas 3.02% cases were affected in the thorax. We observed a higher influence of pathology and shape distortion on the thorax with 24.67% cases being affected compared to 11.86% for the H&N region. Consequently, when OQS is low, both recognition and hence delineation accuracy suffer.

4. Conclusions

The logical predicate can be adapted to the requirements of each application. The proposed holistic assessment of performance may allow for selection of segmentation systems that are optimally suited to the image/object quality distribution underlying a given application/ imaging center. The approach shows promising opportunities for monitoring algorithm performance in an unsupervised setting with future improvements of using machine learning for image quality criteria detection and classification. We now understand that OQS and IQS play different roles in segmentation. They influence object recognition (localization) and delineation in different ways. Multi-object segmentation methods may differ in their performance on different objects which can be well captured via OQS. OQS seems to be a more useful factor than IQS for segmentation evaluation although IQS is useful to understand overall image quality and segmentation performance.

Acknowledgments

This work is supported by grants from the National Science Foundation [IIP1549509] and National Cancer Institute [R41CA199735-01A1].

References

1. Wu X, Udupa JK, Odhner D, Tong Y, Pednekar GV, McLaughlin DJ, Simone CB, , IICamaratta J, Torigian DA. e-Rekha: A high-performance software system for auto contouring head and neck anatomy in adaptive radiation therapy. Abstract accepted for presentation at 59th Annual meeting of American Society for Radiation Oncology; San Diego, ASTRO. 2017.
2. Wu X, Udupa JK, Odhner D, Tong Y, Pednekar GV, McLaughlin DJ, Simone CB II, Camaratta J, Torigian DA. Knowledge-based auto contouring for radiation therapy: Challenges in standardizing object definitions, ground truth delineations, object quality, and image quality, Abstract accepted for presentation at 59th Annual Meeting of American Society for Radiation Oncology, San Diego, ASTRO 2017. Brouwer CL. et al. Radiotherapy and Oncology. 2015; 117:83–90. [PubMed: 26277855]
3. Brouwer CL, Steenbakkers RJHM, Bourhis J, Budach W, Grau C, Grégoire V, van Herk M, Lee A, Maingon P, Nutting C, O'Sullivan B, Porceddu SV, Rosenthal DI, Sijtsema NM, Langendijk JA. CT-based delineation of organs at risk in the head and neck region: DAHANCA, EORTC, GORTEC, HKNPCSG, NCIC CTG, NCRI, NRG Oncology and TROG consensus guidelines. Radiotherapy and Oncology. 2015; 117:83–90. [PubMed: 26277855]
4. Brouwer CL, Steenbakkers RJHM, Bourhis J, Budach W, Grau C, Grégoire V, van Herk M, Lee A, Maingon P, Nutting C, O'Sullivan B, Porceddu SV, Rosenthal DI, Sijtsema NM, Langendijk JA. CT-based delineation of organs at risk in the head and neck region: DAHANCA, EORTC, GORTEC, HKNPCSG, NCIC CTG, NCRI, NRG Oncology and TROG consensus guidelines. Radiotherapy and Oncology. 2015; 117:83–90. Supplemental Material 1 & 2. [PubMed: 26277855]
5. Wu X, Udupa JK, Torigian DA. Definition of objects in the cervical region, Internal Report, Medical Image Processing Group. Department of Radiology; University of Pennsylvania, Philadelphia: Jan, 2017

6. Udupa JK, Odhner D, Zao L, Tong Y, Matsumoto MMS, Ciesielski KC, Falcao AX, Vaideeswaran P, Ciesielski V, Saboury B, Mohammadianrasanani S, Sin S, Arens R, Torigian DA. Body-wide hierarchical fuzzy modeling, recognition, and delineation of anatomy in medical images. Medical Image Analysis. 2014; 18:752–771. [PubMed: 24835182]
7. Grevera G, Udupa JK, Odhner D, Zhuge Y, Souza A, Iwanaga T, Mishra S. CAVASS: A computer-assisted visualization and analysis software system. Journal of Digital Imaging. 2007; 20(Supplement 1):101–118. [PubMed: 17786517]
8. Tong Y, Udupa JK, Wu X, Odhner D, Pednekar GV, Simone CB, McLaughlin D, Apinorasethkul C, Shammo G, James P, Camaratta J, Torigian DA, Torigian. Hierarchical model-based object localization for auto-contouring in head and neck radiation therapy planning. Proceedings of SPIE Medical Imaging. 2018 to appear.
9. Wu X, Udupa JK, Odhner D, Tong Y, Pednekar GV, McLaughlin D, Camaratta J, Torigian D. AAR auto contouring for Head and Neck Region from CT images. Proceedings of SPIE Medical Imaging. 2018 to appear.

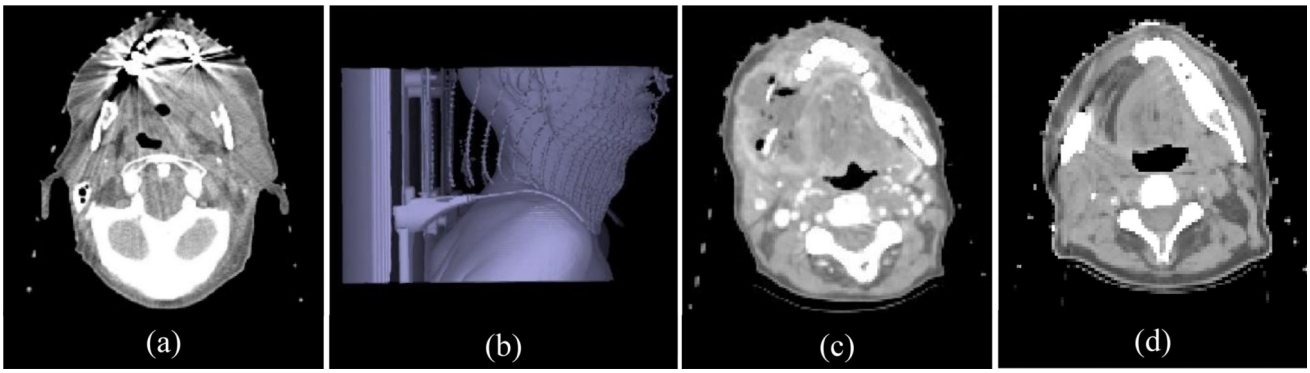

Figure 1.
Illustration of several quality criteria on data sets from our cohort. (a) Streak artifacts, (b) Mouth and neck posture deviation, (c) Pathology, (d) Shape distortion.

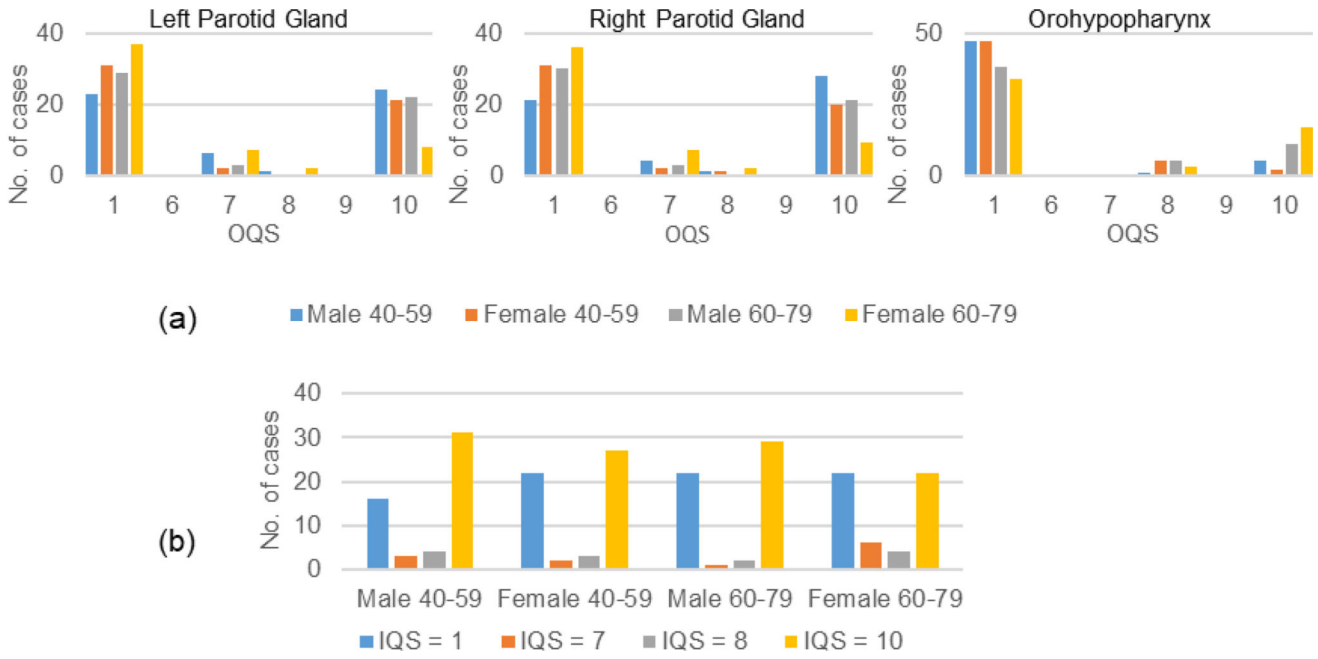

Figure 2.

Distribution of (a) Object quality score (OQS) for three objects, and (b) Image quality score (IQS) for the Head and Neck data set.

Table 1

Location error (LE) in mm and scale error (SE) for recognition and Dice Coefficient (DC) and Hausdorff Distance (HD) for delineation. “High” and “Low” refer to OQS in the high and low ranges in the distributions shown above. Mean and SD values over tested samples are listed. Column labeled All shows mean and SD values over all objects.

		SB	SB superior	SB inferior	SC	LX	LPG	RPG	LSG	RSG	MD	OHP	ES	All
High	LE	4.86	2.84	3.43	3.89	3.23	4.27	4.24	3.37	3.11	4.47	3.79	4.14	3.84
		0.82	1.28	1.3	1.69	2.04	1.73	1.62	1.46	1.64	1.03	1.36	1.39	1.41
	SE	1.0	0.99	1.0	0.99	1.29	1.14	1.11	0.93	0.93	1.03	1.28	0.70	1.03
		0.03	0.02	0.04	0.1	0.12	0.06	0.08	0.12	0.14	0.05	0.18	0.1	0.08
Low	LE	7.07	5.52	5.98	8.29	17.38	21.53	15.15	13.78	12.83	9.24	6.77	17.84	11.87
		1.07	2.79	4.07	4.54	6.30	14.30	7.27	6.58	10.85	5.16	3.47	7.87	6.78
	SE	1.01	0.99	0.99	0.92	1.17	1.28	1.25	1.05	0.88	1.08	1.26	0.77	1.05
		0.03	0.01	0.05	0.03	0.12	0.09	0.15	0.13	0.13	0.08	0.10	0.09	0.08
High	DC	0.98	0.98	0.96	0.75	0.74	0.75	0.72	0.72	0.72	0.88	0.58	0.62	0.78
		0.01	0.0	0.01	0.03	0.04	0.05	0.06	0.05	0.03	0.03	0.04	0.08	0.03
	HD	1.02	1.0	1.0	1.34	1.9	1.73	2.07	1.48	1.47	1.03	1.56	1.51	1.43
		0.0	0.0	0.0	0.12	0.31	0.33	0.69	0.28	0.25	0.08	0.24	0.34	0.22
Low	DC	0.97	0.97	0.95	0.60	0.61	0.52	0.61	0.53	0.45	0.81	0.53	0.36	0.66
		0.01	0.00	0.02	0.12	0.12	0.22	0.14	0.14	0.27	0.06	0.10	0.21	0.12
	HD	1.02	1.00	1.00	1.75	3.72	3.88	2.85	2.58	3.30	1.28	1.78	2.77	2.24
		0.00	0.00	0.00	0.47	1.44	1.89	0.69	0.79	1.90	0.40	0.39	1.19	0.76

Object abbreviations in Table 1: SB: Skin outer boundary, SC: Spinal Canal, LX: Larynx, LPG: Left Parotid Gland, RPG: Right Parotid Gland, LSG: Left Submandibular Gland, RSG: Right Submandibular Gland, MD: Mandible, OHP: Orohypopharynx constrictor muscle, ES: Esophagus.