Title

Flexible methods for segmentation evaluation: results from CT-based luggage screening.

Permalink

https://escholarship.org/uc/item/83f826r8

Journal

Journal of X-Ray Science and Technology, 22(2)

ISSN

0895-3996

Authors

Karimi, Seemeen
Jiang, Xiaoqian
Cosman, Pamela
et al.

Publication Date

2014

DOI

10.3233/xst-140418

Peer reviewed

Published in final edited form as:

J Xray Sci Technol. 2014 ; 22(2): 175–195. doi:10.3233/XST-140418.

Flexible methods for segmentation evaluation: Results from CT-based luggage screening

Seemeen Karimi[a,*], **Xiaoqian Jiang**[a], **Pamela Cosman**[a], and **Harry Martz**[b]

[a]University of California, San Diego, CA, USA

[b]Lawrence Livermore National Laboratories, Livermore, CA, USA

Abstract

BACKGROUND—Imaging systems used in aviation security include segmentation algorithms in an automatic threat recognition pipeline. The segmentation algorithms evolve in response to emerging threats and changing performance requirements. Analysis of segmentation algorithms' behavior, including the nature of errors and feature recovery, facilitates their development. However, evaluation methods from the literature provide limited characterization of the segmentation algorithms.

OBJECTIVE—To develop segmentation evaluation methods that measure systematic errors such as oversegmentation and undersegmentation, outliers, and overall errors. The methods must measure feature recovery and allow us to prioritize segments.

METHODS—We developed two complementary evaluation methods using statistical techniques and information theory. We also created a semi-automatic method to define ground truth from 3D images. We applied our methods to evaluate five segmentation algorithms developed for CT luggage screening. We validated our methods with synthetic problems and an observer evaluation.

RESULTS—Both methods selected the same best segmentation algorithm. Human evaluation confirmed the findings. The measurement of systematic errors and prioritization helped in understanding the behavior of each segmentation algorithm.

CONCLUSIONS—Our evaluation methods allow us to measure and explain the accuracy of segmentation algorithms.

Keywords

Segmentation evaluation; computed tomography; luggage screening; feature recovery

1. Introduction

Modern applications of image segmentation have complex goals, such as extracting multiple objects from a cluttered setting. The objects' image-based features must be computed for decision-making or subsequent processing. For example, in medical imaging, non-

[*]Corresponding author: Seemeen Karimi, University of California, San Diego, California, CA, USA. Tel.: +1 617 953 1662; seemeen.karimi@gmail.com.

destructive testing or luggage scanning, qualitative or theoretical evaluation is not enough, and quantitative evaluation is necessary. Various evaluation methods have been suggested based on the accuracy of edges, goodness measures within labeled regions, feature recovery, or on divergence from a ground truth (GT) [1–24]. The appropriateness of the methods is application-dependent, and frequently, there is no single best measure.

Our objective is the performance evaluation of segmentation algorithms applied to x-ray computed tomography (CT) scans of luggage. In aviation security, checked passenger luggage is imaged by CT-based explosives detection systems (EDS). In EDS, objects are segmented from the CT images and then sent to automatic threat recognition (ATR) algorithms. An important problem in this field is that non-threat objects may produce false alarms. Resolving false alarms involves high labor cost because false-alarm bags must be unpacked or sent for secondary screening. The main difficulties for accurate segmentation in luggage screening are the variety and heterogeneity of non-threat and threat objects found in bags. Another difficulty is image artifacts. These difficulties cause segmentation algorithms to split an object into multiple pieces, or to merge different objects into a single one.

The U.S. Department of Homeland Security has identified requirements for future systems, including increasing threat categories and lowering false alarms [25]. To encourage the development of new segmentation algorithms for CT security systems, a database of CT images of suitcases was generated by the ALERT group at Northeastern University, and distributed to five research groups at universities and corporations [26]. The database contained no threats; the requirement was to segment all objects present in each suitcase. Segmentation results for a sample of this data were obtained for detailed quantitative evaluation. In this project, objects missed by the segmentation algorithm correspond to type II error (false negative) in binary classification and spurious objects created by the segmentation algorithm correspond to type I error (false positive).

Quantitative evaluation of segmentation algorithms is a challenging task in luggage screening because multiple splits and merges are possible. In addition to an accuracy score, we would like to gain a deeper understanding of the algorithms’ behavior. First, we would like to know if an algorithm systematically oversegments or undersegments images or if the error is random. A knowledge of systematic errors allows us to tune the parameters of a segmentation algorithm, or supplement the segmentation algorithm with additional steps such as region merging [27]. Second, the ability of a segmentation algorithm to capture object features must be evaluated, because evaluation of object features is critical in ATR. Third, since it is often more important to correctly segment some objects than others, a method to assign priorities to segments is desirable when evaluating the algorithm. Priorities may be based on image intensity, homogeneity, particular texture or any other image features that define objects of interest. Fourth, a segmentation algorithm may have varying accuracy across the feature range, and this knowledge can be used to establish confidence in a given segment. There can be no restriction on the number or nature of objects. All these considerations are important in luggage scanning but are not adequately addressed by existing evaluation literature.

Various goodness measures have been proposed to evaluate a segmentation without a GT [1,2,19]. The goodness measures are based on entropy, intra-region similarity and inter-region discrepancy, surface smoothness and other properties of regions. However, objects found in luggage are inherently heterogeneous, i.e., made up of different materials that have different textures and attenuation properties. Their sizes and shapes are varied and unpredictable. Therefore, goodness measures are not applicable to our problem.

There are many methods that evaluate segmentation against GT by computing a distance between the sets of edge pixels [1,3,4] or surface voxels [24]. However, edge or surface distances do not measure feature retrieval. Mass or volume may be well retrieved, but have large edge distances due to artifacts or other segmentation errors. Therefore, using discrepancy between sets of edges does not appear to be a good solution for luggage screening.

An error measure was defined to measure the discrepancy among manual segmentations performed by multiple humans. This measure was designed to be unaffected by refinements [5]. In the luggage application, object splitting and merging replace refinement errors, and cannot be considered alternate truths. Therefore, we tested another error measure created with the objective of quantifying the splitting and merging, called the object consistency error (OCE) [6]. OCE is sensitive to refinement. We found that OCE does not perform well with split and merge errors that are not simple refinements. This issue is illustrated in Section 4 using synthetic examples.

Another method breaks down the evaluation problem into the identification of correct detection, over-segmentation, undersegmentation, missing objects, and spurious detection [7]. However, the method depends on the existence of planar surfaces in the image. The measures discussed in [8] interpret the different labeled regions as clusters, and measure distance between clusterings. The wide range of the number of labels from the different machine segmentation (MS) algorithms makes it unsuitable to apply pair-wise clustering interpretations to the labels (such as the Rand index [28]) because some of the labels have small cardinality.

The segmentations may be viewed as different partitionings. A metric was defined as the minimum number of elements that must be moved from one partitioning in order to get to the other partitioning [9]. In these methods, no calculation of systematic errors, no feature-based evaluation and no assignment of priorities among partitions was described, which is needed for our application and may be important in others. Other single-valued measures include [21–23].

The evaluation methods cited above are based on volume or surface overlap. However, there are other features of objects that are more relevant for ATR than volumes or surfaces. A measure called ultimate measurement accuracy (UMA) computes a distance between the measurements of a feature made in the GT and segmented images [10]. This measure works on single foreground objects. A multidimensional evaluation is in [18] but systematic errors are not quantified. The treatment of segmented images as probability mass functions was suggested for a 2-class problem [11]. The divergence measure is similar to the Kullback-

Leibler (KL) divergence. The idea was improved upon, to measure features of collections of similar objects by creating histograms of feature values for populations of similar objects, and comparing them using standard histogram comparison measures [12]. Similarly, image-distance measures were suggested in [14], based on feature distribution similarity.

We propose two new methods of evaluation to meet the application needs described above, and address many limitations of existing methods. In the first method, we calculate a weighted mutual information (WMI) of features from their joint distribution. In the second method, which we call Feature Descriptor Recovery (FDR), we measure systematic and overall errors in feature recovery, and extract additional information about behavior over feature ranges. The two methods provide different evaluation perspectives. They are flexible in that they can operate on features, they impose no restrictions on the type or number of segmented objects, and can prioritize segments by feature values or user preference.

In luggage screening, air, which occupies a large portion of the images, is not segmented. In this project, we treat air differently from other labels so that missing objects are penalized, but spurious objects are not. The spurious objects are bag parts and image quality verification phantoms present in the scans that were labeled by some of the research groups, but were not labeled in the GT. We do not want to penalize (or reward) these spurious objects. Our methods allow us to discard the spurious objects. This is different from the ATR problem, where the spurious objects are analogous to a false alarm. Although we do not specifically discuss false alarms due to the classified nature of threats and ATRs, it is certainly desirable for EDS vendors and testing agencies to assign penalties using appropriate weights within our framework. As we will see later, we cannot treat air as simply another label, because that would allow purely nominal scoring methods to reward missed objects.

We applied our evaluation methods to images from the ALERT dataset. Our evaluation methods were validated (1) by applying them to simple synthetic problems, and (2) by comparing the methods' results on suitcases with an evaluation done by a human observer. In information theory, the F_1 score is an accepted measure of performance for binary classification problems [29]. We have also compared our results against a multi-class generalization of the F_1 score.

2. CT images and ground truths

Suitcases were scanned on a volumetric medical CT scanner. A volume rendering of a CT image is shown in Fig. 1. The suitcases contain objects such as clothing, shoes, electronics, food, books, toys and various contained liquids. These suitcases do not contain threats or simulants. The CT image dimensions were 512 × 512 pixels per image slice, with about 800 slices in each (3D) image. Segmentation algorithms were developed by five different research groups for this project: Siemens Corporate Research, Marquette University, University of East Anglia, Stratovan Corp, and Tele-Security Systems. Each algorithm was run on five test images and generated a label image (except one image by one research group), so we have a total of 24 MS label images. Further details are in [26]. A description of the segmentation algorithms is beyond the scope of this paper but some references are available for the interested reader [30–32].

We developed a computer-assisted method to generate GT label images from volumetric CT images. This method was used by researchers at Northeastern university to generate ground truths from the suitcase images. Objects found in luggage are so varied in size and shape, and heterogenous in material composition, that human interaction is required to define the GT. However, segmentation performed solely by a human is limited in accuracy by the observer's ability to manually contour objects. The objects are not only three-dimensional, but sometimes hollow or thin and with large surface areas. Further, the objects are blurred by the transfer function of the CT scanner. It is therefore impractical to segment these objects by an exclusively manual method. In our method, manual contouring is complemented with manually-seeded region-growing, allowing the observer to segment complicated shapes. Manual segmentation by multiple observers has been addressed [33,34]. However, our challenge comes not from subjective perception, but rather from the difficulty of the manual task, which necessitates some automation. Another 3D GT generation method [17] uses mesh models. Our approach is simpler and does not require modeling.

A unique label is assigned to each object that is individually packed into the suitcase. For example, a liquid is assigned the same label as its container. While the validity of this rule may be argued, it overcomes the issue of subjective perception. There is no soft (probabilistic) label assignment for image voxels. A label value of zero indicates air, which is background. Objects with an average CT value of less than −500 Hounsfield units (HU), such as clothes, were assigned the label of air. In the HU scale, air is −1000 and water is zero.

Each bag image file contains hundreds of image slices. However, the user does not have to contour objects in every slice. Interpolation is performed between contours across slices, allowing the GT segmentation to be completed in roughly two hours per bag. The interpolated contours are filled in and labeled as one object. Objects in a bag are segmented one at a time, given unique labels and accumulated into a GT label image. Figure 2 shows a flowchart of the computer-assisted GT extraction process. There are two parallel paths. The right half of the flowchart illustrates the manual contouring, contour interpolation and contour filling processes. The output of this path is an image of a filled three-dimensional contour, labeled A in the flowchart. The left half illustrates the seeded region-growing path. The user selects seeds and region growing parameters for the current object. The output of this path is a region-grown mask in image B. The common voxels in images A and B are assigned a user-selected label value, Λ. The label image for this object is accumulated with previously segmented labels in image C. We use the maximum operation instead of a binary "or" operation to set a rule for overlapping labels. This rule is useful when there are touching objects. If the first object is labeled Λ_1, as the user segments the second object Λ_2, the two labels may overlap. The user resolves the problem by selecting $\Lambda_2 < \Lambda_1$ for the second object's label if he decides that the common voxels should belong to the first label or $\Lambda_2 > \Lambda_1$ if they should belong to the second one.

The flowchart in Fig. 2 was implemented in MeVisLab [35] a graphical programming language that provides image processing and visualization modules that can be connected together. In our program, DICOM images are read in and multi-planar reformatted (MPR) slices are generated and displayed. The user selects an MPR axis and contours an object in any slices along the chosen axis. We used a drawing tool that incorporates active contours. The active component helps the contour to be attracted to gradients or curvature as determined by user-defined penalties. A contour can be copied and pasted to other slices. The contours are linearly interpolated to all slices between the first and last contour. The observer subjectively decides whether the contour interpolation provides acceptable results. If the results are not acceptable, more contours can be added and interpolation repeated. The interpolated contours are filled in with a user-selected label value. Binary dilation is performed to include edge voxels.

In the second parallel path, the user selects seed voxels from the object, and sets upper and lower thresholds for region growing. The user can overlay the region grown mask on the CT image to decide whether the mask is acceptable, and modify seeds or thresholds if deemed necessary to repeat the region growing process.

3. Segmentation evaluation methods

We first describe the WMI for volume-based evaluation, and then describe our extension for mass-based evaluation. Next, we show the weighting functions that allow us to prioritize objects. Then we describe the FDR method which gives systematic errors and total error. Feature descriptors may also be weighted to prioritize objects. Finally, we describe the multi-class extension of F_1 scores, which we use for comparisons.

Let G and S denote the number of labels in the GT and the MS images, respectively, not including the air segment (label 0). Let $X_G(i)$ be the set of voxels in GT segment i, and $X_S(i)$ be the set of voxels in MS label i. We use the terms segment and label interchangeably.

3.1. Weighted Mutual Information (WMI)

Mutual Information (MI) can be used when the label images are expressed as joint and marginal probability densities [36]. We generate a confusion matrix from the GT and MS images. We first compute the MI and entropies without air so that MI, which is ordinal, does not reward the air label. Then we include the type II errors with a multiplicative factor. We neglect type I errors for the reasons explained in Section 1. Let $N_{G,S}(i, j)$ denote the number of voxels that belong to GT label i and MS label j.

Let $v_{G,S}(i, j)$ denote the joint probability mass function (pmf) based on volume:

$$v_{G,S}(i,j)=\frac{N_{G,S}(i,j)}{\sum_{k=1}^{G}\sum_{l=1}^{S}N_{G,S}(k,l)}, \quad 1 \leq i \leq G, 1 \leq j \leq S \tag{1}$$

We define the marginal pmfs for the GT and MS labels from the joint pmf.

$$v_G\ (i)=\sum_{j=1}^{S} v_{G,S}\ (i,j), \quad 1 \leq i \leq G \tag{2}$$

and

$$v_S\ (i)=\sum_{j=1}^{S} v_{G,S}\ (i,j), \quad 1 \leq i \leq G \tag{3}$$

A normalized mutual information score is generated in the following manner, as first described in [15]:

$$H=\frac{1}{Z}\sum_{i=1}^{G}\sum_{j=1}^{S} v_{G,S}\ (i,j)\log\frac{v_{G,S}\ (i,j)}{v_G\ (i)v_S\ (j)}, \tag{4}$$

where the normalization factor Z is the square root of the product of entropies, or the GT entropy if the MS entropy is zero:

$$Z=\begin{cases} \sqrt{\sum_{i=1}^{G} v_G\ (i)\log\left(\frac{1}{v_G\ (i)}\right)\sum_{j=1}^{S} v_S\ (j)\log\left(\frac{1}{v_S\ (j)}\right)}, & \text{if}\ S>1 \\ \sum_{i=1}^{G} v_G\ (i)\log\left(\frac{1}{v_G\ (i)}\right), & \text{otherwise} \end{cases} \tag{5}$$

We now incorporate type II errors. Referring to Fig. 3, we take the ratio of the total voxels in the inner matrix (dark shaded) to the total voxels in the outer matrix (all shaded).

$$r=\frac{\sum_{i=1}^{G}\sum_{j=1}^{S} N_{G,S}\ (i,j)}{\sum_{k=1}^{G}\sum_{l=0}^{S} N_{G,S}\ (k,l)} \tag{6}$$

This ratio is analogous to recall in a binary classification problem, if all objects were considered to belong to one class and air was considered the second class. Recall is also called sensitivity or true positive rate. The unshaded row contains type I error. We multiply this ratio, r, with H. To make this factor more general, an additional weight can be used so that missed data receive larger or smaller penalty. Our WMI score is given as

$$I=r \times H \tag{7}$$

We now use WMI to measure mass-based score, and then prioritize objects by weighting the confusion matrix. For the mass score, a confusion matrix cell contains not the number of voxels common to a pair of GT and MS labels, but rather, the common mass, which is calculated by summing values from the CT image. Let the CT image be denoted *C*. Then the mass in cell (*i, j*) is given by

$$M_{G,S}\left(i,j\right) \sum_{x \in X_G\ (i) \cap X_S\ (j)} C(x) \tag{8}$$

The joint and marginal pmfs for the mass-feature are computed from the confusion matrix in a manner similar to that shown for volume. The mass WMI score can be considered a weighted volume score, with weights equal to the CT number being assigned to each voxel.

Note that these calculated values of volume and mass should be multiplied by voxel size and CT scaling factors to obtain true volume and mass, but these multiplicative factors are constants and can be neglected. Aside from the constant scaling factors, the mass is not the true physical mass of the object, but an approximation. CT image intensity is proportional to the material linear attenuation coefficient, which itself is proportional to the physical density of the material if we neglect the atomic number of the material and the energy-dependence of the attenuation coefficient.

Next we describe using a weighted confusion matrix (Section 3.1.1) to weight objects and errors. Specifically, we demonstrate uniformity (a regional feature).

3.1.1. Weighted confusion matrix (WCM)—We assign priorities to segments by weighting the cells of the confusion matrix before computing WMI scores. We define weights that assign greater importance to homogenous (also called uniform) objects. Uniformity is not a feature of interest in ATR, but as we will show later, it demonstrates interesting behavior of the segmentation algorithms. The mass confusion matrix rows were weighted by a measure of uniformity. This measure of uniformity can be considered a texture feature, weighted by mass to prioritize heavier objects. Alternately, it can be considered a mass feature, weighted to prioritize homogenous objects.

$$w^{\sigma}_{G,S}\left(i,j\right)=\frac{1}{\sigma_G\left(i\right)}, \tag{9}$$

where $\sigma_G(i)$ is the standard deviation of the CT numbers in GT label *i* and is given by

$$\sigma_G(i)=\sqrt{\frac{1}{N_G\left(i\right)} \sum_{x \in X_G\ (i)} \left(C\left(x\right)-\overline{C_G}(i)\right)^2}, \tag{10}$$

where $N_G(i)$ denotes the number of voxels in GT segment i. In the above equation the mean CT number is given by

$$\overline{C_G}(i)=\frac{M_G(i)}{N_G(i)}, \quad (11)$$

where $M_G(i)$ is the mass within the GT label i given by

$$M_G(i)=\sum_{x\in X_G(i)} C(x) \quad (12)$$

A natural extension of this idea is cell-wise weighting. We also defined cell-wise weights with the goal of assigning non-uniform costs to different classification errors as shown below. The weights of the cells are lower if they are from dissimilar objects:

$$w_{G,S}^{cell}(i,j)=\frac{\min\left(\overline{C_G}(i),\overline{C_S}(j)\right)}{\max\left(\overline{C_G}(i),\overline{C_S}(j)\right)} \quad (13)$$

The cellwise weights were applied to the volume confusion matrix $v_{G,S}(i, j)$.

3.2. Feature descriptor recovery (FDR)

The FDR method measures how well the features of each object are recovered. Feature descriptors have more flexibility than the WMI framework because label-wise features can be used. For example, one can use label-averages or inter-label separation divided by intra-label uniformity. As before, weighting can be incorporated. The weighted features and uniformity are conceptually similar but not identical to those used in the WMI-score, as explained in Section 3.2.1.

The numbering of labels in the GT and MS images is arbitrary. We establish the optimal one-toone correspondence between the GT and MS labels. We used the Hungarian method [37] to maximize the total volume intersection between all GT labels and MS labels. Instead of the volume intersection, another cost function could have been used, such as the mass intersection.

A feature descriptor P_G is generated by calculating some feature within each label in the GT. P_G is a vector, such that for each label $1 \le l \le G$, $P_G(l)$ is the value of the feature computed with respect to the original CT image, within that label. Another feature descriptor P_S is generated for the MS. The feature descriptors P_G and P_S are generated independently of each other.

3.2.1. Features—Analogous to WMI scores, the features we have used are volume, mass, and uniformity, because of their relevance to ATR. As before, the volume of the label is the total number of voxels within the label and the mass of the label is the summed CT value within the label. Similar to the uniformity for WMI in Eq. (9), we define uniformity as the inverse of standard deviation multiplied by the mass, calculated per label. The uniformity feature is shown below for GT labels. It is also calculated for MS labels.

$$W_G^{\sigma}(l)=\frac{M_G(l)}{\sigma_G(l)}, \quad 1 \leq l \leq G \qquad (14)$$

This feature is similar but not identical to the uniformity-weighted mass of the WCM which had a row-wise weighting. In the WCM, it was not meaningful to consider the standard deviation of the voxels in a cell because a cell can have a small number of voxels.

In the FDR method, there is no weighting corresponding to cell-wise weighting w^{cell} of the WMI.

3.2.2. Feature recovery scatter plots—For each object in the MS image and GT image, we generate features as explained in the previous section. For a feature, we generate a scatter plot of the matched labels of P_S against P_G, and call this a feature recovery scatter (FRS) plot. In any bag, the number of GT and MS labels may not be the same, so the minimum is plotted. The data from all the bags were combined. As explained in Section 4, the slope of the line fitted to the data tells us if there are systematic errors. We have used a robust fit to reduce the impact of outliers [38].

3.2.3. Residual errors—In order to compute the residual errors from feature recovery, we applied commonly used error statistics, including Cramer-von-Mises (CVM) [39], Kullback-Leibler divergence (KL) [40] and L_1 error normalized by the sum of GT feature values. The L_1-based score is shown below.

$$R_{L_1}=0.5\frac{|P_G-P_S|_1}{|P_G|_1} \qquad (15)$$

Although the FRS plots contain the minimum of the number of labels in the MS and GT, the residual error is computed on the maximum of the number of labels. Where a label does not exist, its feature value is zero. The slope of the fitted line and the residual error together provide the performance result.

3.2.4. Behavior over feature range—In addition to over and undersegmentation, the pairing of segments allows us to investigate how accuracy changes over a feature range, and to identify outliers. We take the sliding average (geometric mean) of the feature ratio of the label pairs. The ratio is that of the larger to the smaller feature value. We plot this mean as a function of the sliding geometric mean of the GT labels. This plot indicates the average feature retrieval error against average feature value.

$$R(i)=\left(\prod_{j=-\frac{n-1}{2}}^{\frac{n-1}{2}}\frac{\max\left(P_{G}\left(i+j\right),P_{S}\left(i+j\right)\right)}{\min\left(P_{G}\left(i+j\right),P_{S}\left(i+j\right)\right)}\right)^{1/n},\frac{n-1}{2}<i\le\min\left(G,S\right)-\frac{n-1}{2} \qquad (16)$$

Ratios are more meaningful than differences in this computation because of the large dynamic range of the feature. In log-scales, this ratio would be the absolute value of the difference and the geometric mean would be the arithmetic mean, corresponding to taking a sliding L_1 error. This prevents opposite polarity errors from canceling.

From the FRS plots, we can obtain outliers. For each pair of GT and MS points, we compute the following distance.

$$d(i)=\log\left(\frac{P_{S}\left(i\right)}{P_{G}\left(i\right)}\right),\quad 1\le i\le\min\left(G,S\right) \qquad (17)$$

We fit a normal distribution to the distances and obtain its standard deviation, σ. Points $i:\|d(i)\|>3\sigma$ are considered outliers.

3.3. Multiclass F-score (F_1^m)

In information theory, the F-score is an accepted measure of performance for binary classification problems [29]. It is natural to explore a multiclass extension, F_1^m. We generated a multi-class extension of the score to help validate and offer some perspective on our results. The definition of F_1 score is

$$F_1=\frac{2pr}{p+r}, \qquad (18)$$

where r is recall and p is precision (also called positive predictive value). Standard definitions of recall and precision are

$$r=\frac{c}{c+c'}\ \text{and}\ p=\frac{c}{c+d}, \qquad (19)$$

where c is true positive, c' is the type II error, and d is the type I error. The luggage screening application has a multi-class segmentation problem. Therefore, the standard definition of the precision and recall, given in Eq. (19), cannot be used. Our multi-class adaptation defines recall and precision as

$$r=\frac{\sum_{i=1}^{G} N_{G,S}\ (i,j^{'}(i))}{\sum_{k=1}^{G}\sum_{l=0}^{S} N_{G,S}\ (k,l)},$$
$$p=\frac{\sum_{i=1}^{G} N_{G,S}\ (i,j^{'}(i))}{\sum_{k=1}^{G}\sum_{m=1}^{G} N_{G,S}\ (m,j^{'}(k))}$$

In the above equation $j^{'}(i)$ is the MS label that best matches GT label i as per the Hungarian algorithm matching. Using the equation for precision given above, we penalize missing portions of segments, missing segments, and split segments equally. The denominator may not include all the MS labels S, because that would penalize splitting more than missed detection (which is unreasonable).

4. Synthetic problems

To evaluate our measures against intuitive reasoning, we generated simple problems with different kinds of errors. We consider splitting, merging, partial splitting and merging, and missed objects (type II error). We do not consider spurious objects (type I errors) because we do not penalize them, as described earlier. The different test cases illustrate the behavior of the evaluation measures, including singularities, discontinuities and non-linearities. There are eight cases in which the GT has two object labels, each with 500 voxels. The cases are shown as confusion matrices in Fig. 4. In an ideal segmentation, the only populated cells would be along the matrix diagonal. Cases 1–5 consist of errors in which one or more voxels from the first label are misclassified as belonging to the second label as shown below. Consider Case 1: there are two MS labels, but one voxel from segment 1 is misclassified as belonging to segment 2. This error splits GT segment 1 and merges with GT segment 2. The results of applying the various evaluation measures to this case are in the column labeled Case 1 of Table 1. Similarly other columns contain the results for the other cases. In Case 9, one GT label (plus air) is split in two by the MS. We compare our measures against OCE [6] and multiclass F_1.

There are discontinuities in the OCE, but not in the other measures. The OCE jumps from zero at perfect segmentation to 0.25 in our two-label problem when a single pixel is misclassified (Case 1). This is because OCE treats it as a new segment of equal importance as the segment that is a near-perfect match for the GT label. If more voxels are moved over (Cases 2–5), the OCE monotonically increases. However, if instead, one pixel is moved from the second label to the first, as shown by Case 6, there is another jump from 0.25 to 0.5. The discontinuities are an undesirable property of OCE. We contrast Case 6 with Case 2. Intuitively, the error is smaller in Case 6 than 2, and less significant in the luggage screening application, but OCE says the opposite and gives a poorer score to Case 6. In Case 7, there is no penalty for missing an entire object, demonstrating another undesirable property of OCE. F_1^m monotonically decreases as error increases. It penalizes merging more than missing or split segments, as shown by Cases 7–9, according to the argument that we have not only missed one object but expanded another. This is a combined type I and type II error, which does not occur in two-class problems. However, we could argue that there is only one underlying error, and that we want the merged segments to be penalized no worse than the other types of error. But that is a limitation of the F_1-score definition. Note that the confusion matrix can be weighted, e.g., to assign greater penalty to type II error, although we have not done so here.

The WMI scores are intuitive. While there is a degeneracy to zero for single segments in the GT or MS as shown in Cases 7–9, we have not encountered this case in our luggage data.

Now we consider the FDR method comprising residuals and slope. The residuals (CVM, R_{L_1}, KL) report the total error. They do not distinguish between Cases 7–9. They give a perfect score of zero for perfect recovery of the feature (volume in these cases), even if the segmentation boundaries are wrong, as illustrated by Case 6. From the point of view of feature recovery, the error of zero is acceptable. The slope, shown by K in the table, tells us the kind of error, i.e., merging or splitting in Cases 8 and 9. Case 8 shows undersegmentation; the MS labels have larger magnitudes than the GT, which occurs if the algorithm merges objects more than it splits them. Conversely, Case 9 has oversegmentation; the MS labels have smaller magnitudes than their corresponding GT labels, which occurs due to splitting. A slope of one tells us that the errors are random (for non-zero residuals). In these examples, no weighting was applied. Weighting can be applied, for example, to increase the residual penalty for type II errors.

The FDR method is a framework within which we can measure not just point-wise features or features within a fixed neighborhood, but also label-wise features. Therefore, we must use residuals that allow different numbers of labels in each image. R_{L_1} gives us a result that is linear with the error, which makes it easy to understand. The CVM errors are monotonic, but nonlinear because CVM accumulates squared errors. The KL divergence is infinite in Cases 7 and 8. The absence of any one label due to missed objects or merging results in a division by zero and causes the divergence to be infinite. Therefore we can not use the KL divergence on our luggage data. We have not used another common divergence, the Kolmogorov-Smirnoff divergence (KS2) because it gives undue weight to just one object, which is not desirable in our application, where there are multiple errors and where the range of error magnitude is unpredictable.

In summary, we find that WMI and FDR methods provide acceptable and complementary results for the synthetic problems. We use R_{L_1} because it is linear and permits different numbers of labels in the GT and MS. WMI and FDR are flexible because we can use them on features rather than voxels, and can use weighting as will be demonstrated with bag data.

5. Bag data

In this section we present the results of applying our methods to the ALERT luggage images and their segmentations. We first discuss WMI results, then FDR results, and then the human expert validation. In the tables and figures, we name the segmentation algorithms A1–A5 to anonymize the research groups. The bags are named B1–B5.

5.1. WMI results

WMI scores for volume and mass are shown in Tables 2 and 3 respectively. The tables show that the best performer for volume and mass recovery is algorithm A2. Comparing the mass and volume WMI tables, we see that mass and volume give numerically different results. This happens because the GT or MS labels have a mixture of CT densities. For example, in shoes, the upper is responsible for most of the volume, while the sole is responsible for most of the mass. The segmentation algorithms recovered the sole, not the upper. In general, the mass scores are higher than the volume scores.

For comparison, the F_1^m scores for volume and mass are given in Tables 4 and 5 respectively. The F_1^m scores for volume do not yield a clear winner, but the scores for mass are similar to the WMI scores in that the mass scores show the best performer to be A2, and the mass scores are generally higher than the volume scores.

The WMI scores for uniformity are in Table 6. The best performer for the uniformity feature is unclear. Although WMI gave the highest scores to A2 by volume and mass, A2 is not the best algorithm to recover the uniformity feature.

Finally we show the cell-wise WMI weights in Table 7. Some WMI scores increase and some decrease compared to unweighted scores, but are not much different from unweighted scores. The results are discussed in more detail in Section 6. This weighting does not have a counterpart in the FDR method.

5.2. FDR results

An example FRS plot is shown in Fig. 5 for one algorithm. The FRS slopes for volume, mass and uniformity features, for all algorithms are given in Table 8, and the R_{L1}-residuals are given in Table 9 for the combined set of bags. The residual errors per bag for the different features are given in Tables 10 through 12.

Mass and volume features increase monotonically with the number of voxels in a segment, so a slope $K > 1$ indicates systematic undersegmentation and $K < 1$ indicates systematic oversegmentation, including missing parts of segments. For a non-monotonic feature such as uniformity, FRS slope values do not indicate splitting or merging of the object, but rather a systematic over- or under-estimation of the feature. Over or under-segmentation should not be simplistically defined by counting the number of segments. For example, if multiple machine segments exist for a single GT label, there is oversegmentation. However, if most of the feature is recovered in one machine segment, there is less oversegmentation than if the feature is distributed equally among the multiple machine segments.

Among the algorithms, A2 exhibits best mass and volume recovery. Its FRS slopes are closest to one (Table 8), and the residuals are smallest (Table 9). For all algorithms, the mass slopes are closer to one than the volume slopes, and the mass residuals are smaller than the volume residuals. As in the WMI scores, the FRS plots show that it is easier for a segmentation algorithm to recover mass than volume because of the heterogeneity of the material composition of objects and clutter.

Although A2 has the best volume and mass retrieval, it does not show best recovery of uniformity. As shown in Table 8, the uniformity slope for A2 is small (0.51) compared to other algorithms. There is also no clear best performer. The FDR results are in line with the WMI scores.

An instance of poor uniformity recovery was a water bottle touching another liquid-filled container. The MS label for the water bottle included the other liquid, and lost some of the bottle itself, either labeling it as air or as the other liquid as shown in Fig. 6. Also included into the bottle label were voxels metal from a nearby touching object (not shown). The volume and mass were well-recovered because of the exchange of material between the two labels, but the CT number differences of the mixed materials created a high variance in the machine segment. Feature recovery over the feature range is shown in Fig. 7. The sliding average scatter plots show that feature recovery improves as object mass increases for A2. Although A2 has the best mass scores, it is less reliable for low-mass objects than some of the other algorithms. At higher masses, it is more reliable than the other algorithms. In another example, A5 shows that no apparent preference for any range of mass. An example plot of outliers is shown in Fig. 8 for the mass feature, for the A1 algorithm.

5.3. Validation by human expert observer

In order to validate the methods beyond the synthetic problems, a human observer evaluated two MS algorithms, A1 and A2, against GT. The difficulty of the task for the observer arises from the multiple splits and merges and the large number of slices. The comparison was simplified by sampling every fifteenth slice. The observer was presented with corresponding slices of A1, A2, GT and CT images. The MS slices were randomly ordered for blind review. For each pair of slices, the observer selects the MS that he considers a closer match to the GT slice. The results are in Table 13.

In each bag, the observer preferred A2 over A1, which is in agreement with WMI and FDR results. The expert explained some of his decisions. He observed that slices from A1 had more type II error than those from A2. This observation relates to the lower WMI scores (Tables 2 and 3) and smaller slope of A1 (Table 8) compared with A2. The expert selected A1 in some slices where A2 labels appeared jagged. The jagged labels belonged to large liquid-filled containers. These selections agree with the uniformity scores.

There is also a correspondence between expert's preferred percentage and the WMI and R_{L1} results per bag. We do not expect to see perfect correspondence because the expert performed a simplifed evaluation. The slice sampling method favors larger less dense objects over smaller denser ones, the human is imprecise and is influenced by visual appeal, there was no weighting per slice to increase the impact of fuller slices over emptier ones, and no quantification of preference for a given pair of slices.

In addition, we applied WMI slicewise on the same slices evaluated by the human. For all slices, the higher-scoring algorithm was compared with the human preference using McNemar [41] and KS2 tests. The McNemar test yielded a p-value of 0.08 which does not reject the null hypothesis that the human and WMI prefer the same algorithm. The KS2 test-statistic was 0.04, which also does not reject the null hypothesis at a confidence level of 0.05.

In summary, the WMI and FDR results on bag data indicate that these are appropriate evaluation measures. The two methods complement each other. The results showed that the two methods also picked out the same best performer. They can incorporate pointwise or regional weighting, and can operate on features that are more relevant than segment volume.

5.4. Summary

The FDR and WMI both measure feature recovery, unlike existing evaluation methods that compute edge distances or voxel misclassification. Both methods are sensitive to spatial correspondence of labels, unlike histogram comparison methods that measure features. Further, both methods are useful for multiple label segmentation problems. And both allow us to assign priorities to segments. They also gave consistent results in selecting the same best algorithm. However, FDR and WMI have different perspectives. WMI is more sensitive to spatial correspondence than FDR. FDR is more flexible in that data from multiple images can be pooled and trends can be extracted, and a wider variety of features can be used. A human expert validated our methods by visual assessment.

6. Discussion

As discussed in Section 1, many GT-based methods in the evaluation literature use region-based errors when multiple regions of interest are present in the image. This can be thought of as using an indicator function on each voxel for each label. But each voxel and its neighborhood contain additional information we can use instead of just the indicator. In our case, we have used mass and uniformity in addition to volume. In the mass scores, voxels with higher CT number are more important than those with lower CT number. The use of uniformity prioritizes more homogenous objects over less homogenous ones. Mass and uniformity are examples of features that may be useful for a specific application. An EDS may utilize these or other features depending on the ATR algorithm.

The WMI, F_1^m score and FDR results for mass are more consistent with each other (same best performer) than the corresponding volume scores. These discrepancies between volume and mass illustrate the challenges of segmentation of CT images of luggage. The results show that mass is easier to recover than volume, i.e., a meaningful feature within a region is easier to extract than the region itself.

The FDR method is more general and informative than histogram-based methods. A previously published evaluation method for populations of similar objects used histograms [12]. However, in general, the objects in a segmentation problem are not similar, and there are no object-type populations. We generate a bipartite matching and can evaluate any objects. Due the bipartite matching, we can extract information about systematic errors, expected performance as a function of feature value, and outliers. Matching allows object prioritization and non-uniform costs. The residuals include pairwise errors, missed and spurious segments (although we do not penalize the latter here).

The FRS and WMI showed that A2 traded-off region uniformity for better overall segmentation. Note that if the CT number distribution of adjoining objects is the same, the mass or volume FRS plots may not indicate errors (provided the same volumes are displaced from one object to another). If the textures are similar as well, the FRS plot for uniformity will not indicate errors either. This is acceptable from the feature recovery point of view. Another inference we can draw from the uniformity results is that the improvement of the other algorithms relative to A2 shows that they find it easier to segment uniform objects, while A2 is less dependent on object uniformity. We have confirmed this inference by assigning constant weights to the rows of the confusion matrix that represent uniform objects. The WMI only improves slightly for A2, but considerably more for the other algorithms. For brevity, we did not show these results.

Our sliding-average plots show trends in performance as a function of feature value for some algorithms. We show the mass feature, because that has the best WMI and FDR scores. The accuracy of segmentation of an object depends not only on its own features, but those of the surrounding objects. As a result, algorithms may not all show trends with feature value, but if trends are present, they help in the interpretation of segmentation results.

In the cell-wise weighted confusion matrix, we have weighted each cell by a factor representing the similarity of a regional feature. Our factor is the ratio of the smaller mean to the greater mean. For a cell representing some GT and MS labels, if the labels are dissimilar in the regional feature, we assign a smaller weight to the cell, which is to say that this cell does not help us get information about one distribution from the other distribution. This decreases the total WMI. Consider a cell on the diagonal of the confusion matrix. The diagonal represents the matched objects. If the matched objects are dissimilar, then the ratio is small, and the cell loses importance. Here it is easy to see the interpretation that the object represented by the cell in the MS image does not tell us much about the GT image. Considering an off-diagonal cell, it similarly loses importance when the means are dissimilar. At first glance, it seems counter-intuitive that a cell that represents two unmatched objects, should have decreased weight when the objects are dissimilar. But WMI does not measure the ordering of the information. This cell contains the quantity of an intersection that really does exist. So if we decrease (increase) the weighting of that intersection, we decrease (increase) the amount of information one label set tells us about the other label set. In addition, we increase or decrease the entropies of the GT and MS images when we weight cells, depending on what the original image contained. The cell-wise weighting therefore is difficult to control and does not give monotonic results.

Our next goal is to image a larger number of bags with threat simulants and compare our evaluation scores with the probability of detection and probability of false alarm from simulated EDS certification tests.

7. Conclusion

We have developed two flexible parameter-independent methods to evaluate segmentation algorithms. The methods were applied on a test set of luggage images. Our contributions are as follows.

1. We have used a well-accepted measure from information theory to measure feature overlap.

2. **2.** We have developed a new method based on feature recovery that has good agreement with mutual information, but that also identifies systematic errors and allows pointwise or regional features to be used.

3. **3.** We have used weighting functions to prioritize objects based on desired features.

4. **4.** We developed a semi-automatic method to extract GT from three-dimensional CT images.

We used human evaluation of segmentation accuracy and synthetic problems to validate our methods. Our evaluation methods indicated one algorithm, A2, as the best one, and found characteristics of the algorithm: accuracy increased with object mass, and that A2 was less reliant on object uniformity than some of the other algorithms. Given the challenges and requirements for segmentation in luggage scanning, we found our methods to be more suitable to evaluate segmentation algorithms than methods from existing literature.

Acknowledgments

The authors thank Dr. J. Beaty and the ALERT group at Northeastern University for supplying the image sets. The authors are very grateful to Jihoon Kim for help with the statistical testing, and to Drs. A. Boxwala, J. Kallman, S. Azevedo, K. Bond, and C. Crawford. This work was performed under the auspices of Lawrence Livermore National Laboratories, the Science and Technology Directorate of the Department of Homeland Security, and Northeastern University.

References

1. Zhang YJ. A survey on evaluation methods for image segmentation. Pattern Recognit. 1996; 29(8): 1335–1346.
2. Zhang H, Fritts JE, Goldman SA. Image segmentation evaluation: A survey of unsupervised methods. Comput Vis Image Underst. May; 2008 110(2):260–280.
3. Fenster A, Chiu B. Evaluation of segmentation algorithms for medical imaging. Conf Proc IEEE Eng Med Biol Soc. Jan.2005 7:7186–7189. [PubMed: 17281935]
4. Heimann T, van Ginneken B, Styner Ma, Arzhaeva Y, Aurich V, Bauer C, Beck A, Becker C, Beichel R, Bekes G, Bello F, Binnig G, Bischof H, Bornik A, Cashman PMM, Chi Y, Cordova A, Dawant BM, Fidrich M, Furst JD, Furukawa D, Grenacher L, Hornegger J, Kainmüller D, Kitney RI, Kobatake H, Lamecker H, Lange T, Lee J, Lennon B, Li R, Li S, Meinzer H-P, Nemeth G, Raicu DS, Rau A-M, van Rikxoort EM, Rousson M, Rusko L, Saddi Ka, Schmidt G, Seghers D, Shimizu A, Slagmolen P, Sorantin E, Soza G, Susomboon R, Waite JM, Wimmer A, Wolf I. Comparison and evaluation of methods for liver segmentation from CT datasets. IEEE Trans Med Imaging. Aug; 2009 28(8):1251–1265. [PubMed: 19211338]
5. Martin D, Fowlkes C, Tal D, Malik J. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. Proc Eighth IEEE Intl Conf Comp Vis. 2001; 2:416–423.
6. Polak M, Zhang H, Pi M. An evaluation metric for image segmentation of multiple objects. Image Vis Comput. Jul; 2009 27(8):1223–1227.
7. Hoover A, Jean-Baptiste G, Jiang X, Flynn PJ, Bunke H, Goldgof DB, Bowyer K, Eggert DW, Fitzgibbon A, Fisher RB. An experimental comparison of range image segmentation algorithms. IEEE Trans Pattern Anal Mach Intell. Jul; 1996 18(7):673–689.
8. Jiang X, Marti C, Irniger C, Bunke H. Distance measures for image segmentation evaluation. EURASIP J Adv Signal Process. 2006; 2006(1):1–11.
9. Cardoso JS, Corte-Real L. Toward a generic evaluation of image segmentation. IEEE Trans Image Process. Nov; 2005 14(11):1773–1782. [PubMed: 16279178]
10. Zhang Y, Gerbrands J. Segmentation evaluation using ultimate measurement accuracy. Proc SPIE Image Proc Algo Tech III. May.1992 1657:449–460.

11. Pal NR, Bhandari D. Image thresholding: Some new techniques. Signal Processing. Aug; 1993 33(2):139–158.
12. Hagwood C, Bernal J. Evaluation of segmentation algorithms on cell populations using CDF curves. IEEE Trans Med Imaging. Feb; 2011 31(2):380–390. [PubMed: 21965194]
13. Puzicha J, Buhmann JM, Rubner Y, Tomasi C. Empirical evaluation of dissimilarity measures for color and texture. Proc Seventh IEEE Intl Conf Comp Vis. 1999; 2:1165–1172.
14. Rubner Y, Puzicha J, Tomasi C, Buhmann JM. Empirical evaluation of dissimilarity measures for color and texture. Comput Vis Image Underst. Oct; 2001 84(1):25–43.
15. Bai X, Zhao Y, Huang Y, Luo S. Normalized joint mutual information measure for image segmentation evaluation with multiple ground-truth images. Comp Anal Images Patterns. Aug. 2011 :110–117.
16. Cavallaro A, Gelasca ED, Ebrahimi T. Objective evaluation of segmentation quality using spatio-temporal context. IEEE Proc Intl Conf Image Proc. 2002; 3:III-301–III-304.
17. Benhabiles H, Vandeborre J-P, Lavoue G, Daoudi M. A framework for the objective evaluation of segmentation algorithms using a ground-truth of human segmented 3D-models. IEEE Intl Conf Shape Modeling and Appl. Jun.2009 :36–43.
18. Cárdenes R, de Luis-García R, Bach-Cuadra M. A multidimensional segmentation evaluation for medical image data. Comput Methods Programs Biomed. Nov; 2009 96(2):108–124. [PubMed: 19446358]
19. Johnson B, Xie Z. Unsupervised image segmentation evaluation and refinement using a multi-scale approach. ISPRS J Photogramm Remote Sens. Jul; 2011 66(4):473–483.
20. McGuinness K, O’Connor NE. Toward automated evaluation of interactive segmentation. Comput Vis Image Underst. Jun; 2011 115(6):868–884.
21. Monteiro F, Campilho A. Distance measures for image segmentation evaluation. AIP ICNAAM 2012: Intl Conf Num Anal and Appl Math. 2012; 1479(1):794–797.
22. Mezaris V, Kompatsiaris I, Strintzis M. Still image objective segmentation evaluation using ground truth. 5th COST 276 Workshop. 2003:9–14.
23. Rajab M. Feature extraction of dermatoscopic images by iterative segmentation algorithm. J Xray Sci Technol. Mar; 2008 16(1):33–42.
24. Chang Y, Xia J, Yuan P, Kuo T. 3D segmentation of maxilla in cone-beam computed tomography imaging using base invariant wavelet active shape model on customized two-manifold topology. J Xray Sci Technol. May; 2013 21(2):251–282. [PubMed: 23694914]
25. Silevitch, M.Crawford, C., Martz, H., editors. Algorithm Development for Security Applications. 2009. Inal report on algorithm development for security applications.
26. Silevitch, M.Crawford, C., Martz, H., editors. Algorithm Development for Security Applications. 2011. Final report on algorithm development for security applications 6.
27. Haris K, Efstratiadis SN, Maglaveras N, Katsaggelos AK. Hybrid image segmentation using watersheds and fast region merging. IEEE Trans Image Process. Jan; 1998 7(12):1684–1699. [PubMed: 18276235]
28. Rand W. Objective criteria for the evaluation of clustering methods. J Am Stat Assoc. 1971; 66(336):846–850.
29. Olson, D., Delen, D. Advanced Data Mining Techniques. 1. Springer; 2008. p. 138
30. Grady L, Singh V, Kohlberger T, Alvino C, Bahlmann C. Automatic segmentation of unknown objects, with application to baggage security. Proc Eur Conf Comp Vision. 2012:430–444.
31. Wiley D, Ghosh D, Woodhouse C. Automatic segmentation of CT scans of checked baggage. Proc 2nd Intl Meeting Image Formation X-ray CT. 2012:310–313.
32. Southam P, Harvey R. Texture classification via morphological scale-space: Tex-mex features. J Electron Imaging. Oct.2009 18(4):043007.
33. Unnikrishnan R, Pantofaru C, Hebert M. Toward objective evaluation of image segmentation algorithms. IEEE Trans Pattern Anal Mach Intell. Jun; 2007 29(6):929–944. [PubMed: 17431294]
34. Warfield SK, Zou KH, Wells WM. Simultaneous truth and performance level estimation (STAPLE): An algorithm for the validation of image segmentation. IEEE Trans Med Imaging. Jul; 2004 23(7):903–921. [PubMed: 15250643]

35. Mevis Medical Solutions. Bremen, Germany: 2011. MeVisLab medical image processing and visualization.

36. Papoulis, A. Probability, Random Variables and Stochastic Processes. 3. McGraw-Hill; 1991.

37. Kuhn HW. The Hungarian method for the assignment problem. Nav Res Logist Q. Mar; 1955 2(1–2):83–97.

38. Huber, PJ. Robust Statistics. Hoboken, NJ, USA: John Wiley & Sons, Inc; 1981.

39. Anderson TW. On the distribution of the two-sample cramer-von mises criterion. Ann Math Stat. Sep; 1962 33(3):1148–1159.

40. Kullback S, Leibler R. On information and sufficiency. Ann Math Stat. Mar; 1951 22(1):79–86.

41. Bland, M. An Introduction to Medical Statistics. 3. Oxford University Press; 2000. p. 250

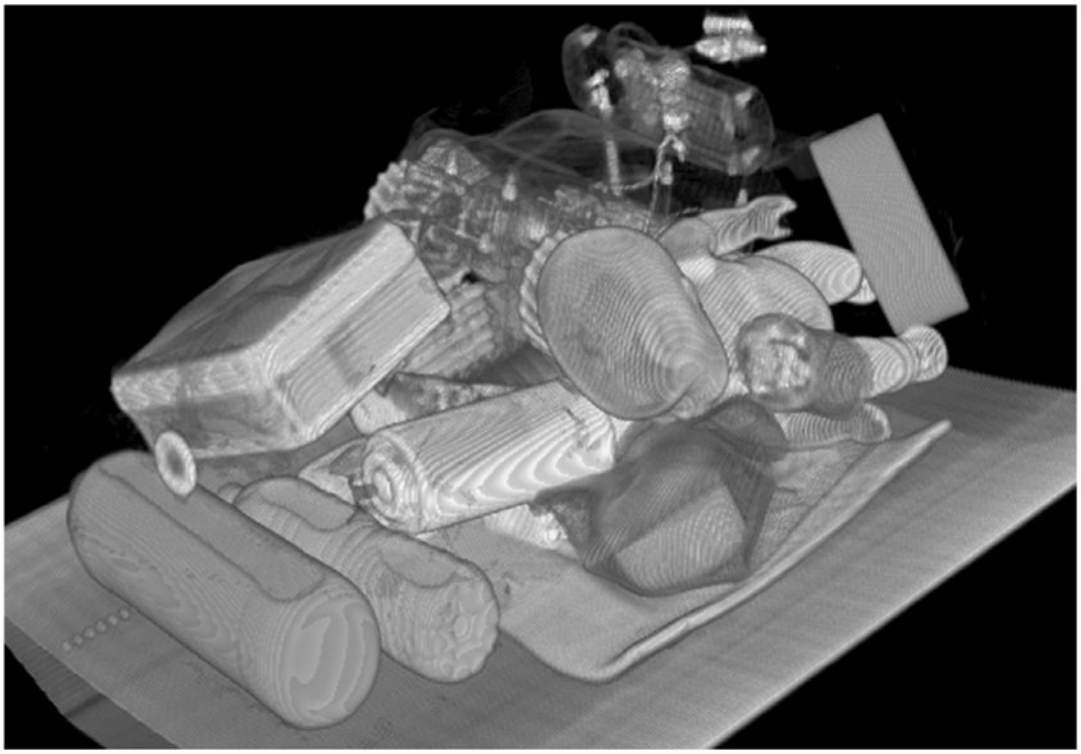


Fig. 1.

A volume rendering of the volumetric CT image of one suitcase. The suitcase is on the patient pallet of a medical CT scanner.

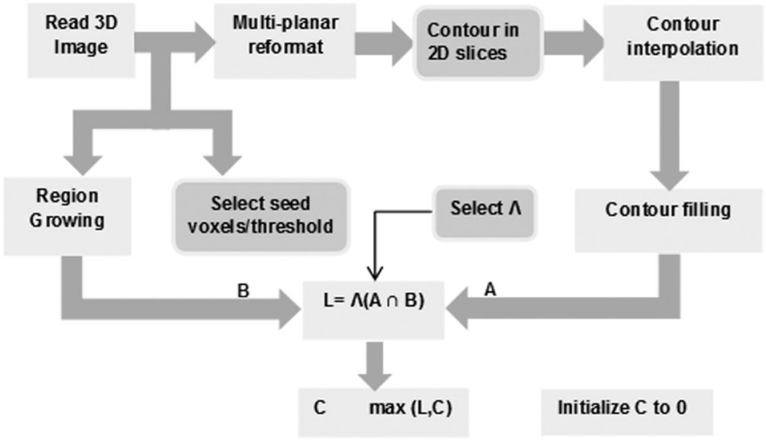


Fig. 2.

Flowchart showing the operations performed to determine GT labels from the volumetric CT image. Manual operations are shown in darker boxes. A and B are segments generated by the two different paths, C is the accumulated set of labels, and Λ is a numeric value assigned to a label.

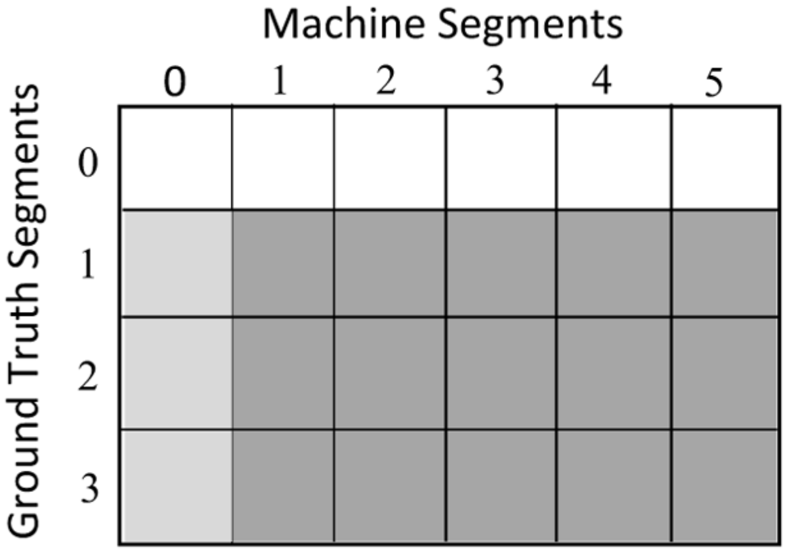
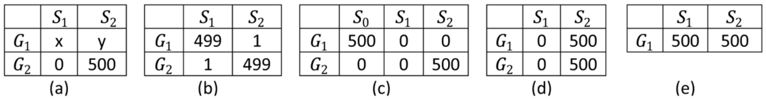


Fig. 3.

Confusion matrix showing inner matrix used in the calculation of entropies, and showing the outer matrix used in *r* in Eq. (6).



	S_1	S_2
G_1	x	y
G_2	0	500

(a)

	S_1	S_2
G_1	499	1
G_2	1	499

(b)

	S_0	S_1	S_2
G_1	500	0	0
G_2	0	0	500

(c)

	S_1	S_2
G_1	0	500
G_2	0	500

(d)

	S_1	S_2
G_1	500	500

(e)

Fig. 4.

Confusion matrices for the synthetic problems. Cases 1–5 are shown in (a). Case 1: $x = 499$, $y = 1$, Case 2: $x = 475$, $y = 25$, Case 3: $x = 450$, $y = 50$, Case 4: $x = 400$, $y = 100$, Case 5: $x = 250$, $y = 250$. Case 6: one pixel from each GT label is misclassified by MS as belonging to the other label (b), Case 7: One GT label is not detected (c), Case 8: Both GT labels are merged by MS (d), Case 9: Single GT label is split by MS (e).

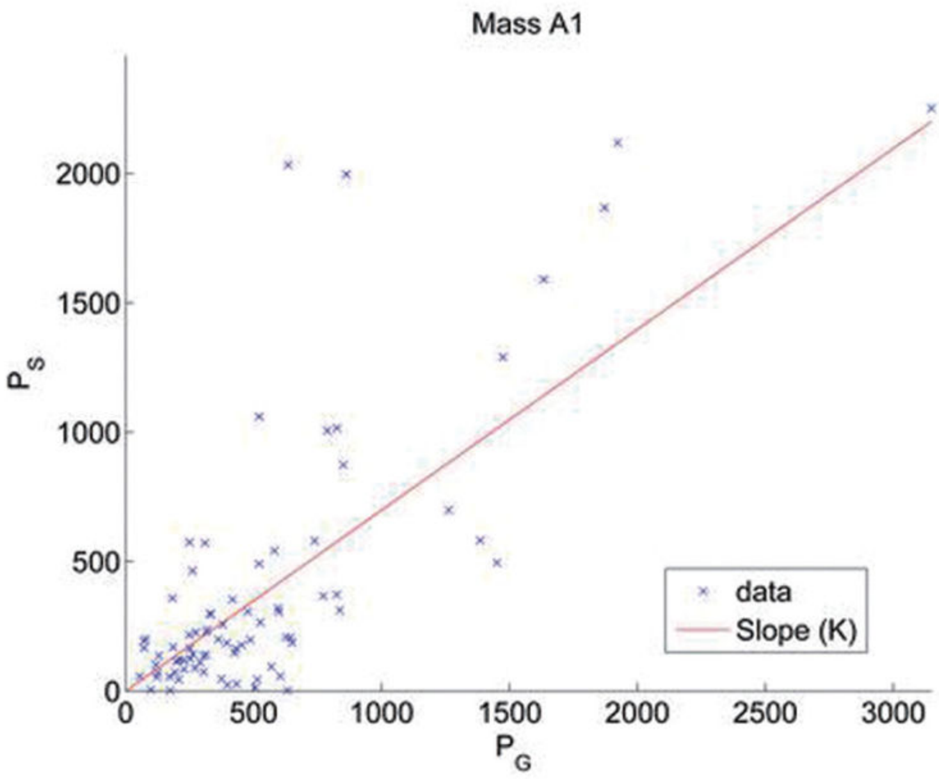


Fig. 5.

The mass scatter plot from algorithm A1. There were 81 GT labels. The fitted line is forced to pass through zero. (Colours are visible in the online version of the article; http://dx.doi.org/10.3233/XST-140418)

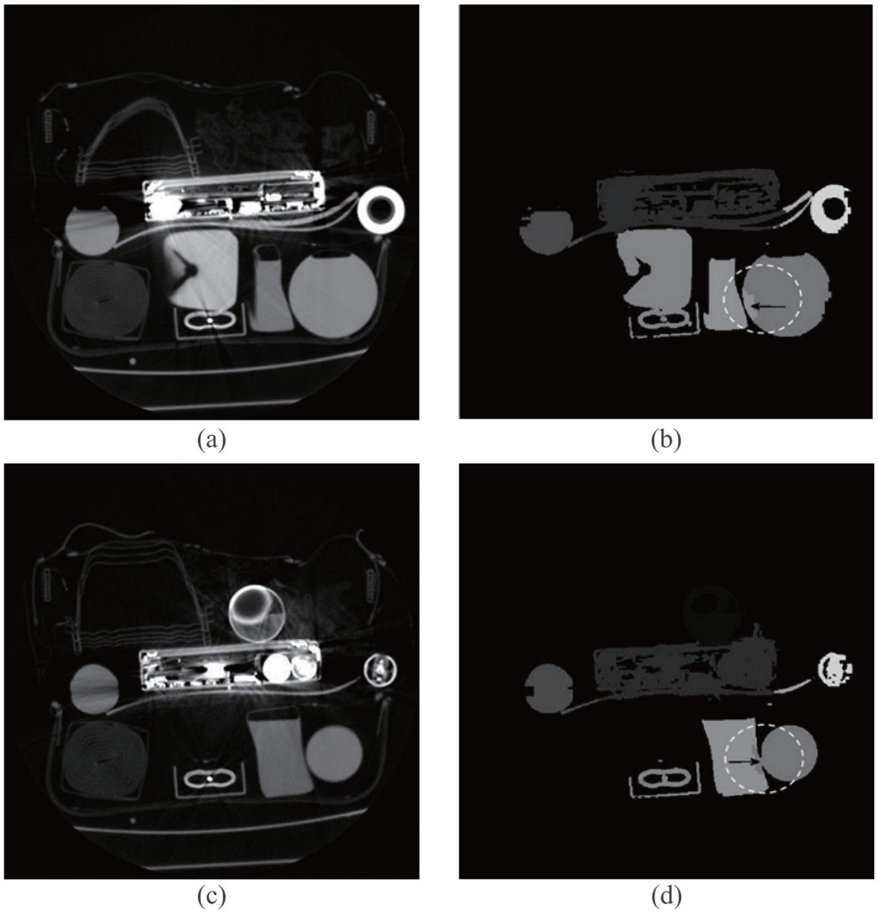


Fig. 6.

Poor uniformity recovery by A2 of a large uniform object. Two CT slices are shown in the left column and label images on the right. Objects circled in the right column are liquid-filled containers. There is misclassification between those two object labels, shown by arrows, as well as one of the objects and air.

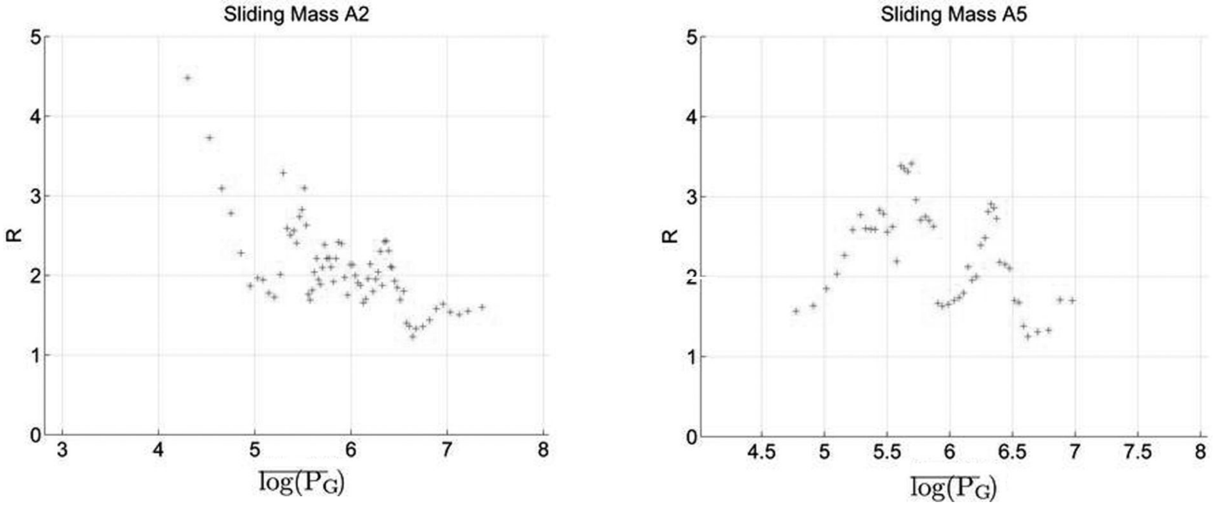


Fig. 7.

The sliding average (Eq. (16)) for the mass feature shown for two algorithms show different characteristics. A2 improves with mass, but A5 does not.

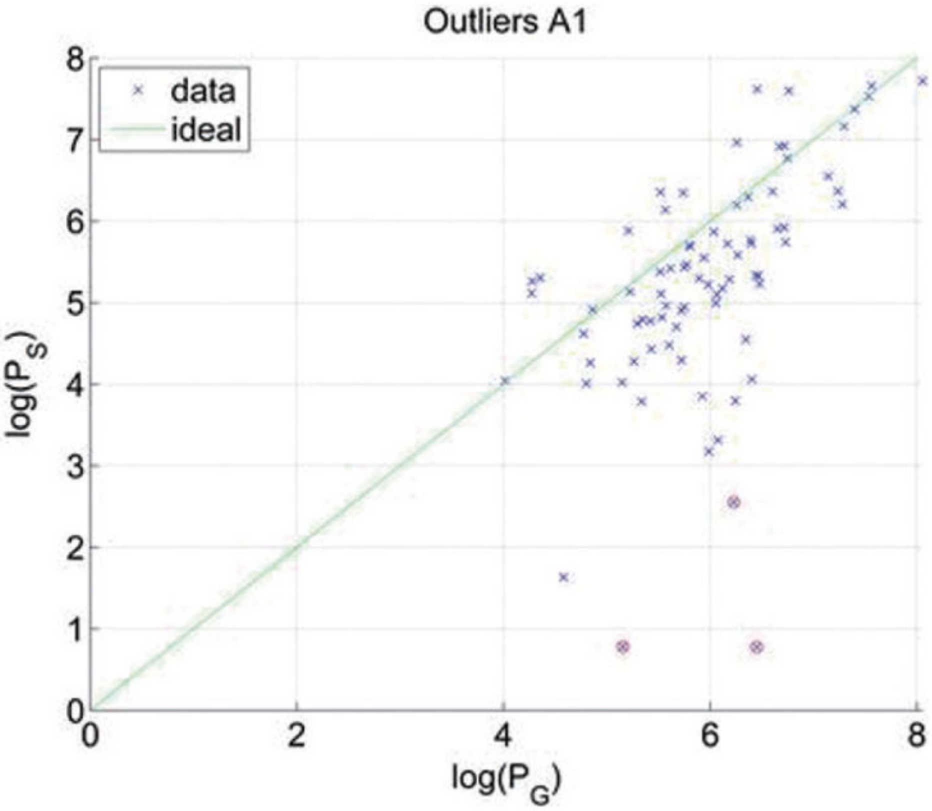


Fig. 8.

Example FRS plot for mass showing encircled outliers (Eq. (18)). (Colours are visible in the online version of the article; http://dx.doi.org/10.3233/XST-140418)

Table 1

Performance values for various test cases considering two GT object labels (and air). The slope of the line fitted to the FRS data is denoted K. CVM, R_{L_1} and KL are the different residual errors. Performance values for perfect MS (no error) are given in the second column as a reference

	Ideal	Case 1	Case 2	Case 3	Case 4	Case 5	Case 6	Case 7	Case 8	Case 9
OCE	0	0.25	0.29	0.33	0.39	0.51	0.5	0	0.5	0.5
F_1^m	1	0.999	0.975	0.95	0.9	0.75	0.998	0.67	0.5	0.67
WMI	1	0.99	0.86	0.76	0.62	0.35	0.98	0	0	0
CVM	0	5×10^{-4}	0.0125	0.025	0.05	0.125	0	0.25	0.25	0.5
R_{L_1}	0	10^{-3}	0.025	0.05	0.1	0.25	0	0.5	0.5	0.5
KL	0	≈0	0.0013	0.005	0.02	0.144	0	∞	∞	0.693
K	1	1	1	1	1	1	1	1	2	0.5

Table 2

WMI scores for volume. The best performance in most bags is from A2

	A1	A2	A3	A4	A5
B1	0.22	0.63	0.54	0.48	0.50
B2	0.45	0.62	0.58	0.48	0.41
B3	0.59	0.69	0.65	0.56	0.38
B4	0.33	0.59	0.65	0.53	0.50
B5	0.60	0.78	0.74	0.68	

Table 3

WMI scores for mass. The best performance in most bags is from A2

	A1	A2	A3	A4	A5
B1	0.27	0.76	0.65	0.58	0.64
B2	0.57	0.74	0.71	0.56	0.58
B3	0.66	0.74	0.69	0.50	0.49
B4	0.40	0.69	0.74	0.63	0.64
B5	0.66	0.84	0.77	0.63	

Table 4

F_1^m scores by volume. It is not clear which is the best performing algorithm

	A1	A2	A3	A4	A5
B1	0.32	0.67	0.56	0.60	0.55
B2	0.53	0.60	0.57	0.61	0.44
B3	0.49	0.62	0.59	0.57	0.47
B4	0.42	0.52	0.65	0.65	0.59
B5	0.67	0.78	0.76	0.79	

Table 5

F_1^m scores by mass. The best performance in most bags is by A2

	A1	A2	A3	A4	A5
B1	0.37	0.73	0.61	0.68	0.60
B2	0.60	0.68	0.62	0.66	0.56
B3	0.54	0.65	0.61	0.54	0.57
B4	0.46	0.61	0.70	0.70	0.70
B5	0.73	0.83	0.73	0.73	

Table 6

WMI score for uniformity. It is not clear which algorithm performs best for this feature

	A1	A2	A3	A4	A5
B1	0.28	0.77	0.69	0.68	0.66
B2	0.66	0.76	0.75	0.68	0.54
B3	0.68	0.67	0.70	0.67	0.50
B4	0.43	0.71	0.78	0.73	0.64
B5	0.78	0.83	0.87	0.90	

Table 7

WMI for cell-wise weighting of volume

	A1	A2	A3	A4	A5
B1	0.20	0.61	0.51	0.46	0.47
B2	0.43	0.59	0.57	0.46	0.37
B3	0.60	0.69	0.65	0.55	0.36
B4	0.29	0.57	0.64	0.51	0.48
B5	0.59	0.78	0.75	0.67	

Table 8

Slopes (K) for FRS fit lines for volume, mass and uniformity features

	A1	A2	A3	A4	A5
Volume	0.59	0.85	0.56	0.73	0.61
Mass	0.70	1.0	0.58	0.67	0.89
Uniformity	1.26	0.51	0.91	1.06	1.5

Table 9

R_{L_1} residuals for all bags combined. The smallest residuals are from A2

	A1	A2	A3	A4	A5
Volume	0.49	0.37	0.48	0.44	0.54
Mass	0.41	0.28	0.45	0.44	0.41
Uniformity	0.60	0.33	0.54	0.51	0.62

Table 10

R_{L_1} residual error by volume. In most bags, the smallest residual is from A2

	A1	A2	A3	A4	A5
B1	0.76	0.46	0.61	0.56	0.51
B2	0.45	0.51	0.58	0.51	0.59
B3	0.37	0.27	0.43	0.48	0.60
B4	0.66	0.44	0.50	0.49	0.45
B5	0.37	0.22	0.36	0.30	

Table 11

R_{L_1} residual error by mass. In most bags, the smallest residual is from A2

	A1	A2	A3	A4	A5
B1	0.71	0.35	0.55	0.48	0.41
B2	0.38	0.41	0.53	0.46	0.44
B3	0.32	0.23	0.39	0.53	0.48
B4	0.63	0.35	0.43	0.43	0.31
B5	0.31	0.15	0.40	0.38	

Table 12

R_{L_1} residual error by uniformity. The smallest residuals are from A2, despite the small slope shown in Table 6

	A1	A2	A3	A4	A5
B1	1.18	0.39	0.71	0.55	0.40
B2	1.00	0.35	1.01	0.97	0.90
B3	0.92	0.24	0.64	0.51	0.42
B4	0.41	0.33	1.03	0.78	0.62
B5	0.42	0.35	0.30	0.35	

Table 13

The human observer evaluation of two MS algorithms. The middle column shows the preferred algorithm and the percentage of slices in which it was preferred. The third column shows the number of slices that were ranked better in A1, in A2 and equal in both

Bag	% A2 by expert	A1/A2/equal	% A2 by MI
B1	96	0/27/1	100
B2	82	5/32/2	85
B3	73	6/22/2	91
B4	76	6/25/2	86
B5	72	8/26/2	92