Evaluation of Volumetric Segmentation for Aviation Security

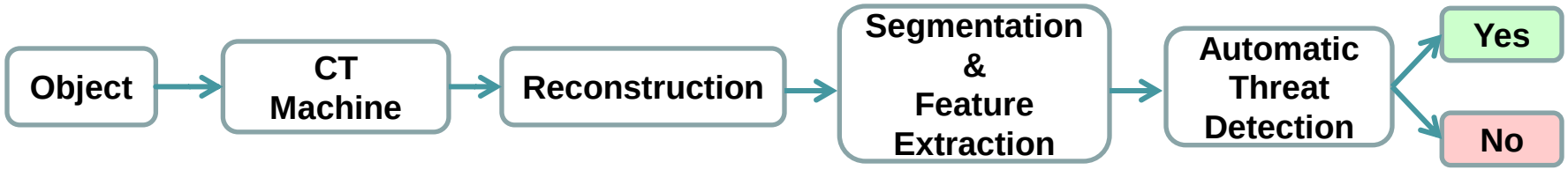
Karina Bond, Jeff Kallman, Steve Azevedo, Harry E. Martz, Jr.
Lawrence Livermore National Laboratory
***LLNL-PRES*-557759**
(IM#618755)

April 22, 2012
Version 4

This work performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.

Summary

- Segmentation of the reconstructed CT images is a key in explosive detection for airport security.
- We have been studying how to measure segmentation performance. It turns out this is not a trivial task.
- We surveyed the published literature on segmentation evaluation metrics and have developed a few ideas of our own.
- We describe one of the segmentation evaluation metrics we developed.
- We present the results of applying this metric to the Segmentation Initiative, organized by Awareness & Location of Explosives-Related Threats(ALERT) .

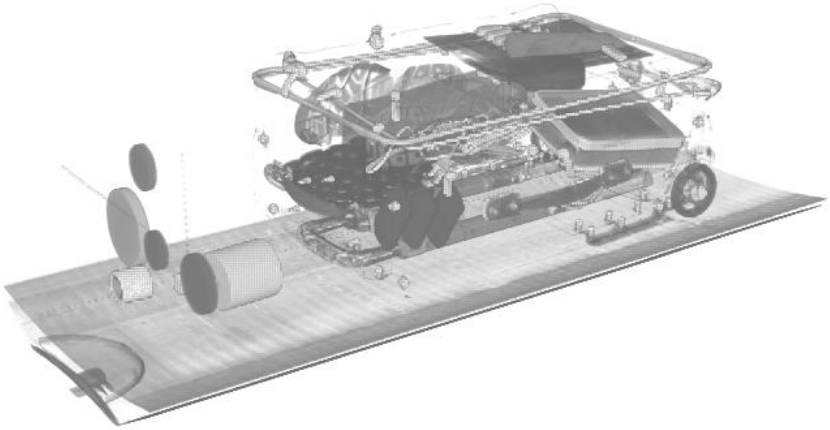


The ALERT Sementation Initiative

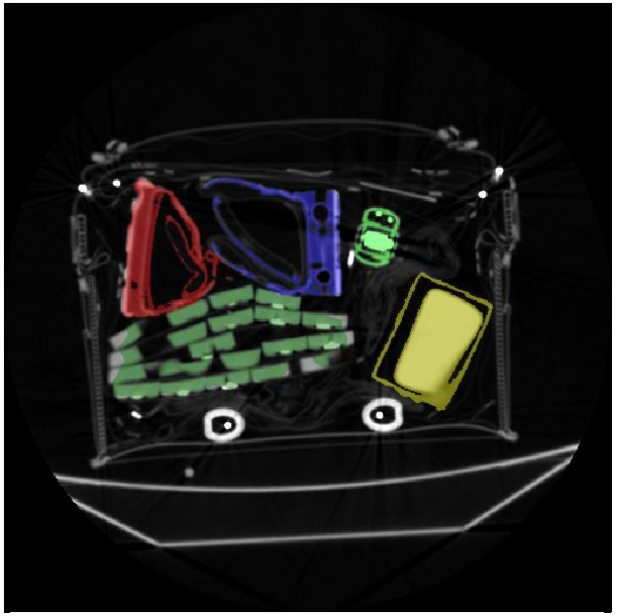
"ALERT, with contract funding from DHS, started a segmentation initiative in which five research groups were asked to adapt or develop algorithms to segment objects contained in scans of luggage on a medical CT scanner."

-Segmentation of Object from Volumetric CT data Final Report, ALERT

Five research groups were selected and subsequently funded by ALERT to develop or refine existing advanced segmentation algorithms using datasets supplied to them by ALERT. The datasets consisted of **scans on a medical CT scanner of luggage**, in addition to **ground truth for the training and evaluations portions of the dataset**.



Example of Training Dataset (3D rendered)



Example of Training Dataset (slice view) Ground truth Objects are overlaid in different colors.

Evaluation of Segmentation Algorithms

Segmentation Evaluation Methods

Subjective Methods
Qualitative evaluation of segmentation results by a human evaluator.

Objective Methods
Quantitative evaluation of segmentation algorithms.

System-Level Methods
Methods that evaluate segmentation on the basis of the larger system's parameters. In the case of CT based images these parameters might be the following.

- μ(i), Linear Attenuation Coefficient of i-th object.
- V(i), volume of i-th object.

Direct Methods
Methods that evaluate segmentation independent of the larger system they are used in.

Analytical Methods
Theoretical evaluation methods that can be calculated without any results solely based on algorithm details.

Empirical Methods
Evaluation Methods that are calculated on the basis of the results of the segmentation algorithm.

Unsupervised Methods
Evaluations methods that are based only on a set of segmentation results (no ground truth).

Supervised Methods
Evaluation methods that are based on the result of the segmentation algorithm and a ground truth image.

- P1\P2 Metric
- Martin Error (GCE\LCE)
- Object Consistency Error (OCE)
- F-Measure

Precision\Recall Definitions

Assume $I_g = \{T_1, T_2, \ldots T_M\}$ is the ground truth image, where T_i is the i-th object in I_g.

Assume $I_s = \{S_1, S_2, \ldots S_N\}$ is the segmented image, where S_j is the j-th segment in I_s.

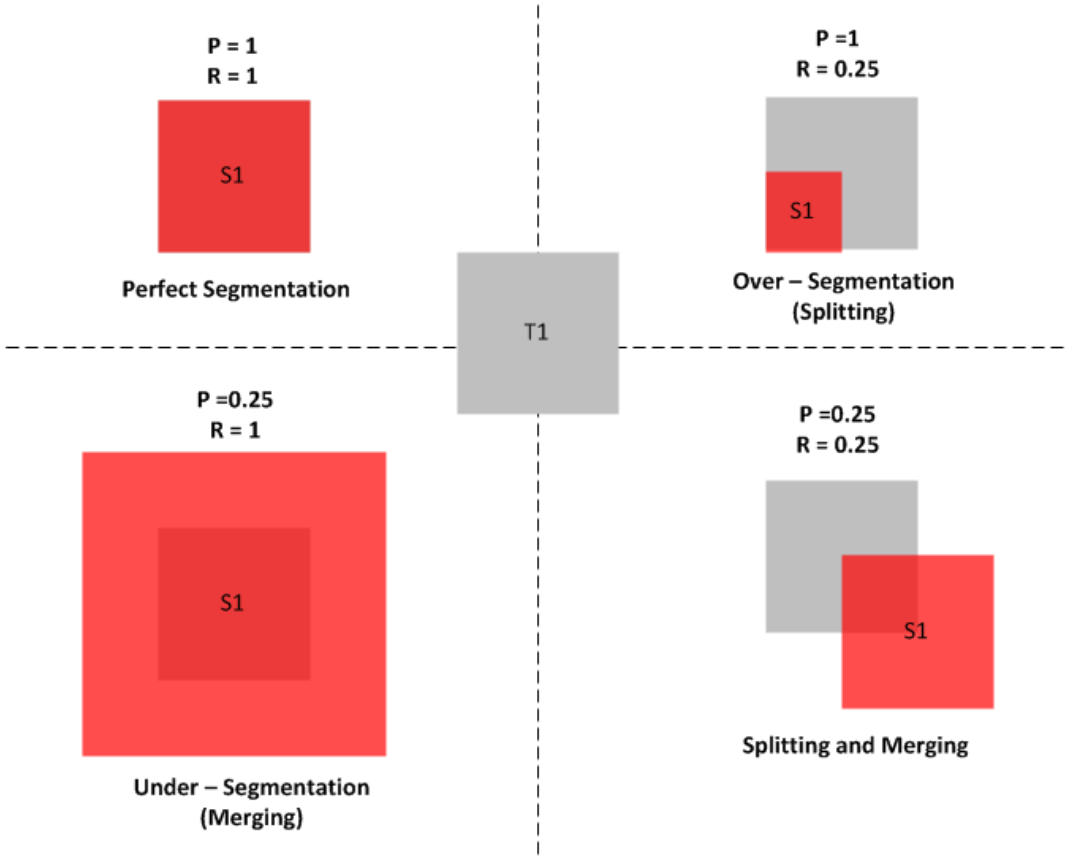
Precision, P_{ij} and Recall, R_{ij} for the ij -th fragment, G_{ij} can be calculated as follows.

For $1 \le i \le M$ and $1 \le j \le N$

$$G_{ij} = T_i \cap S_j$$

$$R_{ij} = \frac{|G_{ij}|}{|T_i|} = \frac{|T_i \cap S_j|}{|T_i|}$$

$$P_{ij} = \frac{|G_{ij}|}{|S_j|} = \frac{|T_i \cap S_j|}{|S_j|}$$



Precision\Recall Definitions

Assume $I_g = \{T_1, T_2, \ldots T_M\}$ is the ground truth image, where T_i is the i-th object in I_g.

Assume $I_s = \{S_1, S_2, \ldots S_N\}$ is the segmented image, where S_j is the j-th segment in I_s.

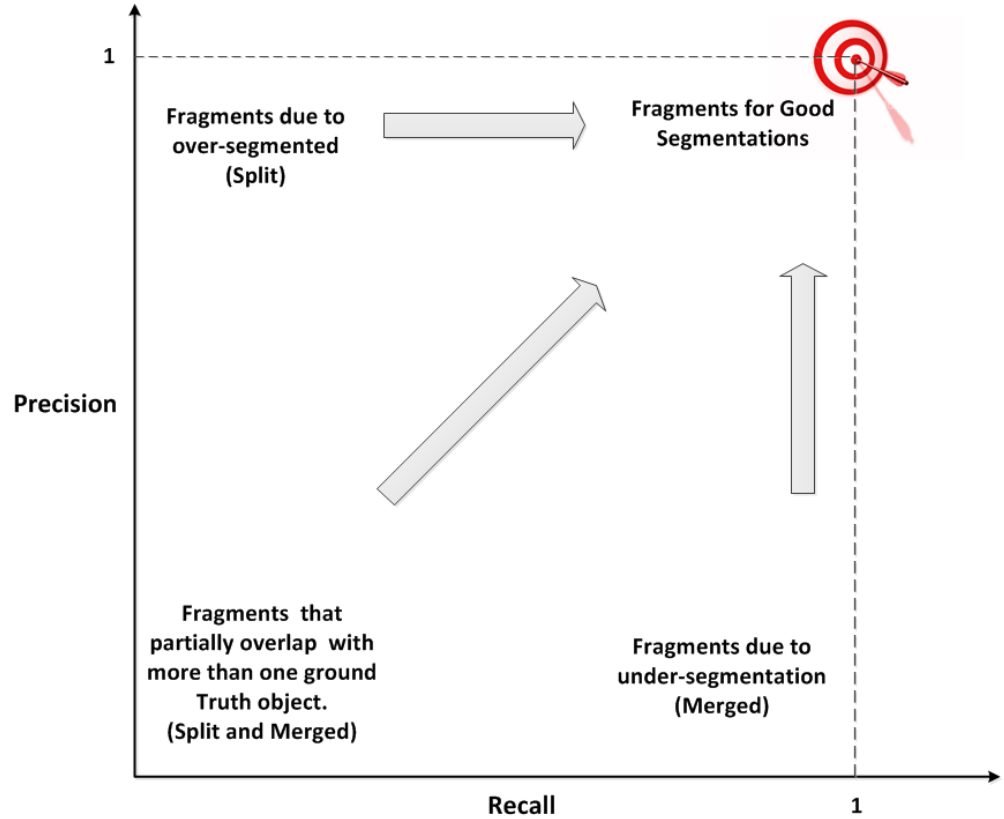
Precision, P_{ij} and Recall, R_{ij} for the ij -th fragment, G_{ij} can be calculated as follows.

For $1 \le i \le M$ and $1 \le j \le N$

$$G_{ij} = T_i \cap S_j$$

$$R_{ij} = \frac{|G_{ij}|}{|T_i|} = \frac{|T_i \cap S_j|}{|T_i|}$$

$$P_{ij} = \frac{|G_{ij}|}{|S_j|} = \frac{|T_i \cap S_j|}{|S_j|}$$



Precision\Recall Definitions

Assume $I_g = \{T_1, T_2, \ldots T_M\}$ is the ground truth image, where T_i is the i-th object in I_g.

Assume $I_s = \{S_1, S_2, \ldots S_N\}$ is the segmented image, where S_j is the j-th segment in I_s.

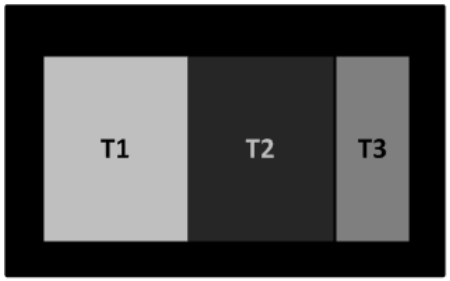
Precision, P_{ij} and Recall, R_{ij} for the ij -th fragment, G_{ij} can be calculated as follows.

For $1 \le i \le M$ and $1 \le j \le N$

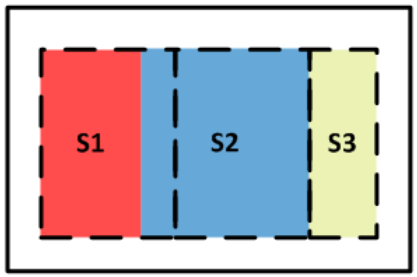
$$G_{ij} = T_i \cap S_j$$

$$R_{ij} = \frac{|G_{ij}|}{|T_i|} = \frac{|T_i \cap S_j|}{|T_i|}$$

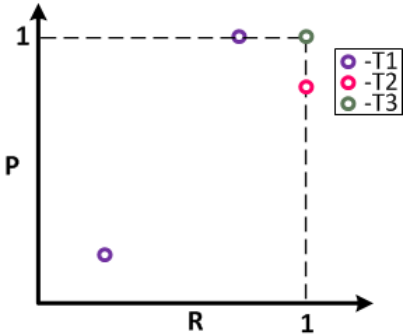
$$P_{ij} = \frac{|G_{ij}|}{|S_j|} = \frac{|T_i \cap S_j|}{|S_j|}$$



CT Data\Ground Truth



Segmented Image



Precision vs. Recall

R	S1	S2	S3
T1	0.75	0.25	0
T2	0	1	0
T3	0	0	1

P	S1	S2	S3
T1	1	0.2	0
T2	0	0.8	0
T3	0	0	1

F-Measure

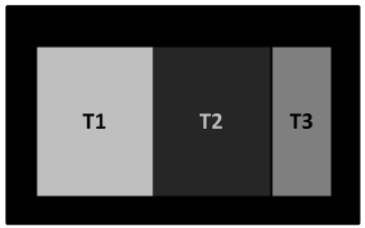
The F-Measure [1] is calculated for each fragment from their precision and recall as follows,

$$F_{ij} = \frac{2P_{ij}R_{ij}}{\left(P_{ij} + R_{ij}\right)} \qquad \text{when} \quad P_{ij} \neq 0, R_{ij} \neq 0$$

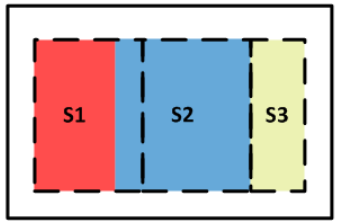
$$F_{ij} = 0 \qquad \text{Otherwise.}$$

In order to get one quantitative metric per dataset, we calculate a combined F-Measure as,

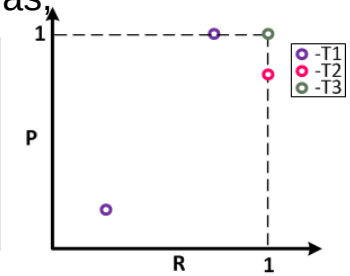
$$F_g = \frac{1}{\sum_{l=1}^{M} |T_l|} \sum_{i=1}^{M} \max_j \left(F_{ij}\right) |T_i|$$



CT Data\Ground Truth



Segmented Image



Precision vs. Recall

R	S1	S2	S3
T1	0.75	0.25	0
T2	0	1	0
T3	0	0	1

P	S1	S2	S3
T1	1	0.2	0
T2	0	0.8	0
T3	0	0	1

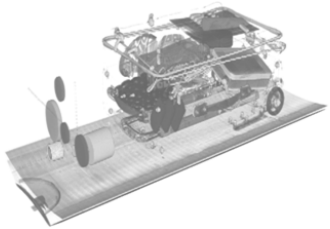
F	S1	S2	S3
T1	0.86	0.22	0
T2	0	0.89	0
T3	0	0	1

Fg = 0.86*0.4 + 0.89*0.4 + 1*0.2

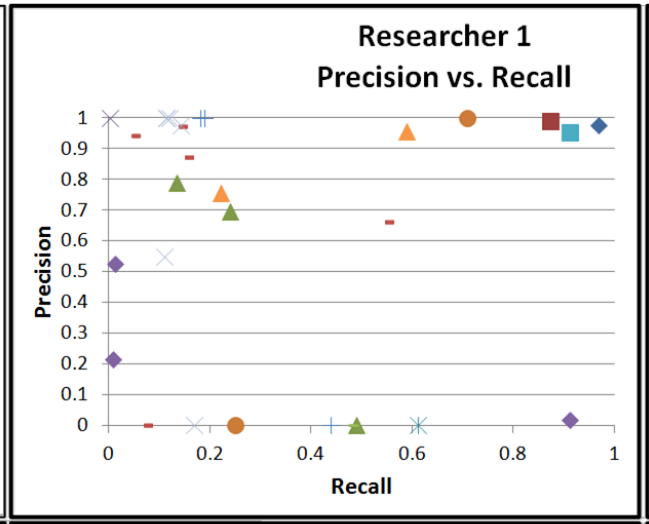
Fg = 0.9

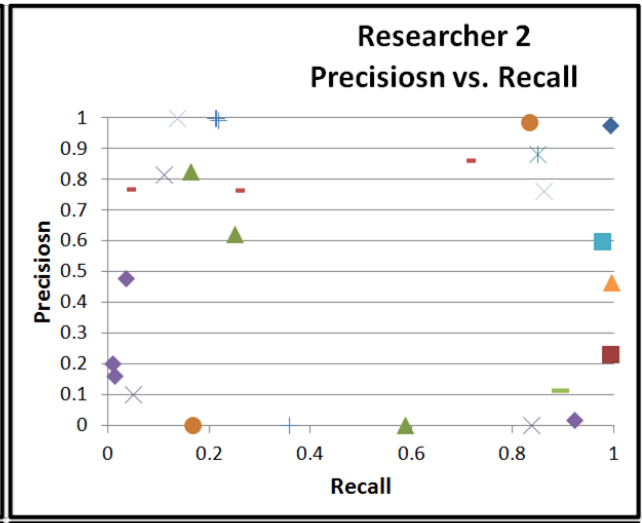
[1] van Rijsbergen, C. J. (1979). *Information Retrieval* (2nd ed.). Butterworth.

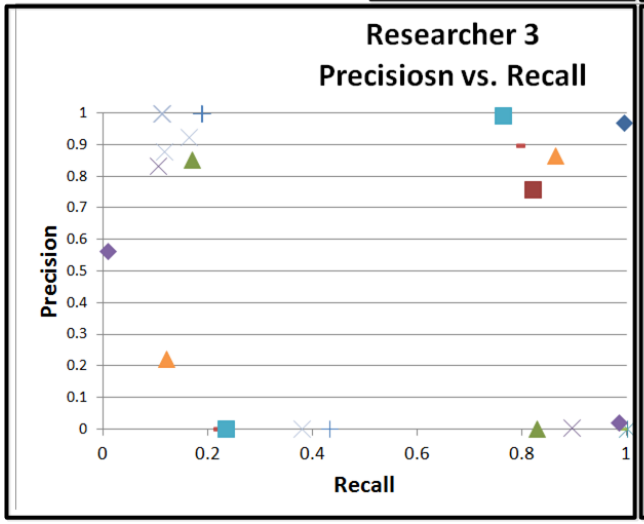
Training Bag 3
Precision vs. Recall

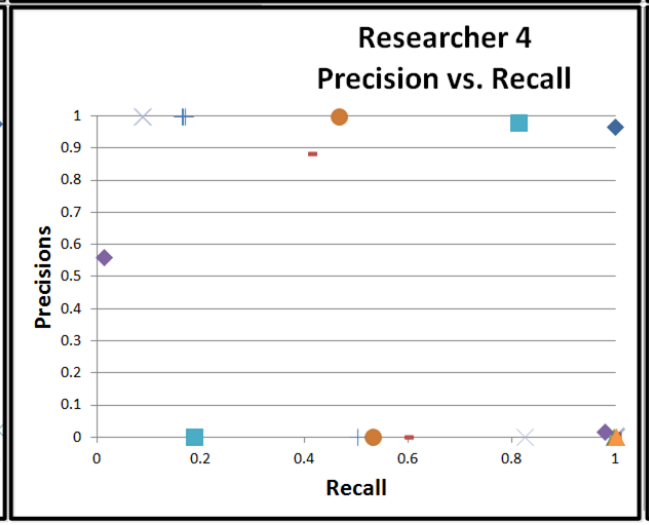


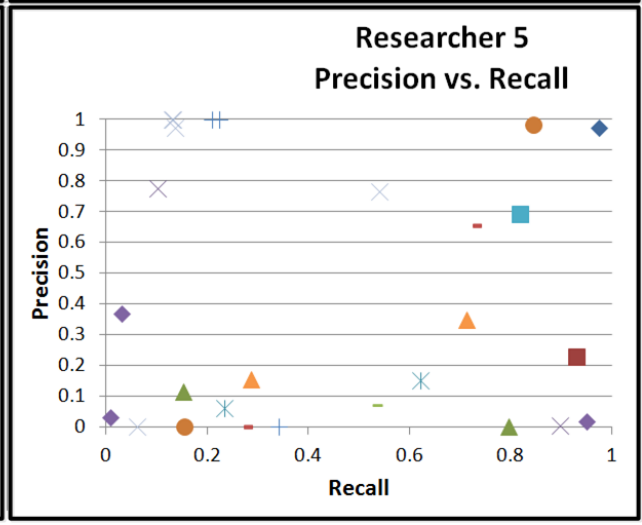
- Back Ground
- toothpaste tube
- Mens Sneaker - R
- Mens Sneaker - L
- Flat Iron
- CD's
- Bar Soap
- Candles - tealight
- Toothbrushes - 4 pk
- Leather Jacket
- Rubber (harder)
- Magazine - GH
- Skip Bo



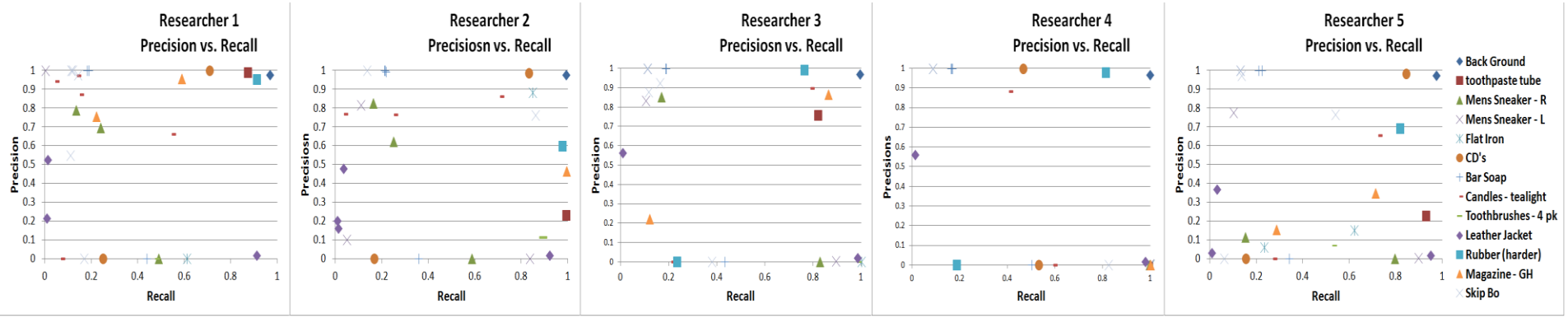


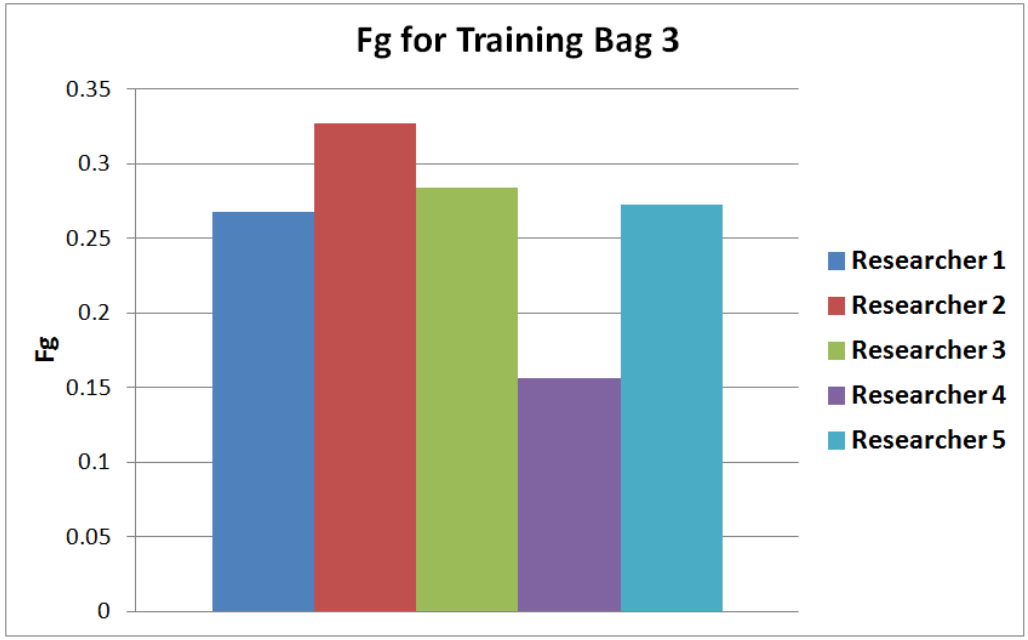


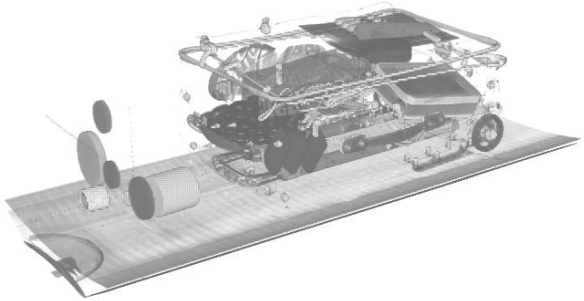




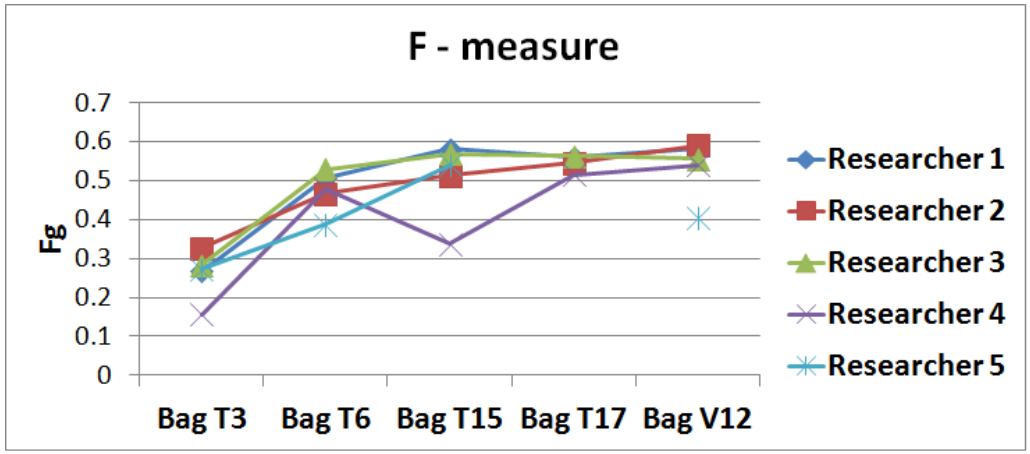
Researchers' Scores for Training Bag 3







Summary of Scores



- Based on the Fg metric, all researcher scores are in the same ball park. There is no one researcher that outshines the others in performance.

- Since we have not been able tie these scores back to system – level performance, we cannot say that small differences in Fg scores make an insignificant difference to over all system performance.

- Researchers 1, 2 & 3 have a similar trend across all the bags. Researchers 4 & 5 have much more variation in their scores across all the bags. This means that the performance of Researcher's 3 & 4 algorithms is not as consistent for varying data as Researcher's 1, 2 & 3.

Applicability to System-level Performance

Segmentation Evaluation Methods

Subjective Methods
Qualitative evaluation of segmentation results by a human evaluator.

Objective Methods
Quantitative evaluation of segmentation algorithms.

System-Level Methods
Methods that evaluate segmentation on the basis of the larger system's parameters. In the case of CT based images these parameters might be the following.

- μ(i), Linear Attenuation Coefficient of i-th object.
- V(i), volume of i-th object.

Direct Methods
Methods that evaluate segmentation independent of the larger system they are used in.

Analytical Methods
Theoretical evaluation methods that can be calculated without any results solely based on algorithm details.

Empirical Methods
Evaluation Methods that are calculated on the basis of the results of the segmentation algorithm.

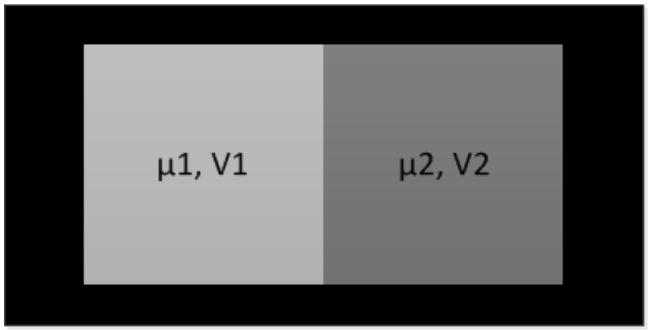
Unsupervised Methods
Evaluations methods that are based only on a set of segmentation results (no ground truth).

Supervised Methods
Evaluation methods that are based on the result of the segmentation algorithm and a ground truth image.

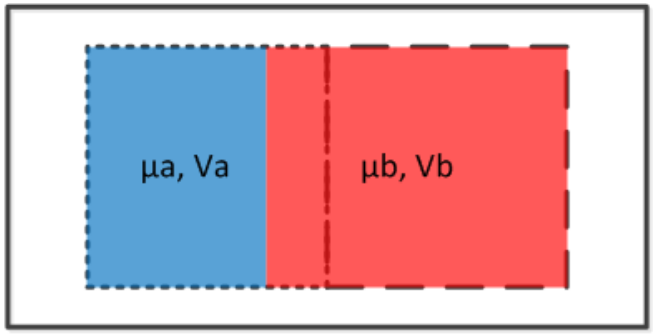
- P1\P2 Metric
- Martin Error (GCE\LCE)
- Object Consistency Error (OCE)
- F-Measure

It is important that supervised metrics correlate well with system performance.

Applicability to System-level Performance



Ground Truth



Segmentation

- **For CT and ATD, we really need to identify threats based on system-level values per segment**
  - **linear attenuation coefficient (*μ*) and**
  - **volume (V) of the segment**

- **As segmentation gets worse a good metric should also get worse.**
  - **Over-segmenting (splitting) can lead to correct *μ* and wrong V, while**
  - **Under-segmenting (merging) can lead to wrong *μ* and wrong V**

- **Current metric definitions allow a segment to match with more than one ground-truth object**
  - **Errors are calculated per ground-truth object (not per segment)**
  - **As the red segment merges more into Ground Truth Object 1, segmentation get worse but the current metrics get better after initially getting worse.**

We will need to modify these supervised metrics to make them more appropriate for system-level and ATD performance.

Summary/Future Work

- Summary
  - Segmentation of the reconstructed CT images is a key in explosive detection for airport security.
  - Studied how to measure segmentation performance and it turns out this is not a trivial task.
  - Surveyed the published literature on segmentation evaluation metrics and have developed a few ideas of our own.
  - Described one of the segmentation evaluation metrics we developed.
  - Present the results of applying this metric to the Segmentation Initiative researchers results

- Future work
  - Develop a segmentation metric that can be related back to system-level parameters.

Backup Slides

Precision\Recall Definitions (Example of Perfect Segmentation)

Assume $I_g = \{T_1, T_2, \ldots T_M\}$ is the ground truth image, where T_i is the i-th object in I_g.

Assume $I_s = \{S_1, S_2, \ldots S_N\}$ is the segmented image, where S_j is the j-th segment in I_s.

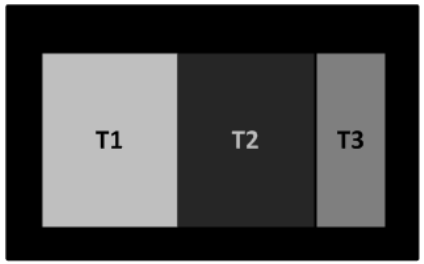
Precision, P_{ij} and Recall, R_{ij} for the ij -th fragment, G_{ij} can be calculated as follows.

For $1 \le i \le M$ and $1 \le j \le N$

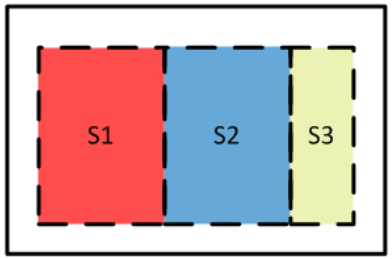
$$G_{ij} = T_i \cap S_j$$

$$R_{ij} = \frac{|G_{ij}|}{|T_i|} = \frac{|T_i \cap S_j|}{|T_i|}$$

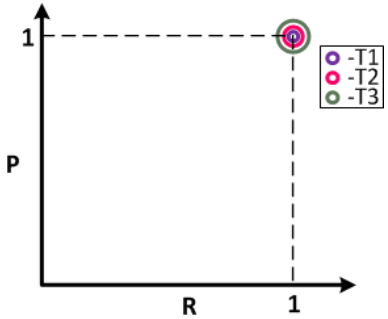
$$P_{ij} = \frac{|G_{ij}|}{|S_j|} = \frac{|T_i \cap S_j|}{|S_j|}$$



CT Data\Ground Truth



Segmented Image



Precision vs. Recall

R	S1	S2	S3
T1	1	0	0
T2	0	1	0
T3	0	0	1

P	S1	S2	S3
T1	1	0	0
T2	0	1	0
T3	0	0	1

Precision\Recall Definitions (Example of Splitting)

Assume $I_g = \{T_1, T_2, \ldots T_M\}$ is the ground truth image, where T_i is the i-th object in I_g.

Assume $I_s = \{S_1, S_2, \ldots S_N\}$ is the segmented image, where S_j is the j-th segment in I_s.

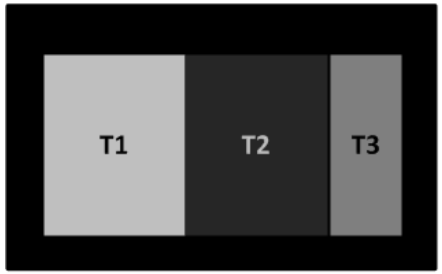
Precision, P_{ij} and Recall, R_{ij} for the ij -th fragment, G_{ij} can be calculated as follows.

For $1 \le i \le M$ and $1 \le j \le N$

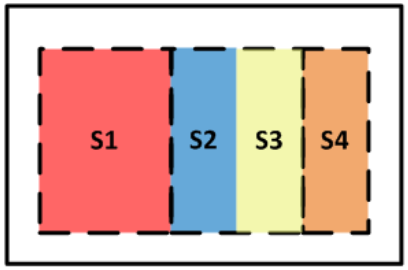
$$G_{ij} = T_i \cap S_j$$

$$R_{ij} = \frac{|G_{ij}|}{|T_i|} = \frac{|T_i \cap S_j|}{|T_i|}$$

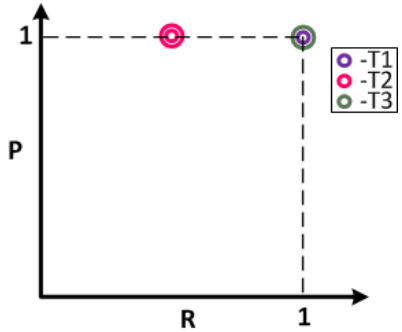
$$P_{ij} = \frac{|G_{ij}|}{|S_j|} = \frac{|T_i \cap S_j|}{|S_j|}$$



CT Data\Ground Truth



Segmented Image



Precision vs. Recall

R	S1	S2	S3	S4
T1	1	0	0	0
T2	0	0.5	0.5	0
T3	0	0	0	1

P	S1	S2	S3	S4
T1	1	0	0	0
T2	0	1	1	0
T3	0	0	0	1

Precision\Recall Definitions (Example of Merging)

Assume $I_g = \{T_1, T_2, \ldots T_M\}$ is the ground truth image, where T_i is the i-th object in I_g.

Assume $I_s = \{S_1, S_2, \ldots S_N\}$ is the segmented image, where S_j is the j-th segment in I_s.

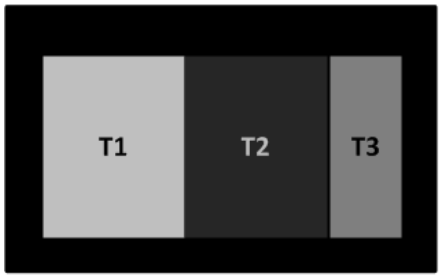
Precision, P_{ij} and Recall, R_{ij} for the ij -th fragment, G_{ij} can be calculated as follows.

For $1 \le i \le M$ and $1 \le j \le N$

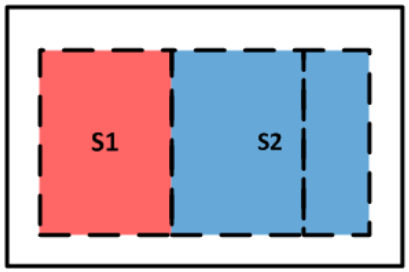
$$G_{ij} = T_i \cap S_j$$

$$R_{ij} = \frac{|G_{ij}|}{|T_i|} = \frac{|T_i \cap S_j|}{|T_i|}$$

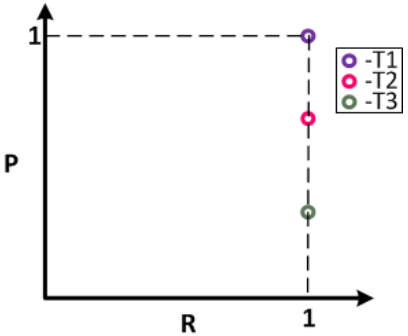
$$P_{ij} = \frac{|G_{ij}|}{|S_j|} = \frac{|T_i \cap S_j|}{|S_j|}$$



CT Data\Ground Truth



Segmented Image

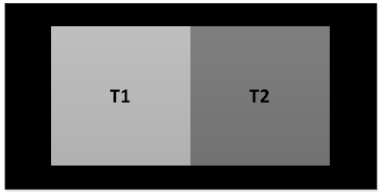


Precision vs. Recall

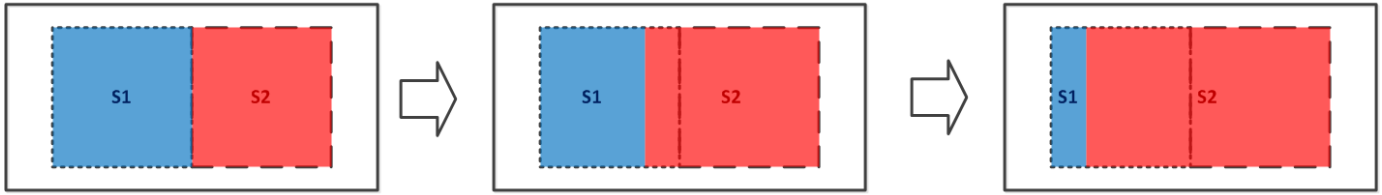
R	S1	S2
T1	1	0
T2	0	1
T3	0	1

P	S1	S2
T1	1	0
T2	0	0.67
T3	0	0.33

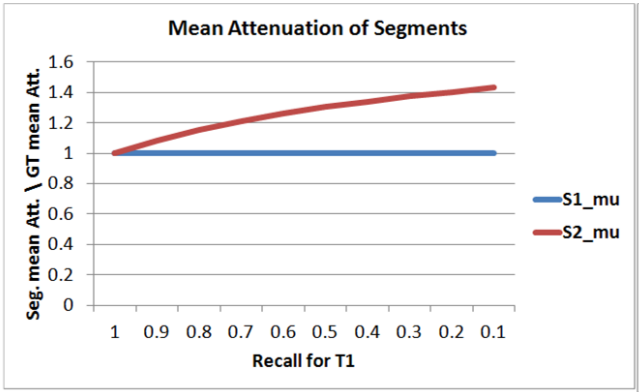
Applicability to System-level Metrics

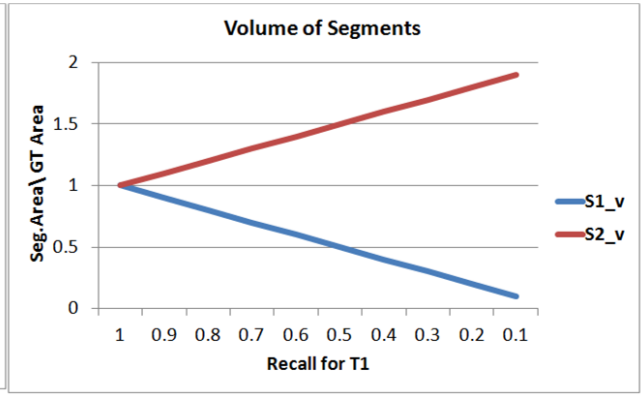


Ground Truth



Segmentation from perfect (left-most) with progressively underestimated ground truth Object 1 (to the right)

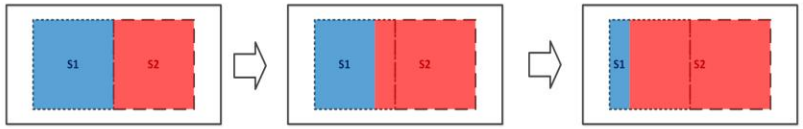




As S2 bleeds into T1, error in volume and mean attenuation for S2 increases. Therefore we should expect that the scoring metric should decrease from left to right.

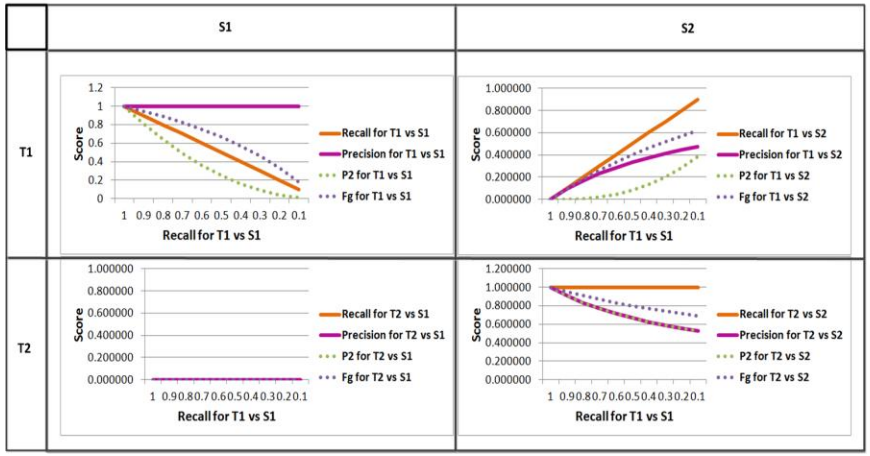
Applicability to System-level Metrics

Ground Truth

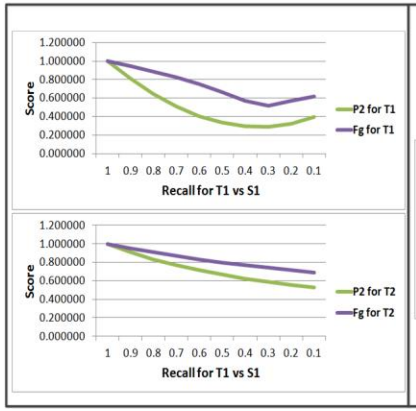


Segmentation from perfect (left-most) with progressively underestimated ground truth Object 1 (to the right)

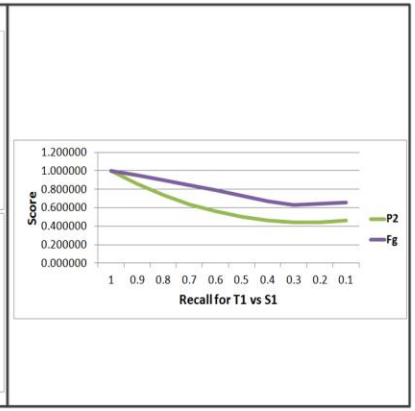
Recall, Precision, P2 and F for all fragments



P2 and F for Ground Truth (T1,T2)



P2 and Fg



- P2 and Fg decrease as the S2 bleeds into T1, until Precision and Recall for T1 are dominated by the Precision and Recall for the T1 vs. S2 fragment. After this point, P2 and Fg starts to increase even though intuitively the score should continue to decrease (since the segmentation continues to get worse).

This occurs because we are allowing the same segment (S2) to contribute to the score of more than one ground truth object (T1 and T2).

Proposed plan for developing a system-level applicable metric

Step 1 : Assign each segment to a single ground truth object.

- Hungarian algorithm to come up with the optimal assignment.

- The cost can be based on the on multiple features such as overlap, distance between centroids, principal axes, distance to mean attenuation etc.

Step 2: Calculate a single metric by combining the individual "score" for each segment (w.r.t. to it's assigned ground truth object from Step 1).

The individual score for each segment could be

- It's F-measure.
- Mathew's Correlation coefficient.
- A multi-feature based error (i.e. error between the segment's mean attenuation \volume and it's assigned ground truth object's mean attenuation\volume).