Evaluation Measures for Segmentation

By: Guillaume Lemaître
Eng Wei Yong

Overview

- Introduction
- Objective
- Evaluation Criteria
- Supervised methods
- Unsupervised methods
- Comparison of different methods
- Conclusion

Introduction

Evaluation Measures for Segmentation

- **Segmentation**: an essential process in image processing, medical imaging, machine vision
- **Evaluation Measures**: Development of tools/techniques to measure & compare the performance of segmentation algorithms
- **Performance**: depends on the application
  - computational efficiency/stability
  - mimics human perceptual segmentation
- Receive less attention than image segmentation itself

Objective

- Essential for application developers & researchers to choose the suitable techniques
- Accurately measures the performance of an algorithm
- To improve & justify new methods via formal comparison with existing methods

Evaluation Criteria

- **Accuracy**: how well the results agree with the human perception
- **Efficiency**: amount of time/effort required for segmentation
- **Precision**: degree to which the same result would be produced over different segmentation sessions

Supervised methods

- Evaluation Metrics Based
  - Rand index
  - Jaccard index
  - Fowkles and Mallows index
- Local and global consistency error (LCE – GCE)
- Huang and Dom evaluation measure

Evaluation metrics based Confusion matrix

		Actual value			
		GT_1	GT_2	……	GT_n
Prediction outcome	S_1	M_{11}	M_{12}	……	M_{1n}
	S_2	M_{21}	M_{22}	……	M_{2n}
	… …	……	……	……	……
	S_m	M_{m1}	M_{m2}	……	M_{mn}

n regions on Ground Truth

m regions on Segmented image

Example of confusion matrix

Evaluation metrics based Confusion matrix

		Actual value			
		GT_1	GT_2	……	GT_n
Prediction outcome	S_1	M_{11}		……	
	S_2		M_{22}	……	
	… …	……	……	……	……
	S_m			……	M_{mn}

$$n_{11} = \frac{1}{2}\left[\sum_{i=1}^{k}\sum_{j=i}^{l} {M_{ij}}^2 - n\right]$$

Evaluation metrics based Confusion matrix

		Actual value			
		GT_1	GT_2	……	GT_n
Prediction outcome	S_1	M_{11}		……	
	S_2	M_{21}		……	
	… …	……	……	……	……
	S_m	M_{m1}		……	

Square sum

Square diagonal element

$$n_{10} = \frac{1}{2}\left[\sum_{i=1}^{k}|GT_i|^2 - \sum_{i=1}^{k}\sum_{j=i}^{l}M_{ij}{}^2\right]$$

Evaluation metrics based Confusion matrix

		Actual value			
		GT_1	GT_2	……	GT_n
Prediction outcome	S_1		M_{12}	……	
	S_2		M_{22}	……	
	… …	……	……	……	……
	S_m		M_{m2}	……	

$$n_{10} = \frac{1}{2}\left[\sum_{i=1}^{k}|GT_i|^2 - \sum_{i=1}^{k}\sum_{j=i}^{l}M_{ij}{}^2\right]$$

Square sum

Square diagonal element

Evaluation metrics based Confusion matrix

		Actual value			
		GT_1	GT_2	……	GT_n
Prediction outcome	S_1	M_{11}	M_{12}	……	M_{1n}
	S_2	M_{21}	M_{22}	……	M_{2n}
	… …	……	……	……	……
	S_m	M_{m1}	M_{m2}	……	M_{mn}

$$n_{10} = \frac{1}{2}\left[\sum_{i=1}^{k}|GT_i|^2 - \sum_{i=1}^{k}\sum_{j=i}^{l} M_{ij}{}^2\right]$$

→ n_{10} = sum of verticals squared minus the sum of the diagonal squared

Evaluation metrics based Confusion matrix

		Actual value			
		GT_1	GT_2	……	GT_n
Prediction outcome	S_1	M_{11}	M_{12}	……	M_{1n}
	S_2			……	
	… …	……	……	……	……
	S_m			……	

Square sum

Square diagonal element

$$n_{01} = \frac{1}{2}\left[\sum_{j=1}^{l}|S_j|^2 - \sum_{i=1}^{k}\sum_{j=i}^{l} M_{ij}{}^2\right]$$

Evaluation metrics based Confusion matrix

		Actual value			
		GT_1	GT_2	……	GT_n
Prediction outcome	S_1			……	
	S_2	M_{21}	M_{22}	……	M_{2n}
	… …	……	……	……	……
	S_m			……	

Square sum

Square diagonal element

$$n_{01} = \frac{1}{2}\left[\sum_{j=1}^{l}|S_j|^2 - \sum_{i=1}^{k}\sum_{j=i}^{l} M_{ij}{}^2\right]$$

Evaluation metrics based Confusion matrix

		Actual value			
		GT_1	GT_2	……	GT_n
Prediction outcome	S_1	M_{11}	M_{12}	……	M_{1n}
	S_2	M_{21}	M_{22}	……	M_{2n}
	… …	……	……	……	……
	S_m	M_{m1}	M_{m2}	……	M_{mn}

$$n_{01} = \frac{1}{2}\left[\sum_{j=1}^{l}\left|S_j\right|^2 - \sum_{i=1}^{k}\sum_{j=i}^{l} {M_{ij}}^2\right]$$

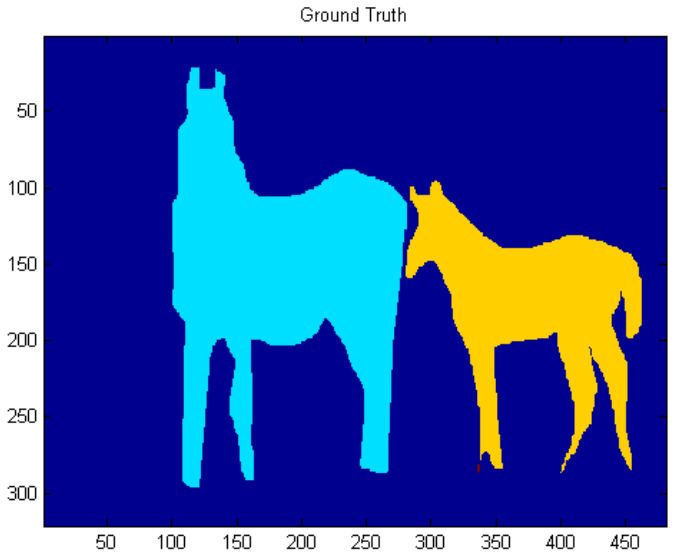
→ n_{10} = sum of horizontals squared minus the sum of the diagonal squared

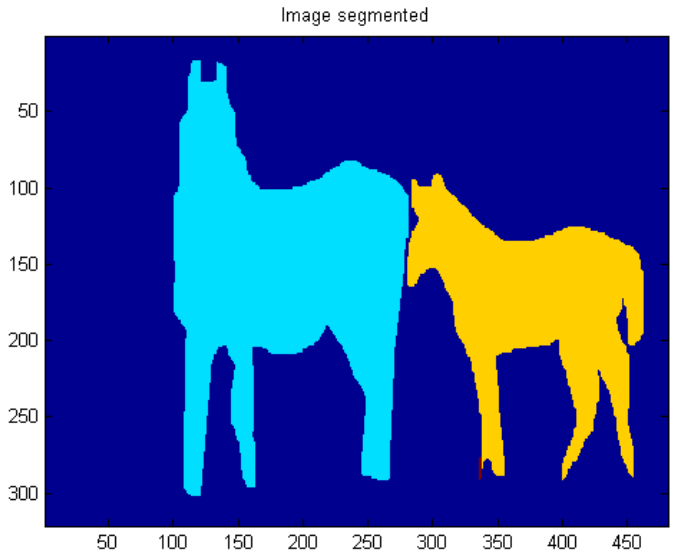
Evaluation metrics based Confusion matrix

		Actual value			
		GT_1	GT_2	……	GT_n
Prediction outcome	S_1	M_{11}	M_{12}	……	M_{1n}
	S_2	M_{21}	M_{22}	……	M_{2n}
	… …	……	……	……	……
	S_m	M_{m1}	M_{m2}	……	M_{mn}

$$n_{00} = \frac{n(n-1)}{2} - n_{11} - n_{10} - n_{01}$$

Evaluation metrics based Confusion matrix

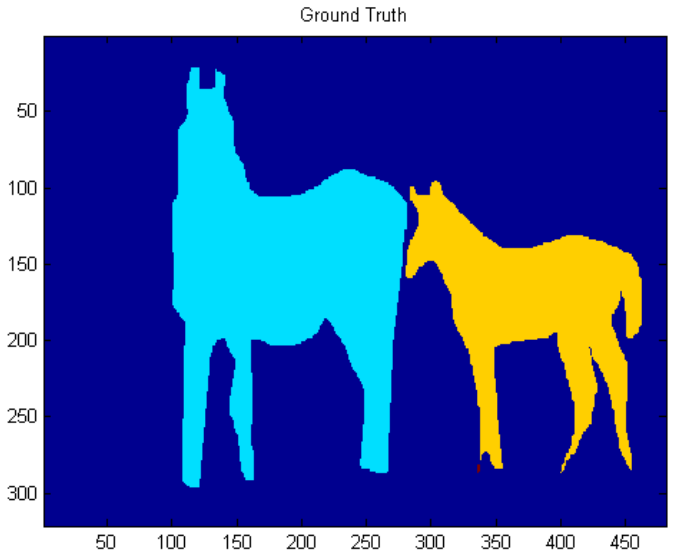


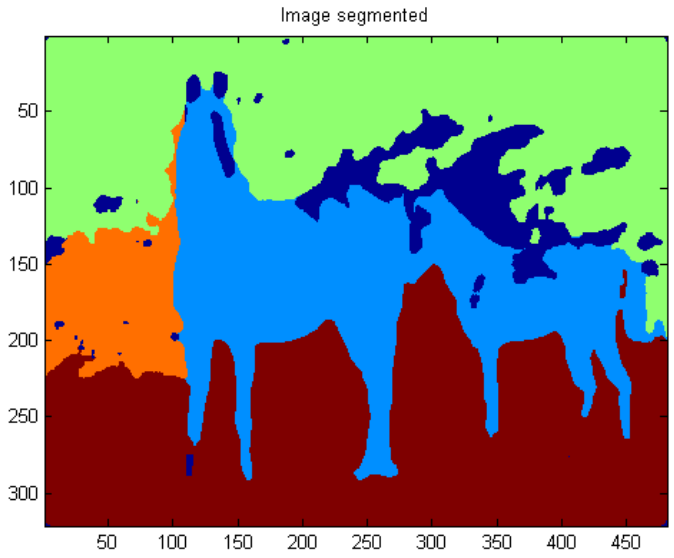


Evaluation metrics based Confusion matrix

		Actual value			
		GT_1	GT_2	GT_3	GT_4
Prediction outcome	S_1	110107	0	0	0
	S_2	1970	25447	0	0
	S_3	2282	0	14566	0
	S_4	20	0	0	9

Evaluation metrics based Confusion matrix





Evaluation metrics based Confusion matrix

		Actual value			
		GT_1	GT_2	GT_3	GT_4
Prediction outcome	S_1	7780	1771	1249	0
	S_2	1301	22027	11405	0
	S_3	46681	276	42	0
	S_4	9263	147	0	0
	S_5	49354	1226	1870	9

Evaluation metrics based Rand index

$$\mathcal{R}(GT,S) = 1 - \frac{n_{11} + n_{00}}{n(n-1)/2}$$

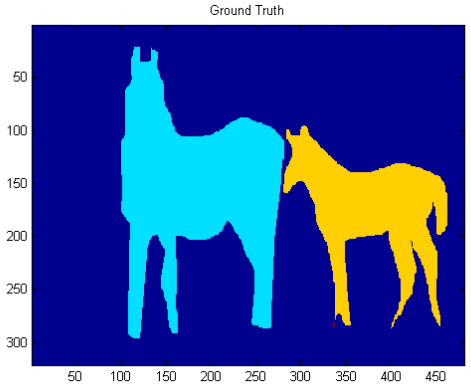
Typical values:

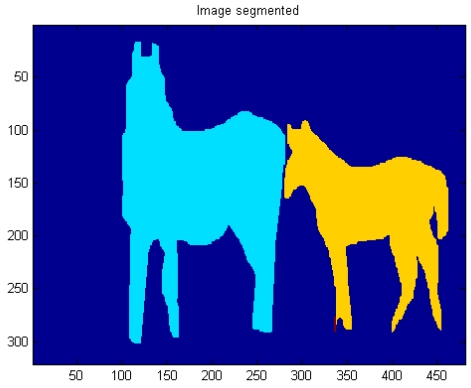
- 1: Large error
- 0: Identical images

→ Give the accuracy of the segmentation:

- *Represents the closeness of the Ground Truth and the segmented image*

Evaluation metrics based Rand index

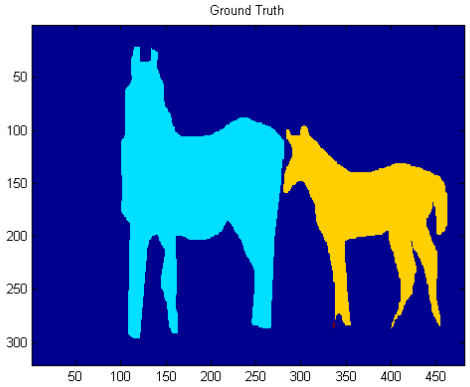


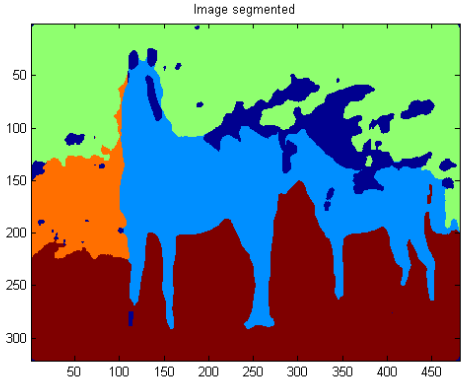


		Actual value			
		GT1	GT2	GT3	GT4
Prediction outcome	S1	110107	0	0	0
	S2	1970	25447	0	0
	S3	2282	0	14566	0
	S4	20	0	0	9

$$\mathcal{R}(GT, S) = 0.0476$$

Evaluation metrics based Rand index





		Actual value			
		GT1	GT2	GT3	GT4
Prediction outcome	S1	7780	1771	1249	0
	S2	1301	22027	11405	0
	S3	46681	276	42	0
	S4	9263	147	0	0
	S5	49354	1226	1870	9

$$\mathcal{R}(GT, S) = 0.7951$$

Evaluation metrics based Jaccard index

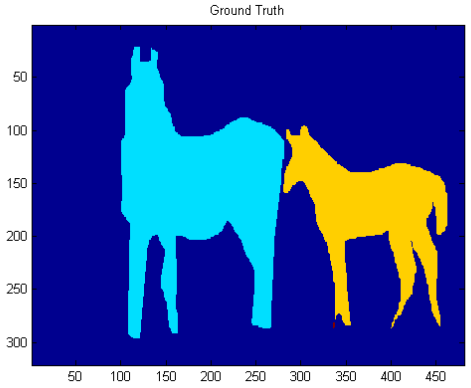
$$\Im(GT,S) = 1 - \frac{n_{11}}{n_{11} + n_{01} + n_{10}}$$

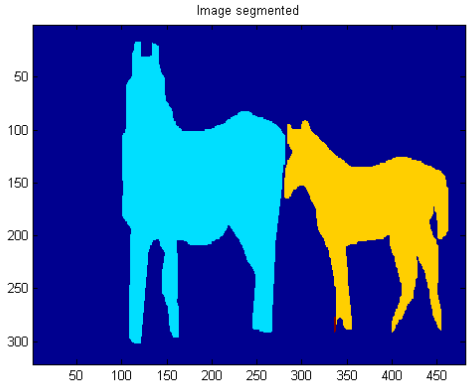
Typical values:

- 1: Large error
- 0: Identical images

→ Give the similarities between the Ground Truth and the segmented image

Evaluation metrics based Jaccard index

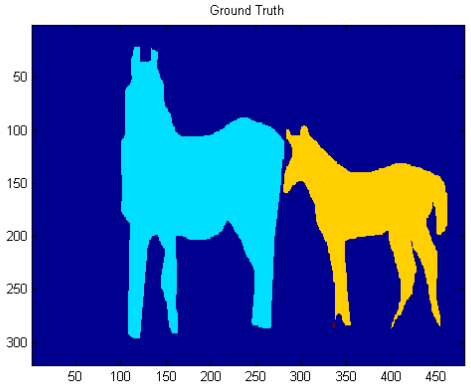


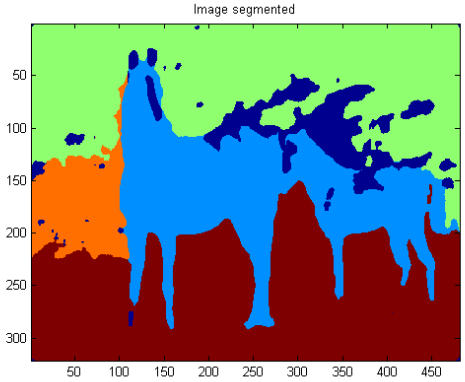


		Actual value			
		GT1	GT2	GT3	GT4
Prediction outcome	S1	110107	0	0	0
	S2	1970	25447	0	0
	S3	2282	0	14566	0
	S4	20	0	0	9

$$\Im(GT, S) = 0.0804$$

Evaluation metrics based Jaccard index





		Actual value			
		GT1	GT2	GT3	GT4
Prediction outcome	S1	7780	1771	1249	0
	S2	1301	22027	11405	0
	S3	46681	276	42	0
	S4	9263	147	0	0
	S5	49354	1226	1870	9

$$\Im(GT, S) = 0.9653$$

Evaluation metrics based Fowkles and Mallows index

$$\mathcal{F}(GT,S) = 1 - \sqrt{W_1(GT,S)W_2(GT,S)}$$

$$W_1(GT,S) = \sum_{i=1}^{k} \frac{n_{11}}{|GT_i|(|GT_i|-1)/2}$$

$$W_2(GT,S) = \sum_{j=1}^{l} \frac{n_{11}}{|S_j|(|S_j|-1)/2}$$

Evaluation metrics based Fowkles and Mallows index

		Actual value			
		GT_1	GT_2	……	GT_n
Prediction outcome	S_1	M_{11}		……	
	S_2	M_{21}		……	
	… …	……	……	……	……
	S_m	M_{m1}		……	

$$W_1(GT,S) = \sum_{i=1}^{k} \frac{n_{11}}{|GT_i|(|GT_i|-1)/2}$$

Sum = GT_1

Give the probability that the number of points in one cluster of GT are also in the same cluster of S

Evaluation metrics based Fowkles and Mallows index

		Actual value			
		GT_1	GT_2	……	GT_n
Prediction outcome	S_1	M_{11}	M_{12}	……	M_{1n}
	S_2			……	
	… …	……	……	……	……
	S_m			……	

→ Sum = S_1

$$W_2(GT,S) = \sum_{j=1}^{l} \frac{n_{11}}{|S_j|(|S_j|-1)/2}$$

→ Give the probability that the number of points in one cluster of S are also in the same cluster of GT

Evaluation metrics based Fowkles and Mallows index

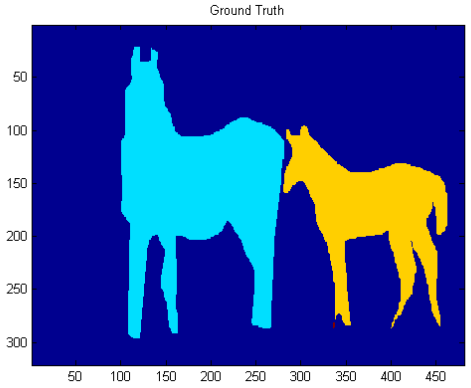
$$\mathcal{F}(GT,S) = 1 - \sqrt{W_1(GT,S)W_2(GT,S)}$$

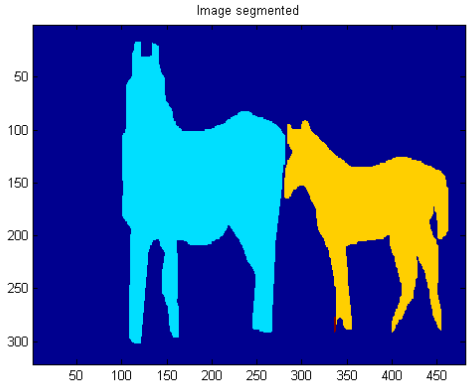
Typical values:

- 1: Large error
- 0: Identical images

→ Give the similarities between the Ground Truth and the segmented image

Evaluation metrics based Fowkles and Mallows index

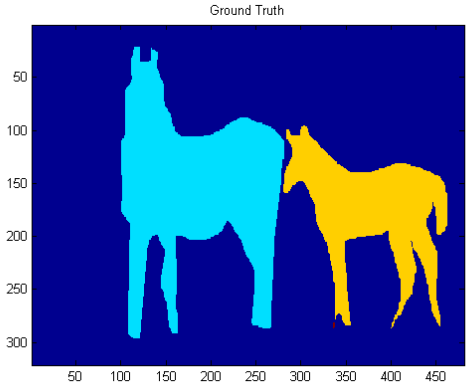


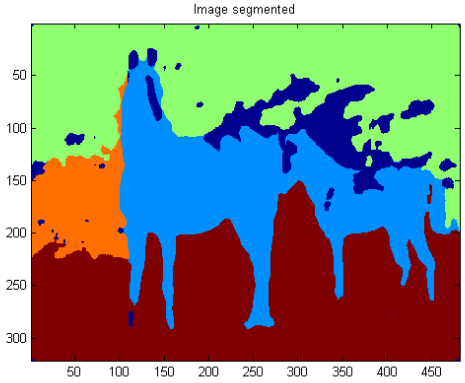


		Actual value			
		GT1	GT2	GT3	GT4
Prediction outcome	S1	110107	0	0	0
	S2	1970	25447	0	0
	S3	2282	0	14566	0
	S4	20	0	0	9

$$\mathcal{F}(GT, S) = 0.0415$$

Evaluation metrics based Fowkles and Mallows index





		Actual value			
		GT1	GT2	GT3	GT4
Prediction outcome	S1	7780	1771	1249	0
	S2	1301	22027	11405	0
	S3	46681	276	42	0
	S4	9263	147	0	0
	S5	49354	1226	1870	9

$$\mathcal{F}(GT, S) = 0.9278$$

Local and global consistency error

- Previous measures: Ground Truth was the reference
- Ground Truth can depend of the human perception
- Local and Global Consistency Error evaluate the dissimilarities between the Ground-Truth and the segmented image but between the segmented image and the Ground truth.

Local and global consistency error

- Local refinement error between clusters of the Ground Truth and the segmented image:

$$E(GT, S, p_i) = \frac{|\mathcal{R}(GT, p_i) \setminus \mathcal{R}(S, p_i)|}{|\mathcal{R}(GT, p_i)|}$$

- Local refinement error between clusters of the segmented image and the Ground Truth:

$$E(S, GT, p_i) = \frac{|\mathcal{R}(S, p_i) \setminus \mathcal{R}(GT, p_i)|}{|\mathcal{R}(S, p_i)|}$$

Local and global consistency error

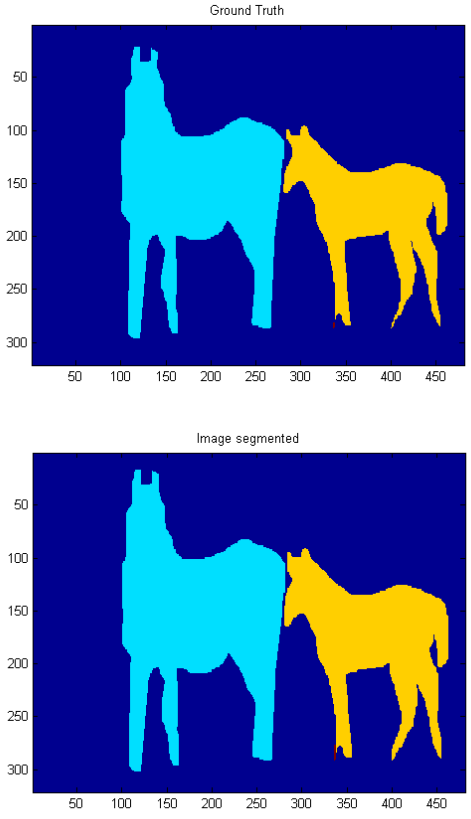
- Local Consistency Error – LCE:

$$LCE = \frac{1}{n}\sum_{p_i} \min\{E(GT, S, p_i), E(S, GT, p_i)\}$$

n is the number of pixels
p_i is a pixel of the image

Typical values:
- 1: Large error
- 0: Identique images

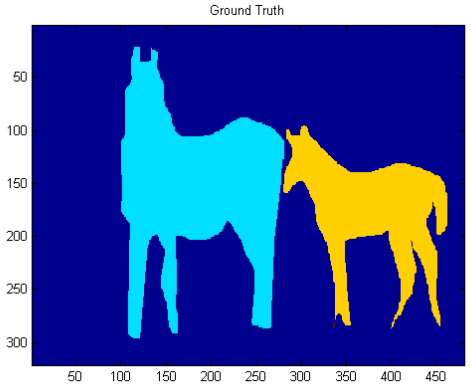
Local Consistency Error

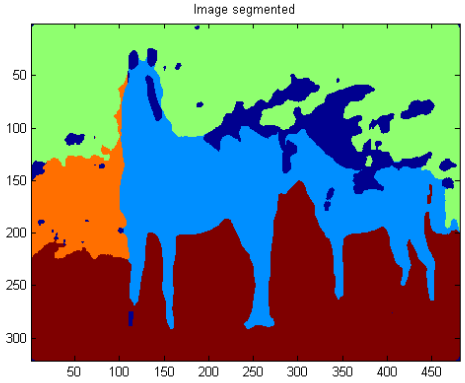


		Actual value			
		GT1	GT2	GT3	GT4
Prediction outcome	S1	110107	0	0	0
	S2	1970	25447	0	0
	S3	2282	0	14566	0
	S4	20	0	0	9

$$LCE = 0.0247$$

Local Consistency Error





		Actual value			
		GT1	GT2	GT3	GT4
Prediction outcome	S1	7780	1771	1249	0
	S2	1301	22027	11405	0
	S3	46681	276	42	0
	S4	9263	147	0	0
	S5	49354	1226	1870	9

$$LCE = 0.1171$$

Local and global consistency error

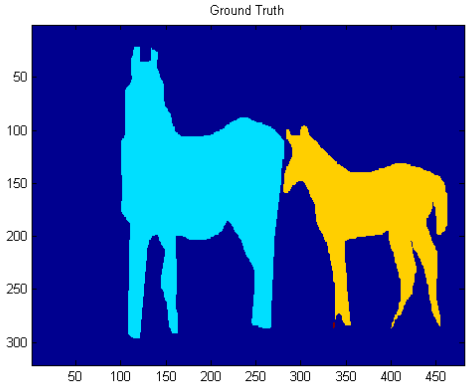
- Global Consistency Error – GCE:

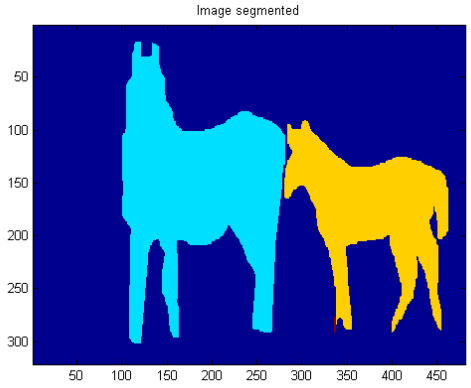
$$GCE = \frac{1}{n} \min \left\{ \sum_{p_i} E(GT, S, p_i), \sum_{p_i} E(S, GT, p_i) \right\}$$

n is the number of pixels
p_i is a pixel of the image

Typical values:
- 1: Large error
- 0: Identical images

Global Consistency Error

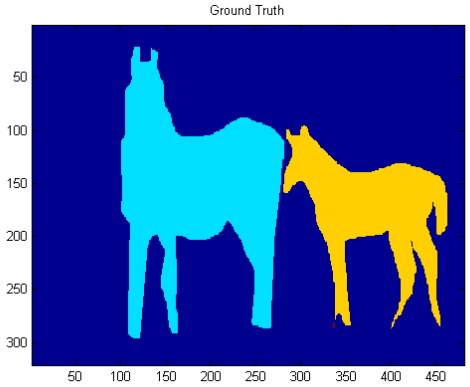


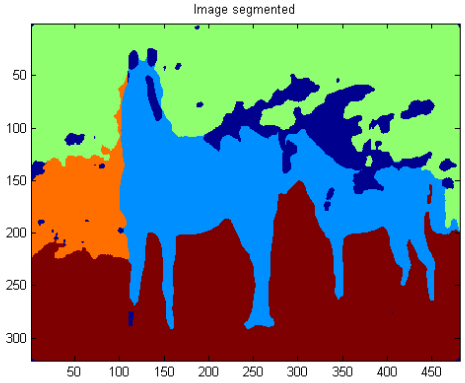


		Actual value			
		GT1	GT2	GT3	GT4
Prediction outcome	S1	110107	0	0	0
	S2	1970	25447	0	0
	S3	2282	0	14566	0
	S4	20	0	0	9

$$GCE = 0.0493$$

Global Consistency Error





		Actual value			
		GT1	GT2	GT3	GT4
Prediction outcome	S1	7780	1771	1249	0
	S2	1301	22027	11405	0
	S3	46681	276	42	0
	S4	9263	147	0	0
	S5	49354	1226	1870	9

$$GCE = 0.1851$$

Huang and Dom evaluation measure

Ignore refinement an degree of under or over-segmentation are important

$$HD = 1 - \frac{D_H(GT \rightarrow S) + D_H(S \rightarrow GT)}{2A}$$

D_H is the Hamming distance
A is the area of the image

Typical values:
- 1: Identical images
- 0: Large error

Huang and Dom evaluation measure

		Actual value			
		GT_1	GT_2	……	GT_n
Prediction outcome	S_1	M_{11}	M_{12}	……	M_{1n}
	S_2	M_{21}	M_{22}	……	M_{2n}
	… …	……	……	……	……
	S_m	M_{m1}	M_{m2}	……	M_{mn}

Sum – max(GT_1)

$$D_H(GT \rightarrow S) = \sum_i \sum_{j \neq \max(i)} M_{ij}$$

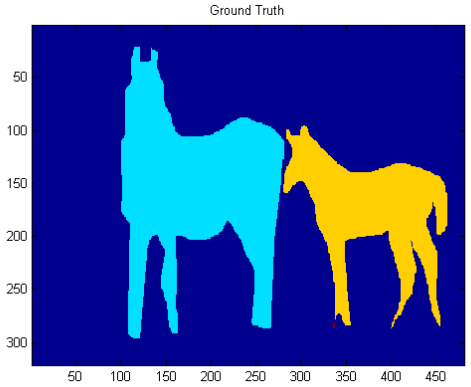
Huang and Dom evaluation measure

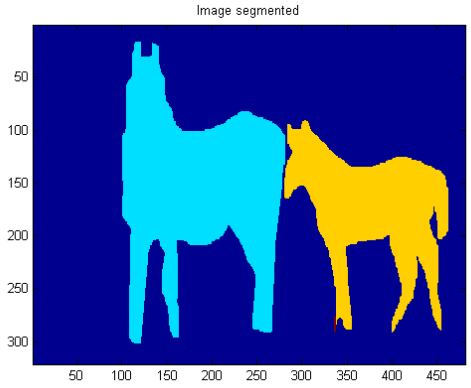
		Actual value			
		GT_1	GT_2	……	GT_n
Prediction outcome	S_1	M_{11}	M_{12}	……	M_{1n}
	S_2	M_{21}	M_{22}	……	M_{2n}
	… …	……	……	……	……
	S_m	M_{m1}	M_{m2}	……	M_{mn}

$$D_H(GT \rightarrow S) = \sum_i \sum_{j \neq \max(i)} M_{ij}$$

Sum – max(GT_2)

Huang and Dom evaluation measure

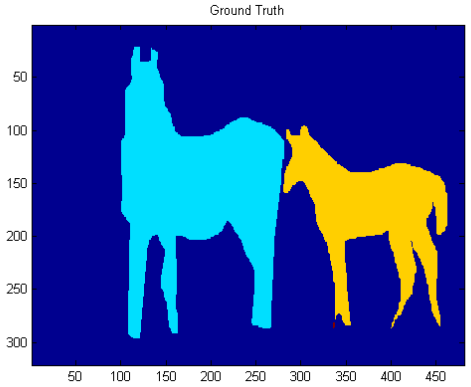


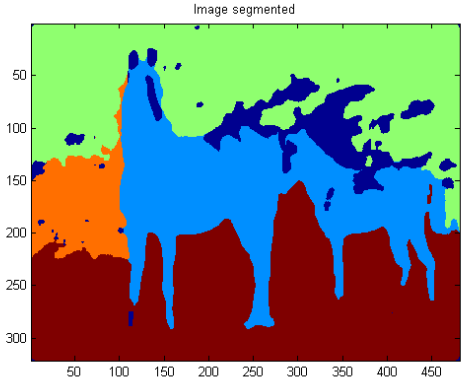


		Actual value			
		GT1	GT2	GT3	GT4
Prediction outcome	S1	110107	0	0	0
	S2	1970	25447	0	0
	S3	2282	0	14566	0
	S4	20	0	0	9

$$HD = 0.9724$$

Huang and Dom evaluation measure





		Actual value			
		GT1	GT2	GT3	GT4
Prediction outcome	S1	7780	1771	1249	0
	S2	1301	22027	11405	0
	S3	46681	276	42	0
	S4	9263	147	0	0
	S5	49354	1226	1870	9

$$HD = 0.7056$$

Unsupervised Evaluation

- Do not use a reference ground truth
- Intuitively observe output of a segmentation algorithm
  - over-segmentation, under-segmentation
  - jagged, or very symmetrical
- Fundamental understanding of human perceptual grouping
  - Focus on objective rather than method
- However, objective is hard to formalize.
- Successful measures
  - **Entropy Based Evaluation: information theory**
  - Visible Color Distance Based Evaluation: Perceived color distances.

Entropy Based Evaluation

- Measure pixel uniformity within a region & complexity of overall partitioning
  - number of regions
  - differences between adjacent regions
- Region Entropy – Region homogeneity
- Layout Entropy – No. of region

Region Entropy

$$H(R_i) = -\sum_x \frac{N_i(x)}{|R_i|} log \frac{N_i(x)}{|R_i|}$$

- $H(R_i)$ = Entropy of Region i
- $N_i(x)$ = No. of pixel in region i, R_i with value x
- Weighted sum of individual region entropies for image I

$$H_r(I) = \sum_i \frac{|R_i|}{|I|} H(R_i)$$

- ↑ equal feature points in a region, ↓ region entropy
- Biased towards oversegmentation
- ↑ no. of region, ↑ equal feature points, ↓ region entropy

Layout entropy

- no. of bits required to specify region to which each pixel belongs

$$H_l(I) = -\sum_i \frac{|R_i|}{|I|} log \frac{|R_i|}{|I|}$$

- Biased towards under segmentation
- ↓ No. of region, ↓ layout entropy
- Balanced by combining region & layout entropy

$$E = H_r(I) + H_l(I)$$

Comparison

Supervised Evaluations	Unsupervised Evaluations
Need ground truth	Not need ground truth
Ground truth maybe ambiguous for a complex scenery	Avoid ambiguity in ground truth
No training phase	Explicitly allowed a training phase
Cannot find optimal parameterization automatically	Automatically find optimal parameterization of a segmentation

Conclusion

- Two types of segmentation evaluation measures :
  - Supervised methods:
    - Metrics Based Evaluation
    - Local and Global Consistency
    - Huang and Dom Evaluation
  - Unsupervised methods
    - Entropy based methods

Bibliography

[1] Kevin McGuinness, “Image Segmentation, Evaluation, and Applications”, November 2009

[2] Yu Jin Zhang, “A Review of Recent Evaluation Methods for Image Segmentation”, August 2001

[3] Aaron Fenster, Bernard Chiu, “Evaluation of Segmentation algorithms for Medical Imaging”, September 2005

[4] Tom Fawcett, “An introduction to ROC analysis”, December 2005

[5] R. Unnikrishnan, C. Pantofaru, M. Hebert, “A Measure for Objective Evaluation of Image Segmentation Algorithms”, 2005