Meta Measures

How to Quantitatively Compare Evaluation Measures

Jordi Pont-Tuset

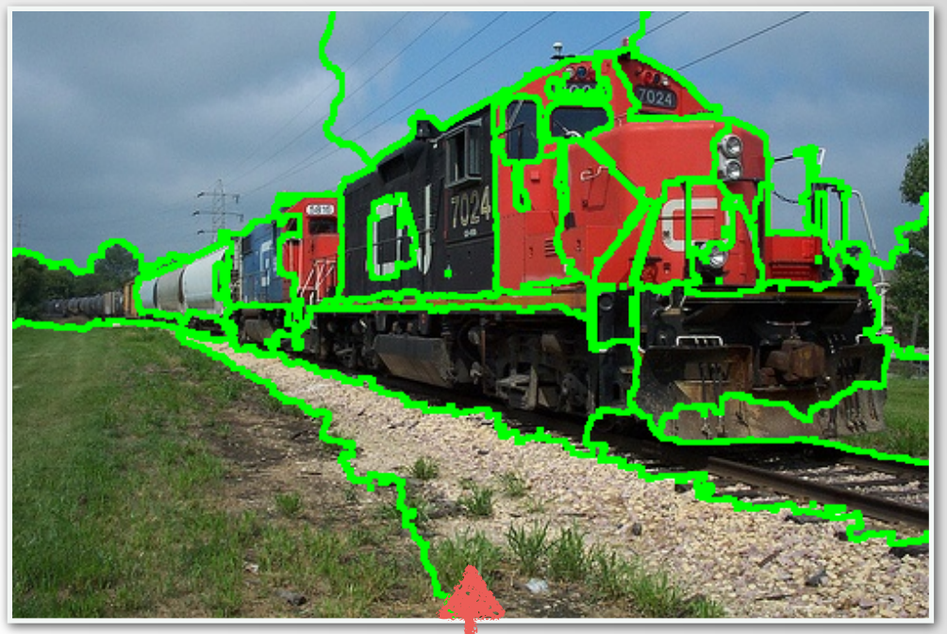
ECCV Workshop on Datasets and Performance Analysis in Early Vision
Amsterdam, 8th October 2016

ETH zürich
CVL Computer Vision Lab

Meta Measures

How to Quantitatively Compare Evaluation Measures

Result



Evaluation Measure

Ground Truth

Meta Measures

How to Quantitatively Compare Evaluation Measures

Result

Evaluation Measure

Ground Truth

- Distance Signature [Huang1995]
- Fuzzy Jaccard Index [McGuinnes2010]
- ROC Curves for Boundaries [Estrada2005][Estrada2009]
- Greedy Boundary Matching [Bowyer2001]
- Precision-Recall for Boundaries [Martin2003][Martin2004]
- Rand Index [Rand1971, Ben-Hur2002]
- Probabilistic Rand Index (PRI) [Unnikrishnan2005]
- Normalized Probabilistic Rand index (NPR) [Unnikrishnan2007]
- Precision-Recall for Regions [Martin2003]
- Directional Hamming distance [Kanungo94], [Huang95]
- Asymmetric Partition distance [Cardoso2005]
- Projection Number [Dongen2000], [Jiang2006]
- Larsen measure [Larsen1999]
- Segmentation Covering [Arbeláez2011]
- Symmetric Partition Distance [Gusfield2002], [Cardoso2005]
- Bipartite Graph Matching [Jiang2006]
- Classification Error Distance [Meila2005]
- Bidirectional, Local, and Global Consistency Errors [Martin2003]
- Variation of Information (VoI) [Meila2003], [Meila2005]
- Mirkin metric [Mirkin1996]
- Hoover Measures [Hoover1996], [Min2004]

Meta Measures

How to Quantitatively Compare Evaluation Measures

Evaluation Measure

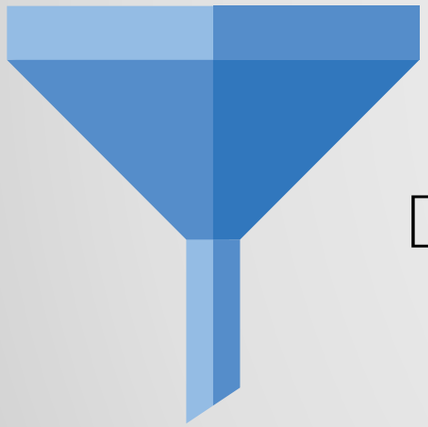
- Distance Signature [Huang1995]
- Fuzzy Jaccard Index [McGuinnes2010]
- ROC Curves for Boundaries [Estrada2005][Estrada2009]
- Greedy Boundary Matching [Bowyer2001]
- Precision-Recall for Boundaries [Martin2003][Martin2004]
- Rand Index [Rand1971, Ben-Hur2002]
- Probabilistic Rand Index (PRI) [Unnikrishnan2005]
- Normalized Probabilistic Rand index (NPR) [Unnikrishnan2007]
- Precision-Recall for Regions [Martin2003]
- Directional Hamming distance [Kanungo94], [Huang95]
- Asymmetric Partition distance [Cardoso2005]
- Projection Number [Dongen2000], [Jiang2006]
- Larsen measure [Larsen1999]
- Segmentation Covering [Arbeláez2011]
- Symmetric Partition Distance [Gusfield2002], [Cardoso2005]
- Bipartite Graph Matching [Jiang2006]
- Classification Error Distance [Meila2005]
- Bidirectional, Local, and Global Consistency Errors [Martin2003]
- Variation of Information (VoI) [Meila2003], [Meila2005]
- Mirkin metric [Mirkin1996]
- Hoover Measures [Hoover1996], [Min2004]

Meta Measures

How to Quantitatively Compare Evaluation Measures

Evaluation Measure

- Distance Signature [Huang1995]
- Fuzzy Jaccard Index [McGuinnes2010]
- ROC Curves for Boundaries [Estrada2005][Estrada2009]
- Greedy Boundary Matching [Bowyer2001]
- Precision-Recall for Boundaries [Martin2003][Martin2004]
- Rand Index [Rand1971, Ben-Hur2002]
- Probabilistic Rand Index (PRI) [Unnikrishnan2005]
- Normalized Probabilistic Rand index (NPR) [Unnikrishnan2007]
- Precision-Recall for Regions [Martin2003]
- Directional Hamming distance [Kanungo94], [Huang95]
- Asymmetric Partition distance [Cardoso2005]
- Projection Number [Dongen2000], [Jiang2006]
- Larsen measure [Larsen1999]
- Segmentation Covering [Arbeláez2011]
- Symmetric Partition Distance [Gusfield2002], [Cardoso2005]
- Bipartite Graph Matching [Jiang2006]
- Classification Error Distance [Meila2005]
- Bidirectional, Local, and Global Consistency Errors [Martin2003]
- Variation of Information (VoI) [Meila2003], [Meila2005]
- Mirkin metric [Mirkin1996]
- Hoover Measures [Hoover1996], [Min2004]



Deduplicate and Structure

- Distance Signature [Huang1995]
- Fuzzy Jaccard Index [McGuinnes2010]
- ROC Curves for Boundaries [Estrada2005][Estrada2009]
- Greedy Boundary Matching [Bowyer2001]
- Precision-Recall for Boundaries [Martin2003][Martin2004]

- Rand Index [Rand1971, Ben-Hur2002]
- Probabilistic Rand Index (PRI) [Unnikrishnan2005]
- Normalized Probabilistic Rand index (NPR) [Unnikrishnan2007]
- Precision-Recall for Regions [Martin2003]

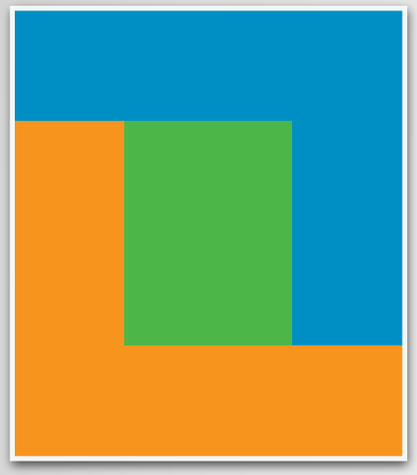
- Directional Hamming distance [Kanungo94], [Huang95],
- Asymmetric Partition distance [Cardoso2005]
- Projection Number [Dongen2000], [Jiang2006]
- Larsen measure [Larsen1999]
- Segmentation Covering [Arbeláez2011]
- Symmetric Partition Distance [Gusfield2002], [Cardoso2005],
- Bipartite Graph Matching [Jiang2006]
- Classification Error Distance [Meila2005]
- Bidirectional, Local, and Global Consistency Errors [Martin2003]
- Variation of Information (VoI) [Meila2003], [Meila2005]
- Mirkin metric [Mirkin1996]
- Hoover Measures [Hoover1996], [Min2004]

Exhaustive

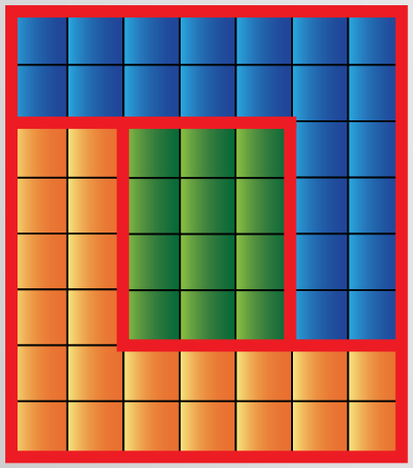
Structured

Deduplicate and Structure

Interpretations of a Partition

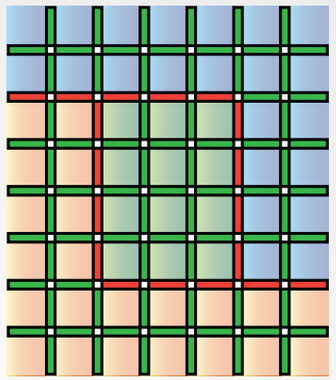


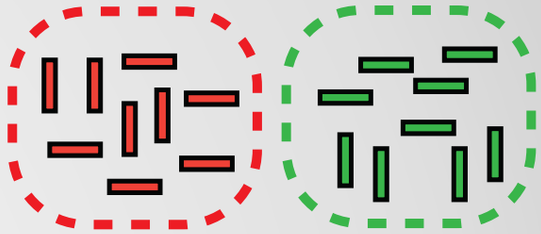
Image



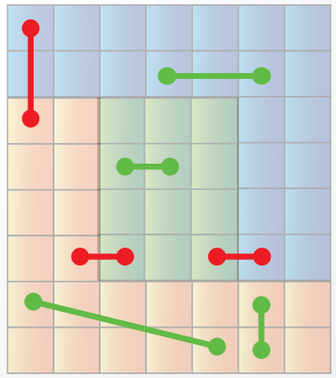
Partition

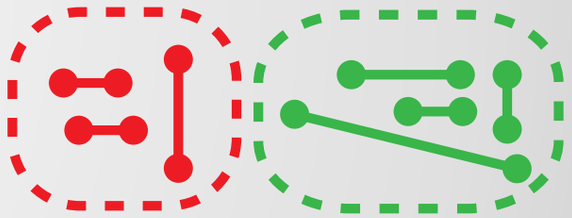
Two-class clustering of the pixel contour segments (boundary based)



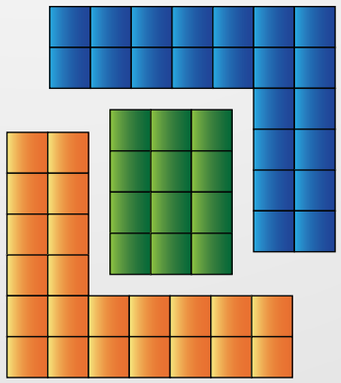


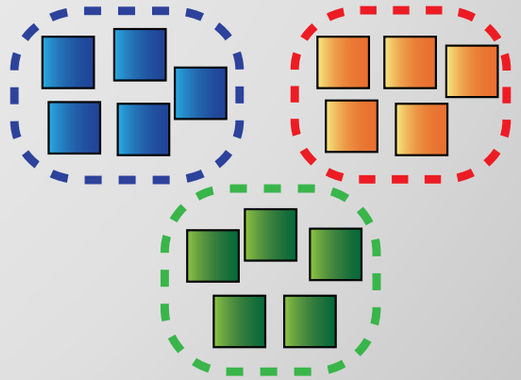
Two-class clustering of the pairs of pixels





Clustering of the set of pixels (region based)

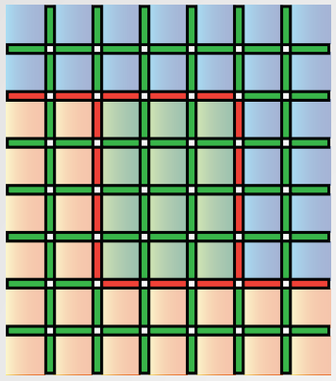




Deduplicate and Structure

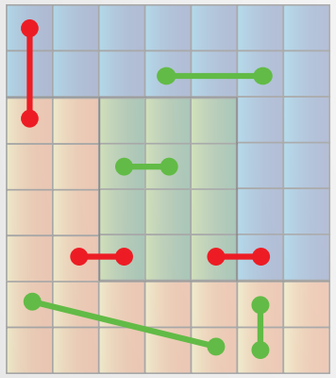
Interpretations of a Partition

Two-class clustering of the pixel contour segments (boundary based)



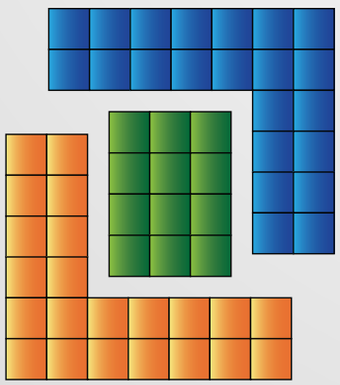
- Distance Signature [Huang1995]
- Fuzzy Jaccard Index [McGuinnes2010]
- ROC Curves for Boundaries [Estrada2005][Estrada2009]
- Greedy Boundary Matching [Bowyer2001]
- Precision-Recall for Boundaries [Martin2003][Martin2004]

Two-class clustering of the pairs of pixels



- Rand Index [Rand1971, Ben-Hur2002]
- Probabilistic Rand Index (PRI) [Unnikrishnan2005]
- Normalized Probabilistic Rand index (NPR) [Unnikrishnan2007]
- Precision-Recall for Regions [Martin2003]

Clustering of the set of pixels (region based)

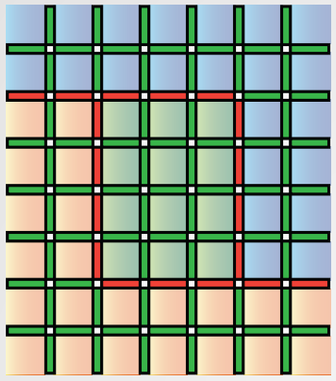


- Directional Hamming distance [Kanungo94], [Huang95],
- Asymmetric Partition distance [Cardoso2005]
- Projection Number [Dongen2000], [Jiang2006]
- Larsen measure [Larsen1999]
- Segmentation Covering [Arbeláez2011]
- Symmetric Partition Distance [Gusfield2002], [Cardoso2005],
- Bipartite Graph Matching [Jiang2006]
- Classification Error Distance [Meila2005]
- Bidirectional, Local, and Global Consistency Errors [Martin2003]
- Variation of Information (VoI) [Meila2003], [Meila2005]
- Mirkin metric [Mirkin1996]
- Hoover Measures [Hoover1996], [Min2004]

Deduplicate and Structure

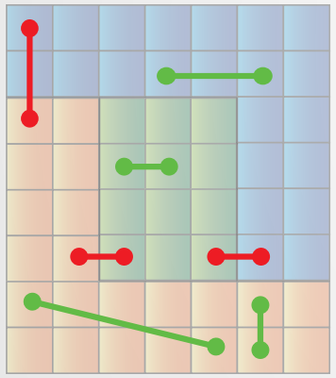
Interpretations of a Partition

Two-class clustering of the pixel contour segments (boundary based)



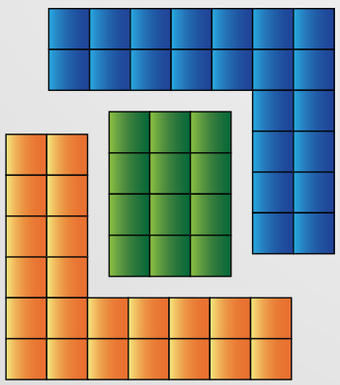
- Distance Signature [Huang1995]
- Fuzzy Jaccard Index [McGuinnes2010]
- ROC Curves for Boundaries [Estrada2005][Estrada2009]
- Greedy Boundary Matching [Bowyer2001]
- **Precision-Recall for Boundaries [Martin2003][Martin2004]**

Two-class clustering of the pairs of pixels



- Rand Index [Rand1971, Ben-Hur2002]
- Probabilistic Rand Index (PRI) [Unnikrishnan2005]
- Normalized Probabilistic Rand index (NPR) [Unnikrishnan2007]
- **Precision-Recall for Regions [Martin2003]**

Clustering of the set of pixels (region based)



- Directional Hamming distance [Kanungo94], [Huang95],
- Asymmetric Partition distance [Cardoso2005]
- Projection Number [Dongen2000], [Jiang2006]
- Larsen measure [Larsen1999]
- Segmentation Covering [Arbeláez2011]
- Symmetric Partition Distance [Gusfield2002], [Cardoso2005],
- Bipartite Graph Matching [Jiang2006]
- Classification Error Distance [Meila2005]
- Bidirectional, Local, and Global Consistency Errors [Martin2003]
- Variation of Information (VoI) [Meila2003], [Meila2005]
- Mirkin metric [Mirkin1996]
- Hoover Measures [Hoover1996], [Min2004]

Deduplicate and Structure

Interpretations of a Partition

To sum up…

…we had N measures…

…now we have N+1 measures…

META MEASURES!

Meta Measures

How to Quantitatively Compare Evaluation Measures

Axiom — *"An axiom is a statement that is taken to be true, to serve as a premise for further reasoning"*

Axiom 1 — *m(P,P) = 1*

Axiom 2 — *m(P,Q) ~ 1* — If *P* and *Q* are *very similar*

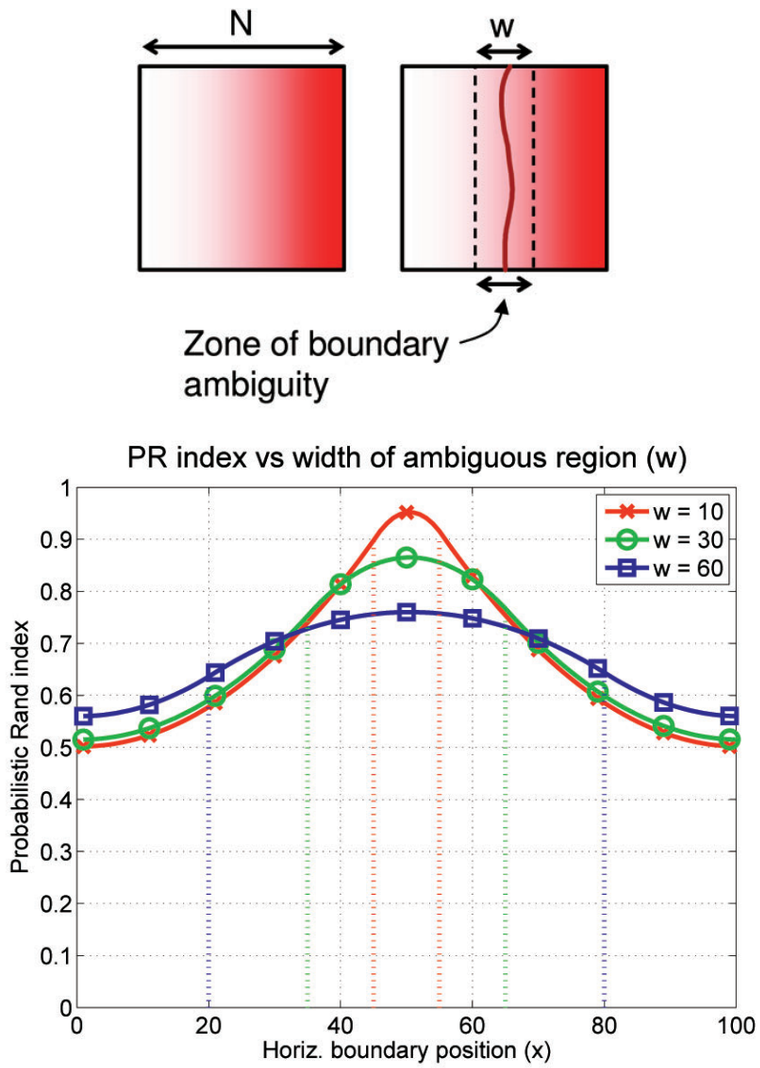
Axiom 3 — *m(P,Q) << 1* — If *P* and *Q* are *very different*

Meta Measures

How to Quantitatively Compare Evaluation Measures

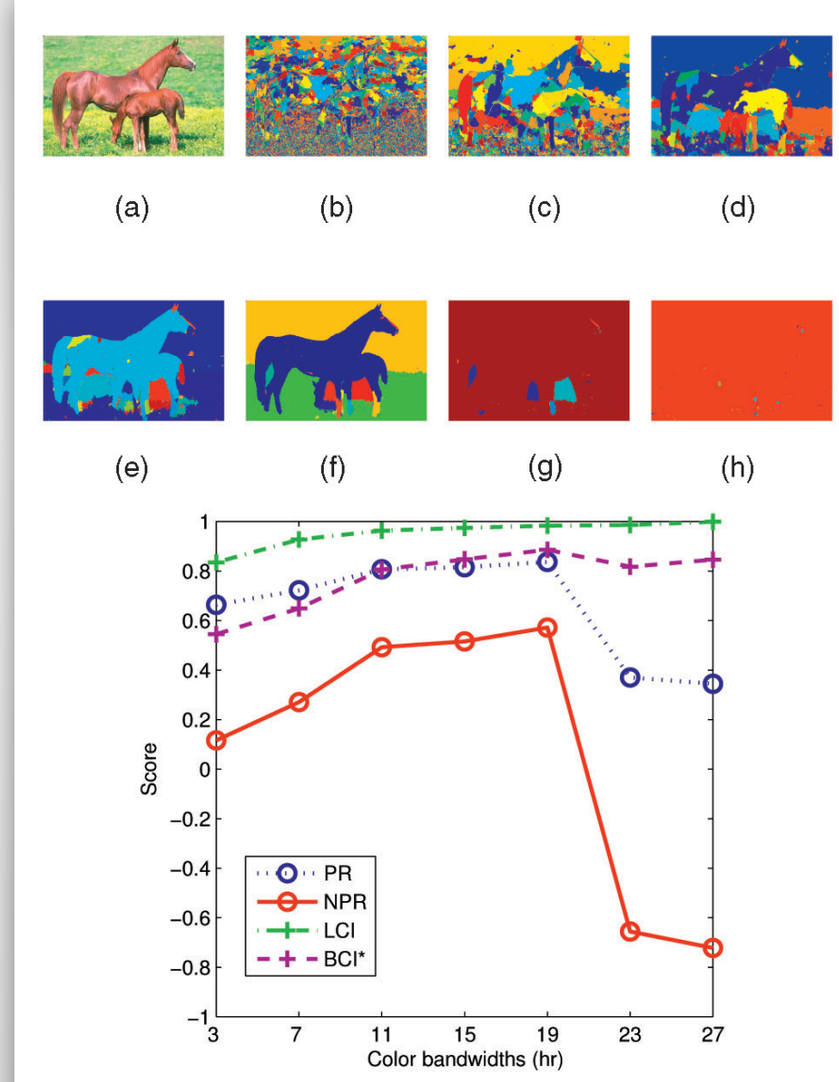
Axiom 2

m(P,Q) ~ 1 If *P* and *Q* are *very similar*



Axiom 3

m(P,Q) << 1 If *P* and *Q* are *very different*

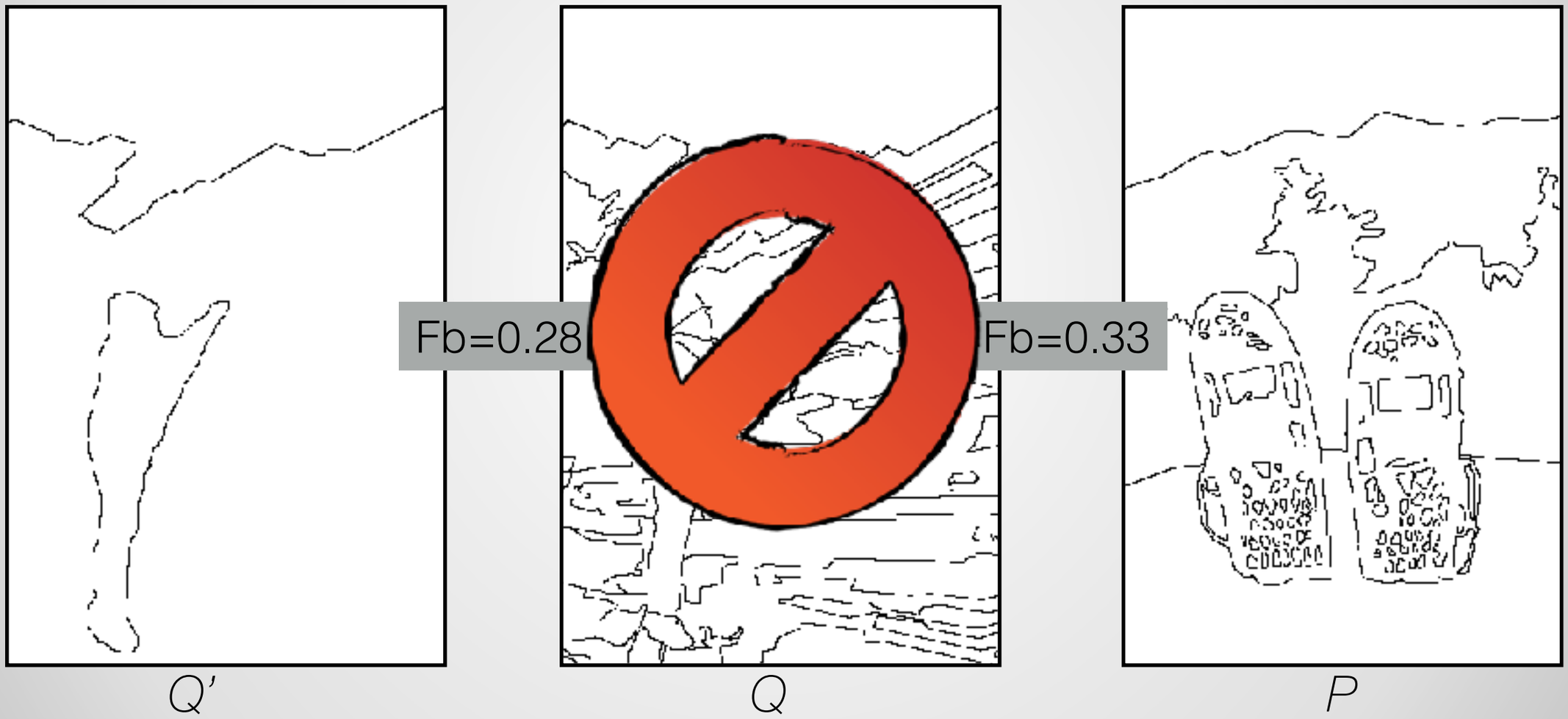


R. Unnikrishnan, **C. Pantofaru**, and **M. Hebert. Toward Objective Evaluation of Image Segmentation Algorithms**, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 29, No. 6, June, 2007, pp. 929-944.

Meta Measures

How to Quantitatively Compare Evaluation Measures

Axiom 4 $m(P,Q) < m(Q,Q')$ If Q and Q' should be more similar that P and Q



Swapped-Image Human Discrimination (SIHD)

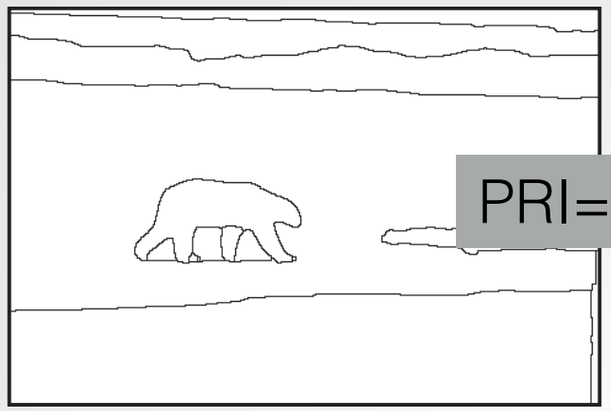
David Martin, An empirical approach to grouping and segmentation,
Ph.D. thesis, EECS Department, University of California, Berkeley, Aug 2003.

Meta Measures

How to Quantitatively Compare Evaluation Measures

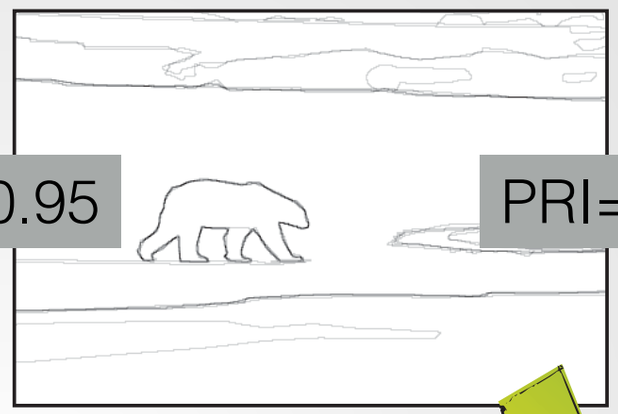
Axiom 4 $m(P,Q) < m(Q,Q')$ If Q and Q' should be more similar that P and Q

Image



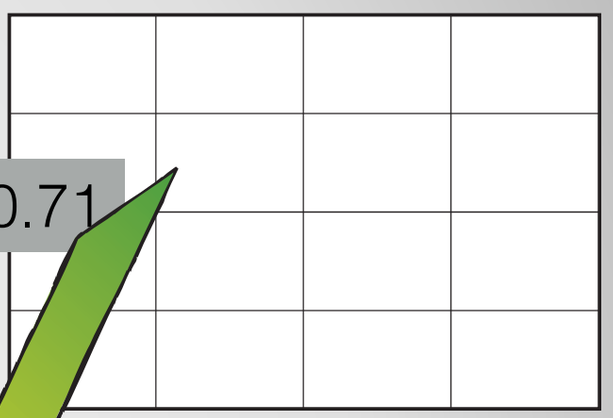
gPb-UCM result

PRI=0.95



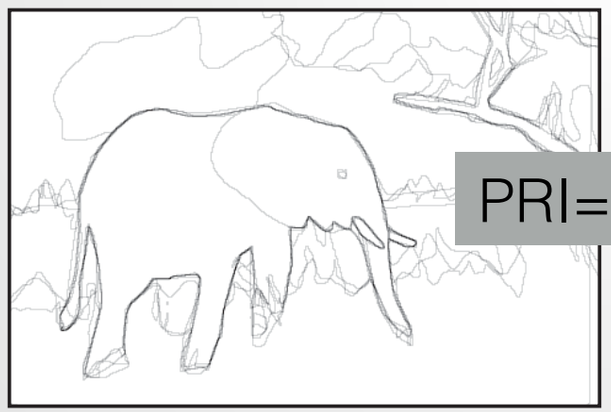
Ground Truth

PRI=0.71



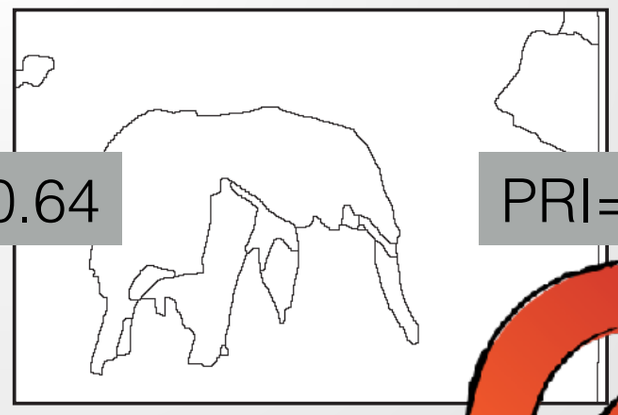
Baseline result

Image



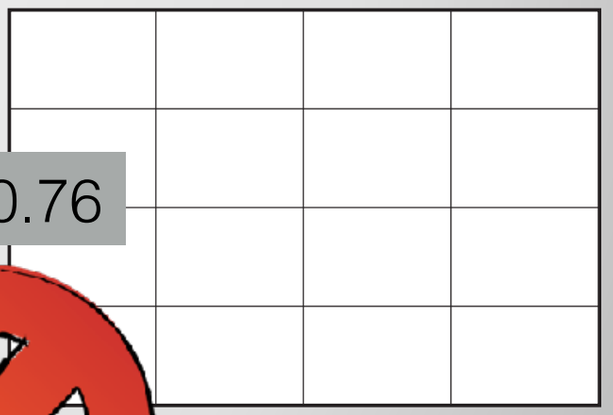
Ground Truth

PRI=0.64



gPb-UCM result

PRI=0.76



Baseline result

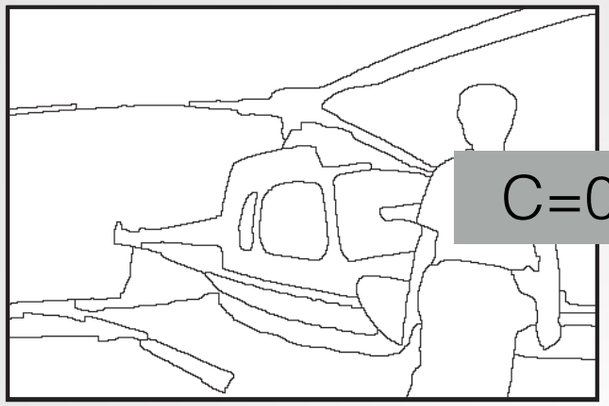
State-of-the-Art Baseline Discrimination (SABD)

Meta Measures

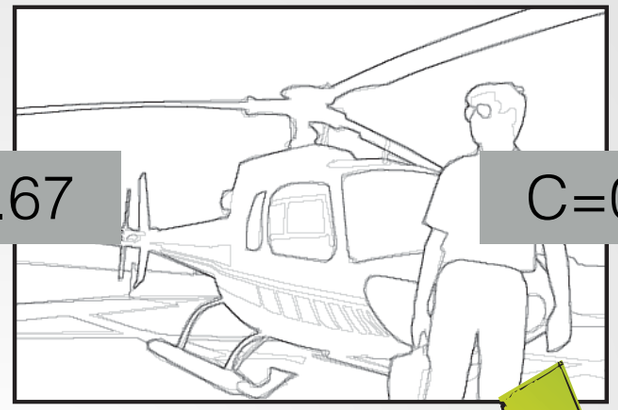
How to Quantitatively Compare Evaluation Measures

Axiom 4 $m(P,Q) < m(Q,Q')$ If Q and Q' should be more similar that P and Q

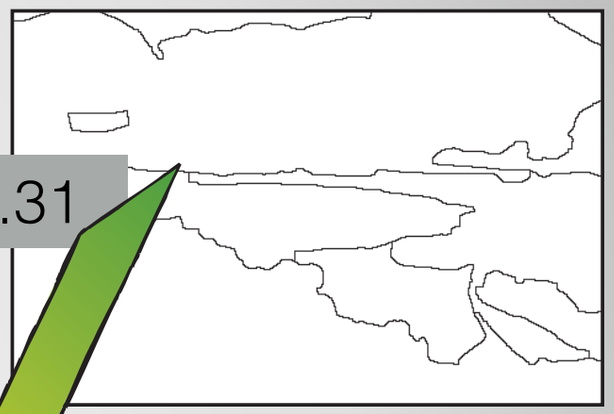
Image



gPb-UCM result

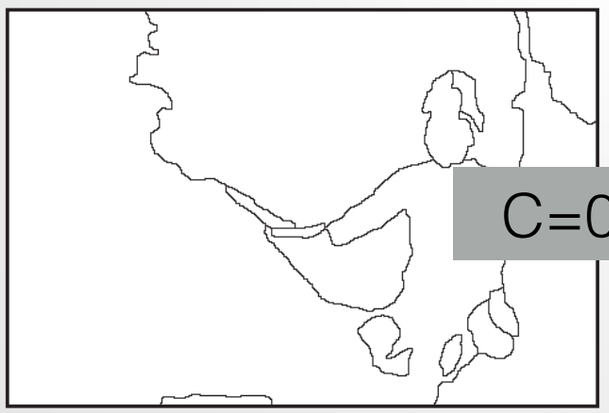


Ground Truth

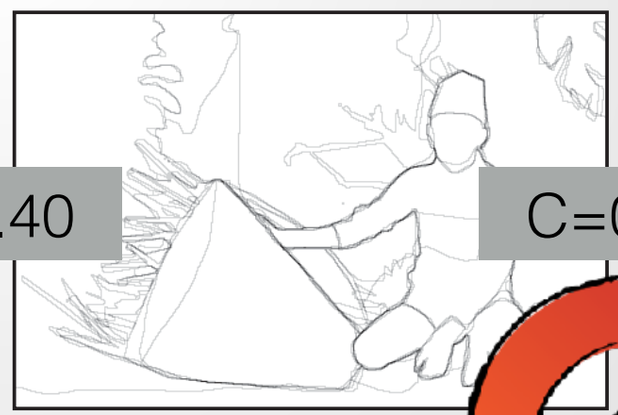


UCM swapped result

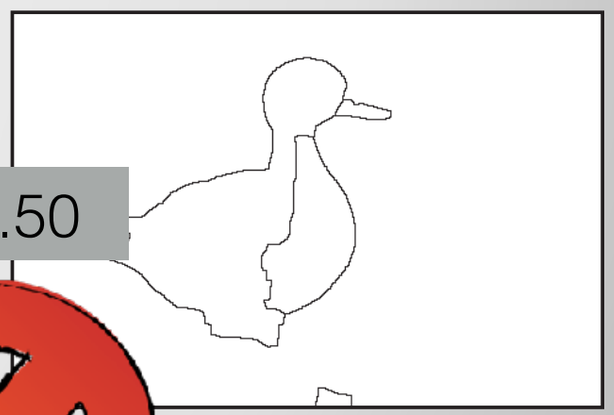
Image



gPb-UCM result



Ground Truth



CM swapped result

Swapped Image State-of-the-art Discrimination (SISD)

Meta Measures

How to Quantitatively Compare Evaluation Measures

Results

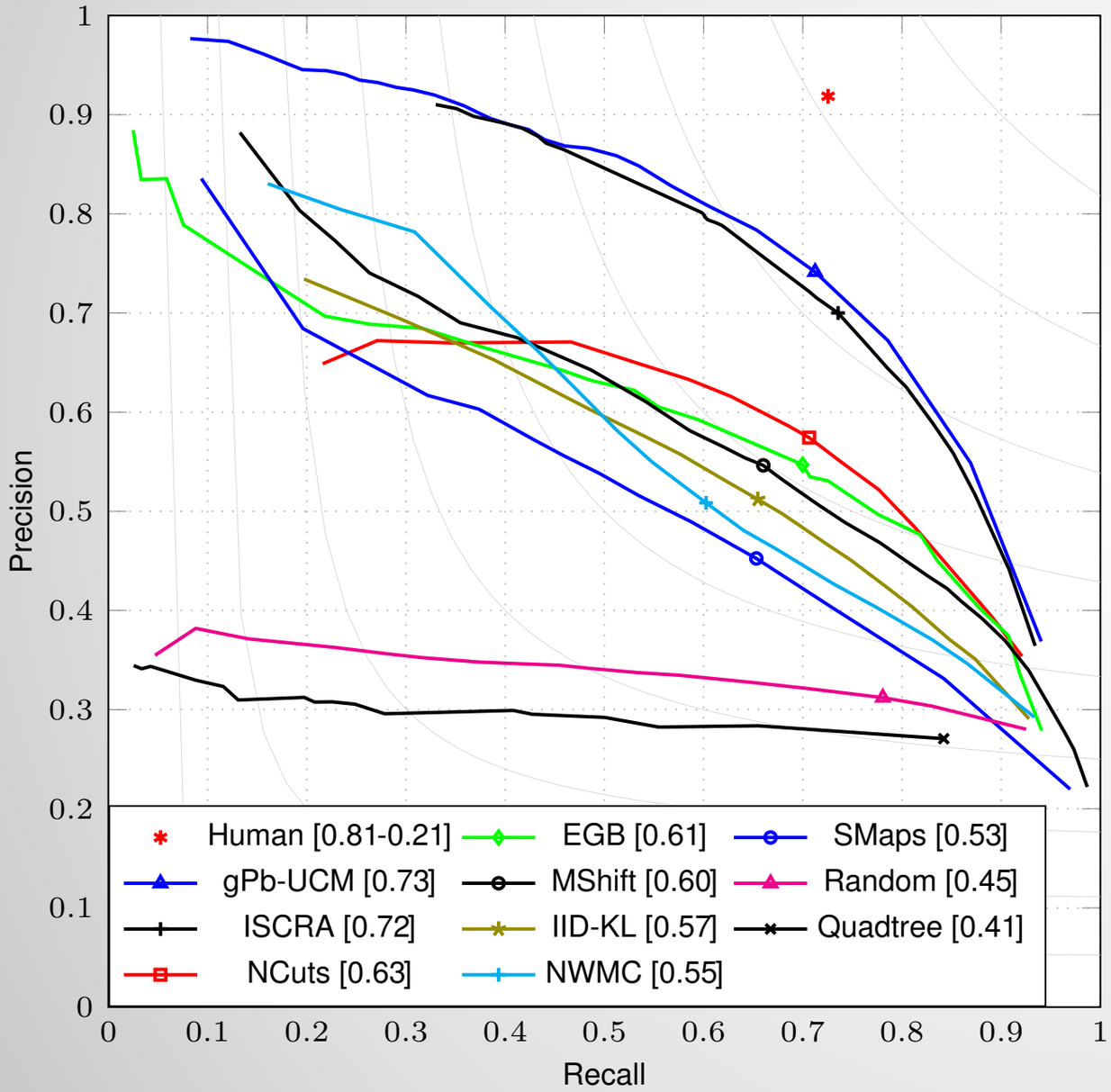
Measure	Quant. Meta-Meas. SIHD	SABD	SISD	Global
F_b	**99.5**	92.9	**99.9**	**97.4**
F_{op}	98.4	**94.2**	97.7	96.8
NVI	96.7	81.4	96.5	91.5
$\mathcal{C}(S \to \{G_i\})$	92.7	84.2	95.1	90.7
BCE	93.1	77.9	95.2	88.8
PRI	78.8	88.7	93.9	87.1
d_{vD}	95.0	74.6	90.5	86.7
BGM	90.2	77.4	92.5	86.7
$D_H(S \Rightarrow \{G_i\})$	78.1	81.7	99.1	86.3
F_r	89.3	74.6	92.6	85.5
$\mathcal{C}(\{G_i\} \to S)$	91.4	69.8	90.0	83.7
$D_H(\{G_i\} \Rightarrow S)$	73.8	56.4	77.7	69.3

Meta Measures

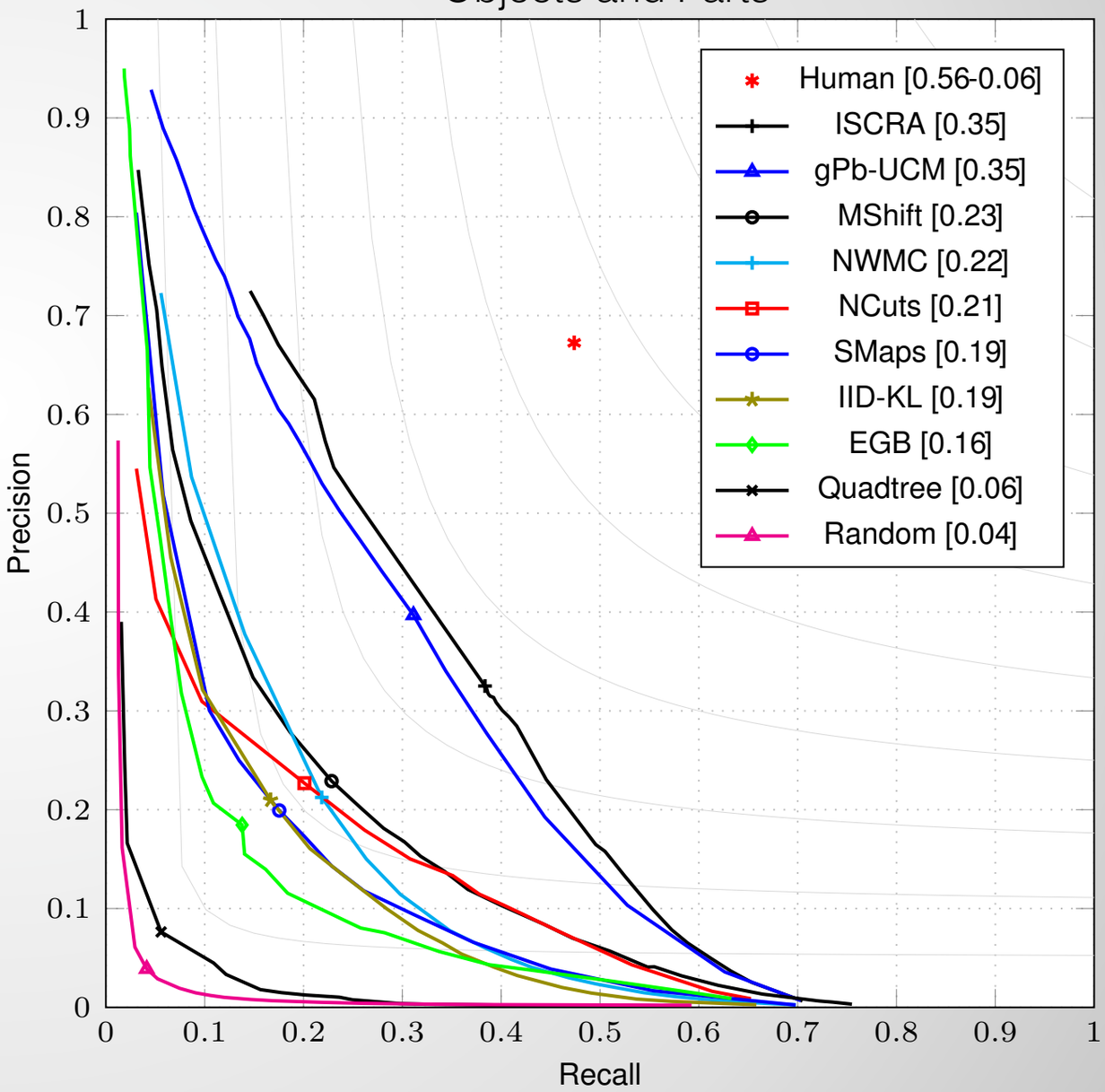
How to Quantitatively Compare Evaluation Measures

Results

Boundaries



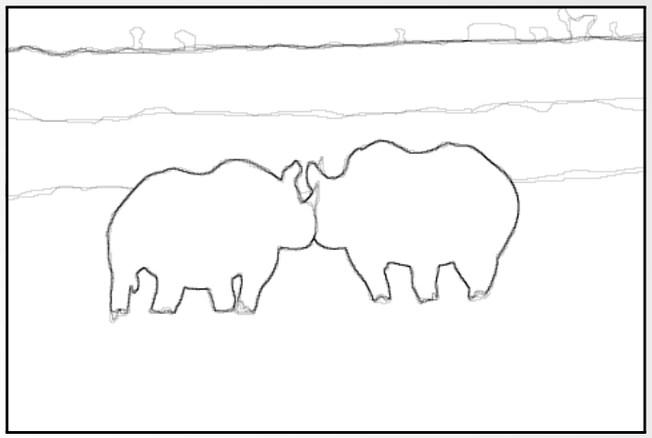
Objects and Parts



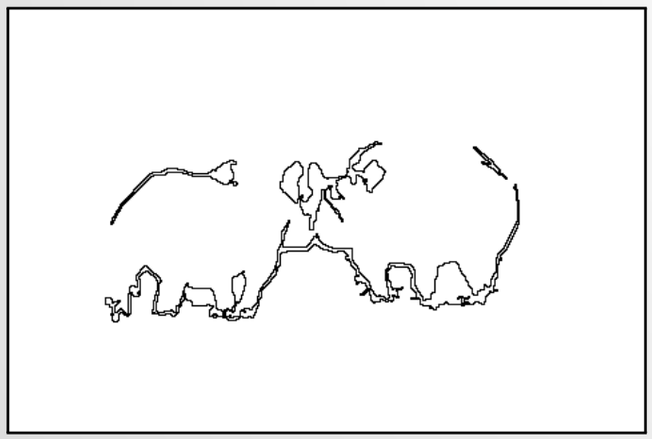
Meta Measures

How to Quantitatively Compare Evaluation Measures

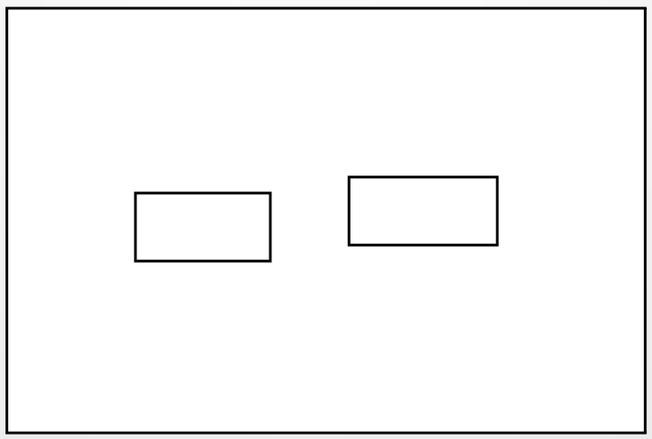
Image



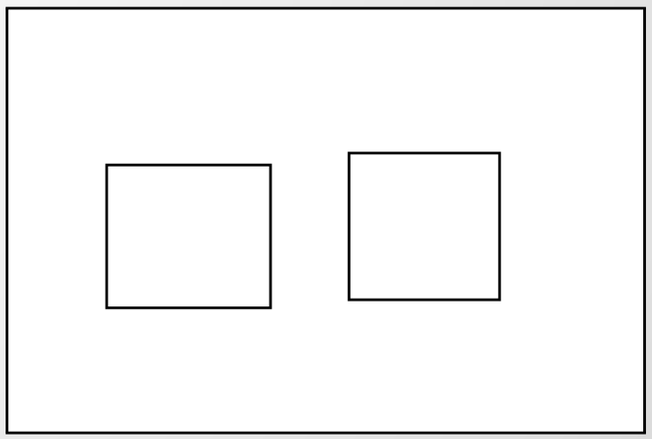
Ground Truth



(a) $F_b = 0.62$
$F_{op} = 0.05$



(b) $F_b = 0.01$
$F_{op} = 0.25$



(c) $F_b = 0.28$
$F_{op} = 0.04$

Meta Measures

How to Quantitatively Compare Evaluation Measures

Beyond Image Segmentation

1. The ranking of an evaluation measure should agree with the preferences of an application that uses the map as input.

2. A measure should prefer a good result by an algorithm that considers the content of the image, over an arbitrary map [22].

3. The score of a map should decrease when using a wrong ground-truth map [22].

4. The ranking of an evaluation measure should not be sensitive to inaccuracies in the manually marked boundaries in the ground-truth maps.

Ran Margolin, Lihi Zelnik-Manor, Ayellet Tal **How to evaluate foreground maps?** CVPR 2014

Meta Measures

How to Quantitatively Compare Evaluation Measures

Take-Home Messages

Deduplicate and **structure** measures in the literature
Are there possible *gaps* to fill?

Meta-measures: Set axioms and check them

- Multiple annotations of the ground truth
- Discrimination of **baseline results**
- Discrimination of **swapped results**

Check results **qualitatively** and **quantitatively**
on a wide range of **state-of-the-art results** and **baselines**

J. Pont-Tuset and F. Marques
Measures and Meta-Measures for the Supervised Evaluation of Image Segmentation
Computer Vision and Pattern Recognition (CVPR), 2013

J. Pont-Tuset and F. Marques
Supervised Evaluation of Image Segmentation and Object Proposal Techniques
IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), vol. 38, no. 7, pp. 1465-1478, 2016.

THANK YOU

Jordi Pont-Tuset

ECCV Workshop on Datasets and Performance Analysis in Early Vision
Amsterdam, 8th October 2016

@jponttuset

http://jponttuset.github.io

ETH zürich

CVL Computer Vision Lab