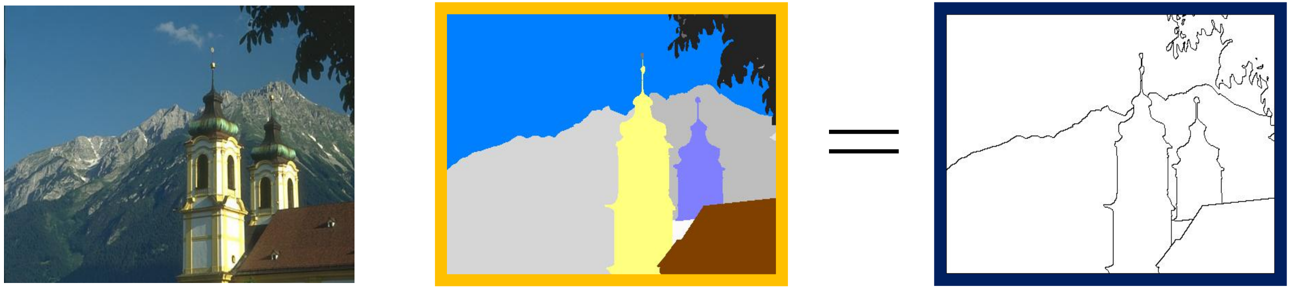
Evaluation of Segmentation Quality via Adaptive Composition of Reference Segmentations

Bo Peng[1], Lei Zhang[2], Xuanqin Mou[3], and Ming-Hsuan Yang[4]

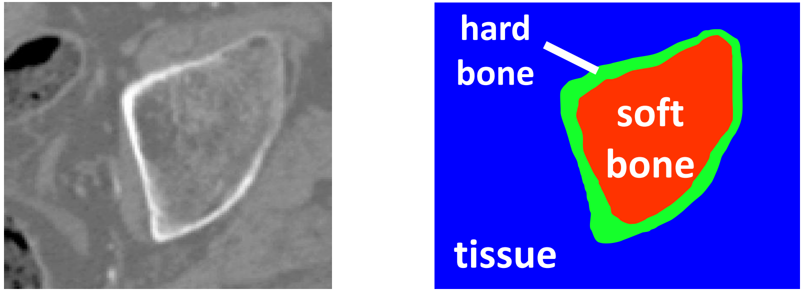
[1]Southwest Jiaotong University, [2]Hong Kong Polytechnic University,
[3]Xi'an Jiaotong University, [4]University of California at Merced

Introduction

- ➢ What is image segmentation?



Extract “regions” or “boundaries”

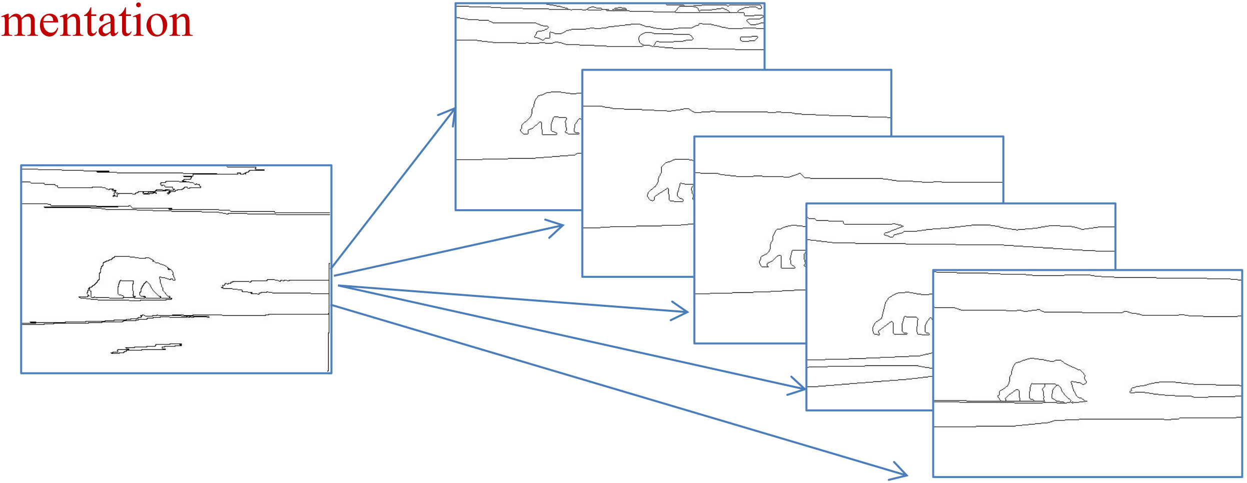


labeling

Introduction

- Evaluation of image segmentation quality

Reference-based segmentation evaluation



Machine segmentation

Hand-labeled segmentation (ground truth)

Introduction

- Applications
  - Performance evaluation of segmentation algorithms.
  - Proper parameter values can be determined based on reliable quantitative evaluation of image segmentation.

Related work

- Variation of Information metric (VOI)

It measures the distance between two segmentations in terms of their average conditional entropy.

$$VOI(S_1, S_2) = H(S_1 \mid S_2) + H(S_2 \mid S_1) - 2I(S_1, S_2)$$

- Segmentation Covering (SC)

It measures the similarity between segmentations by weight averaging the overlaps of regions in two segmentations.

$$C(S_1 \rightarrow S_2) = \frac{1}{N} \sum_{R \in S_1} |R| \cdot \max_{R' \in S_2} \frac{|R \cap R'|}{|R \cup R'|}$$

1. M. Meila. Comparing clusterings: an axiomatic view. In International Conference on Machine Learning, pages 577-584, 2005
2. P. Arbelaez, M. Maire, C. C. Fowlkes, and J. Malik. Contour detection and hierarchical image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 33(5):898–916, 2011

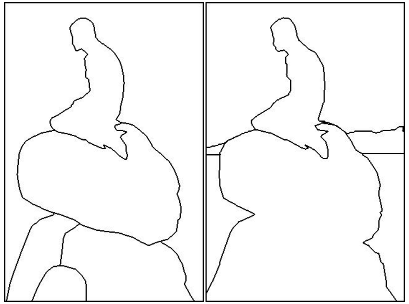
Related work

- Global Consistency Error (GCE)

It measure to which degree the segmentations S_1 and S_2 agree with each other.

$$E(S_1, S_2, p_i) = \frac{|R(S_1, p_i) \setminus R(S_2, p_i)|}{|R(S_1, p_i)|}$$

$$GCE(S_1, S_2) = \frac{1}{N} \min\{\sum_i E(S_1, S_2, p_i), \sum_i E(S_2, S_1, p_i)\}$$



- F-measure

A combination of precision and recall leads to the F-measure.

$$F = \frac{PR}{\tau R + (1-\tau)P}$$

D. Martin. An Empirical Approach to Grouping and Segmentation. PhD thesis, EECS Department, University of California, Berkeley, 2002

西南交通大學
Southwest Jiaotong University

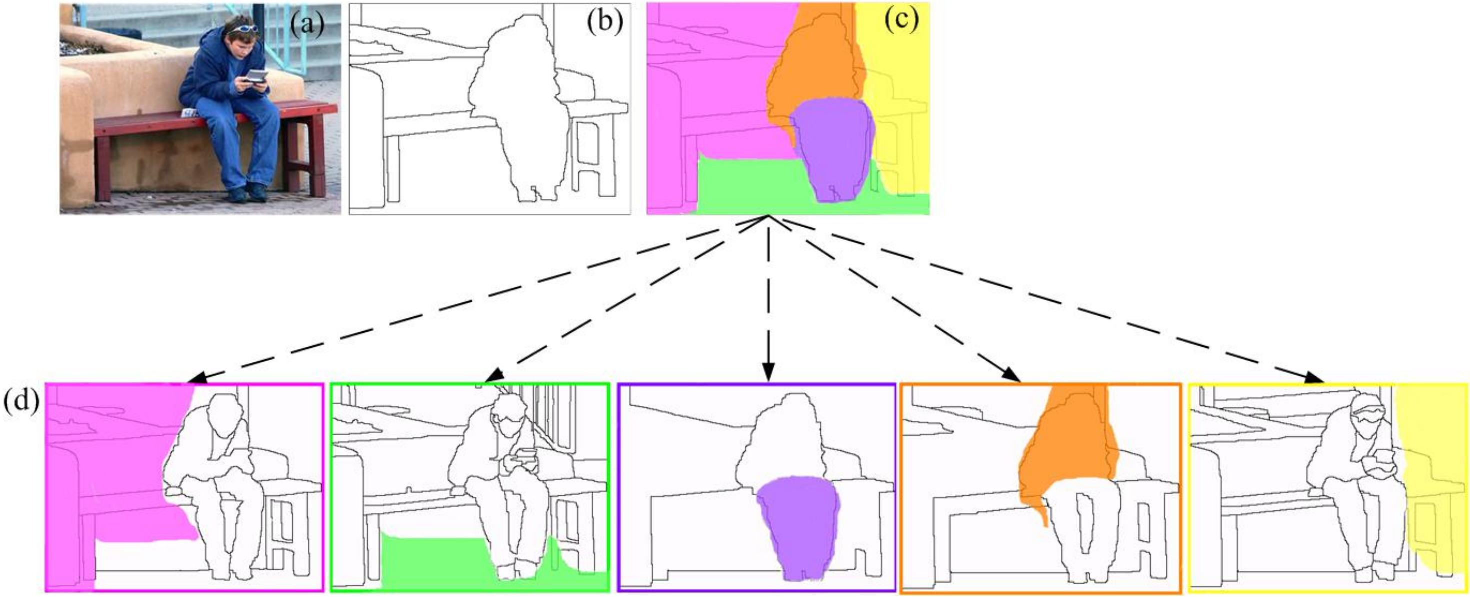
Related work

➢ Global comparison strategy

Elements (e.g. pixels) from one segmentation are fully compared with those of another segmentation (i.e. the ground truth).

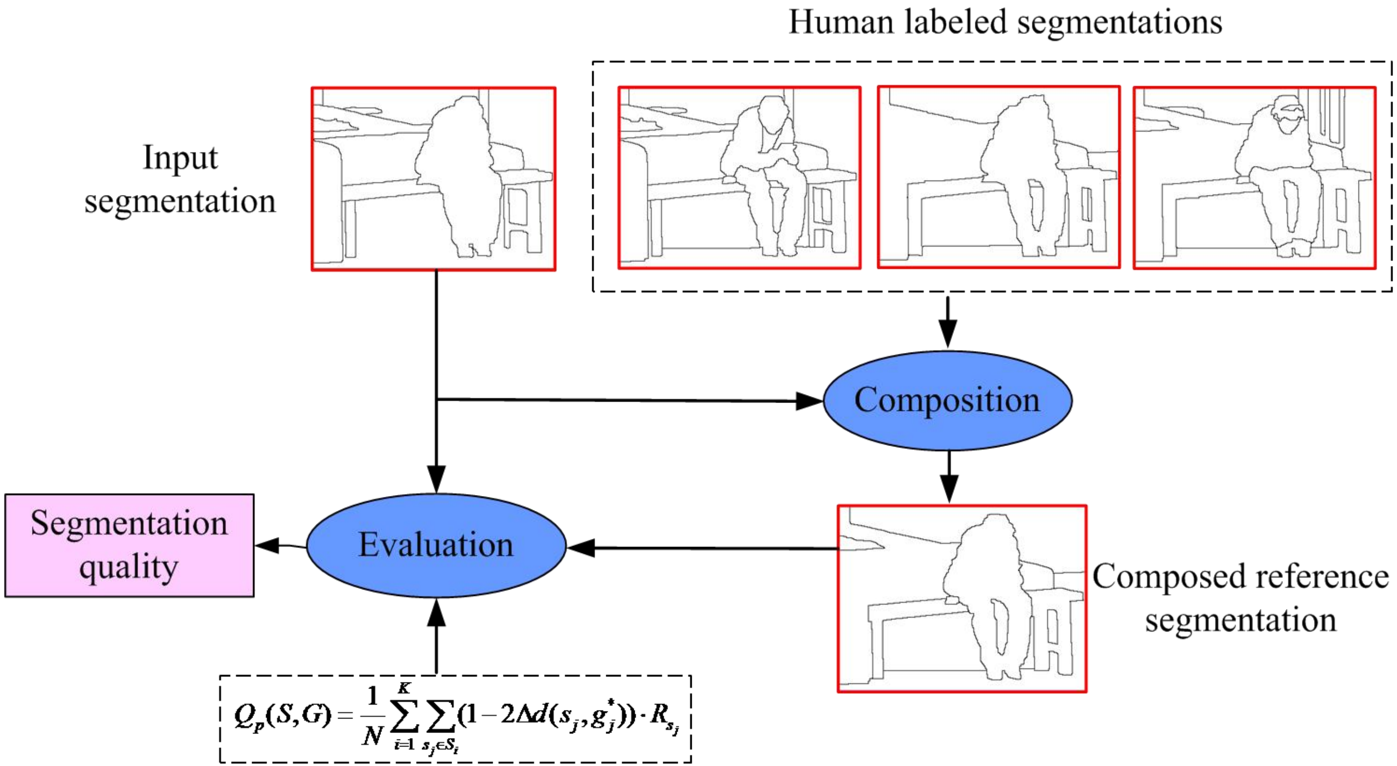
- ➢ The human visual system (HVS): highly adapted to extract structural information from natural scenes.
- ➢ Human observers may pay different attentions to different parts of the images.
- ➢ Ground truths of the same image therefore present various granularities in the object parts. This fact makes them rarely identical in the global view, while highly **consistent** in the local structures.

Motivation



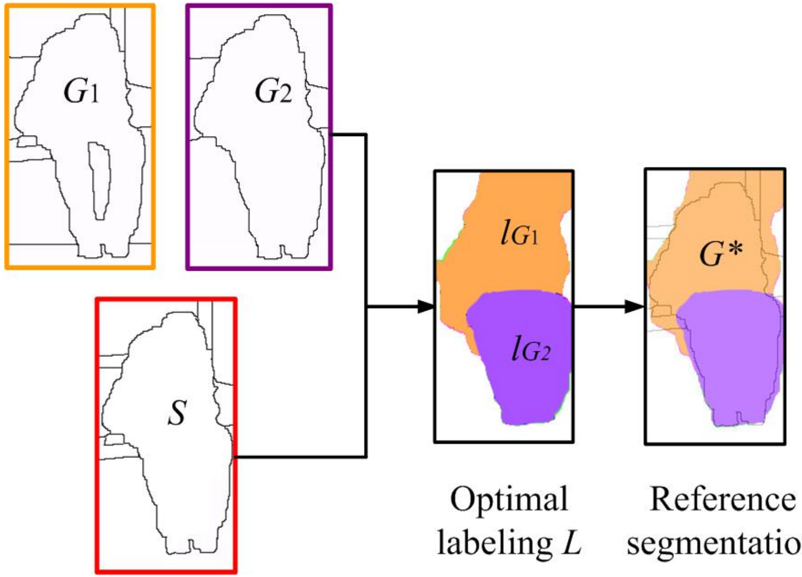
An illustrative example between a machine segmentation and labeled segmentations by humans.

Proposed evaluation framework



$$Q_p(S,G)=\frac{1}{N}\sum_{i=1}^{K}\sum_{s_j\in S_i}(1-2\Delta d(s_j,g_j^*))\cdot R_{s_j}$$

Composing Reference Segmentations



Optimal labeling L

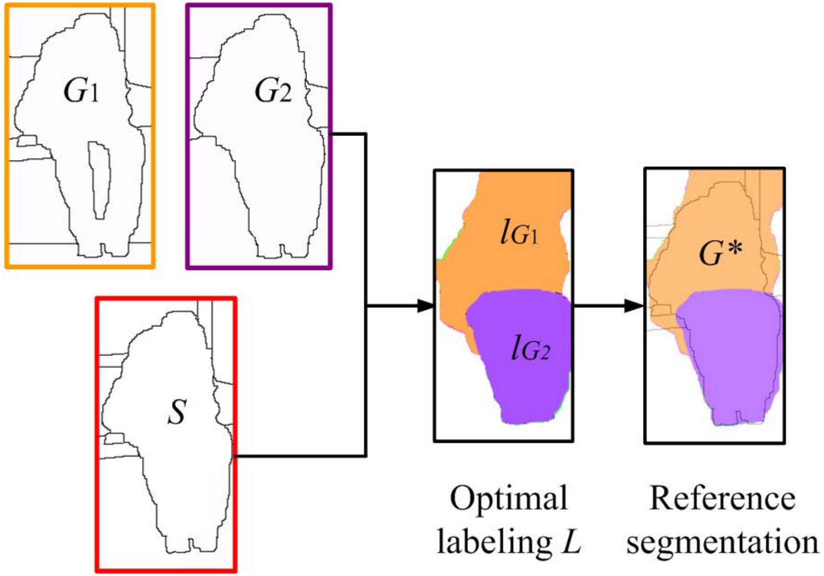
Reference segmentation

Seek for the labeling that minimizes the energy:

$$E(l) = \sum_i D(l_{gj}) + \lambda \cdot \sum_{\{g_j, g_{j'}\} \in M} u_{\{g_j, g_{j'}\}} \cdot T(l_{g_j} \neq l_{g_{j'}})$$

We use l labels, where each label corresponds to one reference segmentation, to compose G^*.

Composing Reference Segmentations



Seek for the labeling *l* that minimizes the energy:

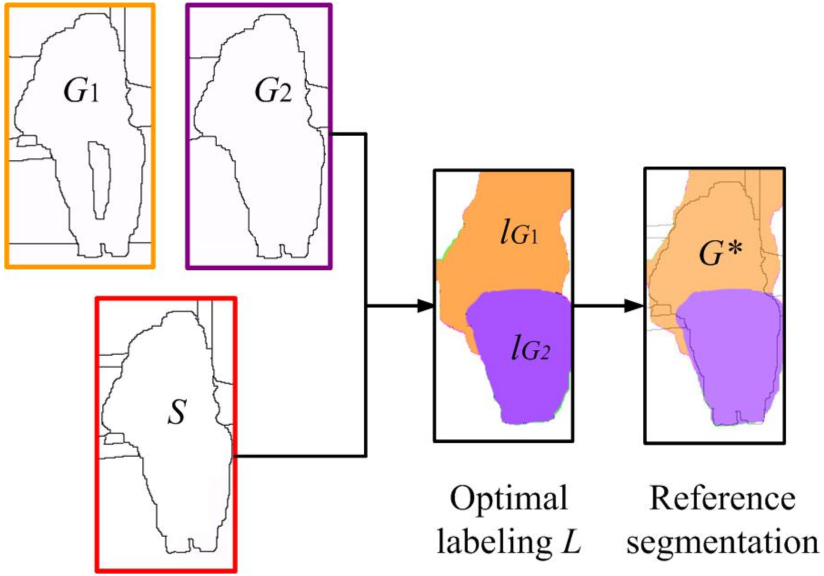
$$E(l)=\sum_i D(l_{gj})+\lambda\cdot\sum_{\{g_j,g_{j'}\}\in M} u_{\{g_j,g_{j'}\}}\cdot T(l_{g_j}\neq l_{g_{j'}})$$

$$D(l_{g_j})=\boxed{\Delta d(s_j,g_j)}$$

$$T(l_{g_j}\neq l_{g_{j'}})=\begin{cases}1 \text{ if } l_{g_j}\neq l_{g_{j'}}\\ 0 \text{ otherwise}\end{cases}$$

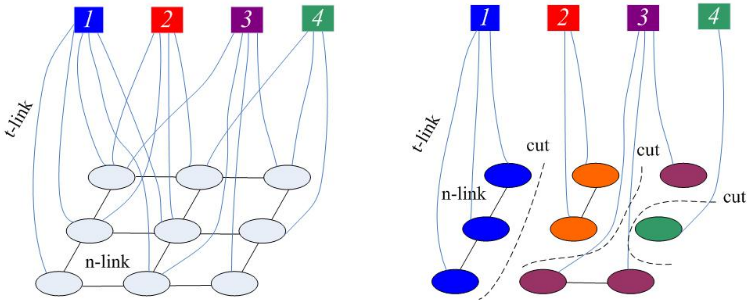
$$u_{\{g_j,g_{j'}\}}=\min\{\overline{\Delta d_j},\overline{\Delta d_{j'}}\}$$

Composing Reference Segmentations



Seek for the labeling that minimizes the energy.

$$E(l)=\sum_{i} D(l_{gj})+\lambda\cdot \sum_{\{g_j,g_{j'}\}\in M} u_{\{g_j,g_{j'}\}}\cdot T(l_{g_j}\neq l_{g_{j'}})$$

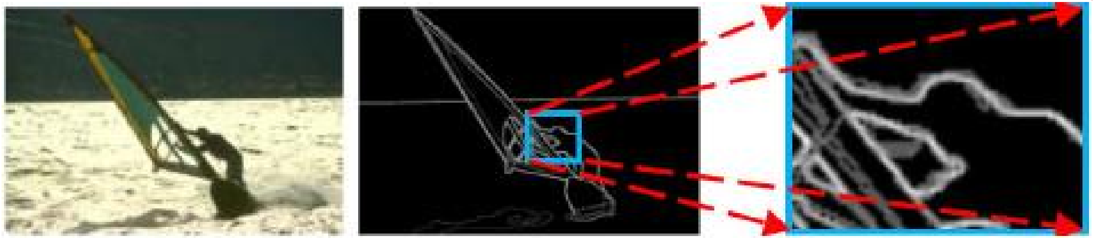


Multi-label graph cuts [Y.Boykov et al. 2001]

Composing Reference Segmentations

$$D(l_{g_j}) = \Delta d(s_j, g_j)$$

Localization errors from human labeling process



Structural similarity index: define a pixel-based distance, which uses the complex Gabor transform coefficients.

$$\Delta d(c_s, c_{s'}) = 1 - \overline{H}(c_s, c_{s'})$$

$$H(\mathbf{c}_x, \mathbf{c}_y) = \frac{2\sum_{i=1}^{N} |c_{x,i}||c_{y,i}^{*}| + \alpha}{\sum_{i=1}^{N} |c_{x,i}|^2 + \sum_{i=1}^{N} |c_{y,i}|^2 + \alpha} \cdot \frac{2|\sum_{i=1}^{N} c_{x,i} c_{y,i}^{*}| + \beta}{2\sum_{i=1}^{N} |c_{x,i} c_{y,i}^{*}| + \beta}$$

wavelet transform coefficients: CW-SSIM [M. Sampat, 2009]

Measuring Segmentation Quality

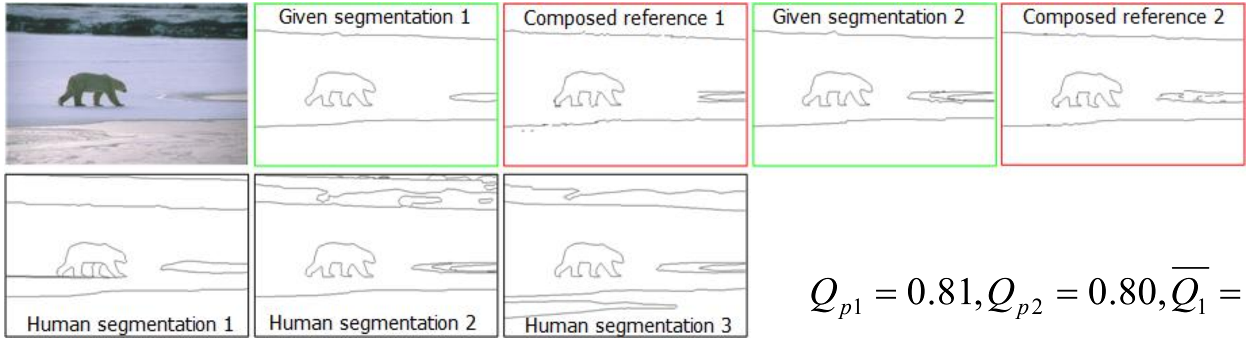
- Compute the similarity (or distance) between S and the reference G^*

$$\boxed{Q_p(S,\mathbf{G}) = \frac{1}{N}\sum_{i=1}^{K}\sum_{s_j \in S_i}(1-2\Delta d(s_j, g_j^*))\cdot R_{s_j}}$$

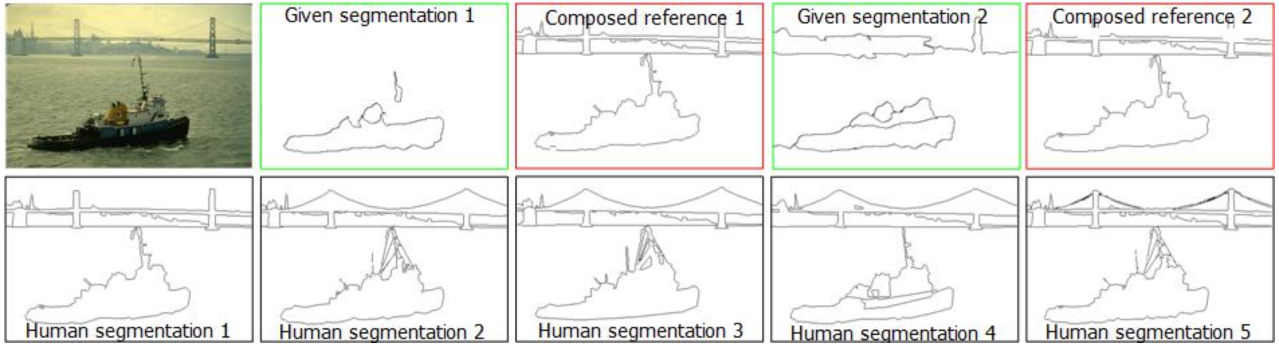
The similarity between S and the composite ground truth G^*: $\Delta d(s_j, g_j^*)$

The empirical global confidence of G^*: $R_{s_j} = 1 - \overline{\Delta d_j}$

Examples of composed references



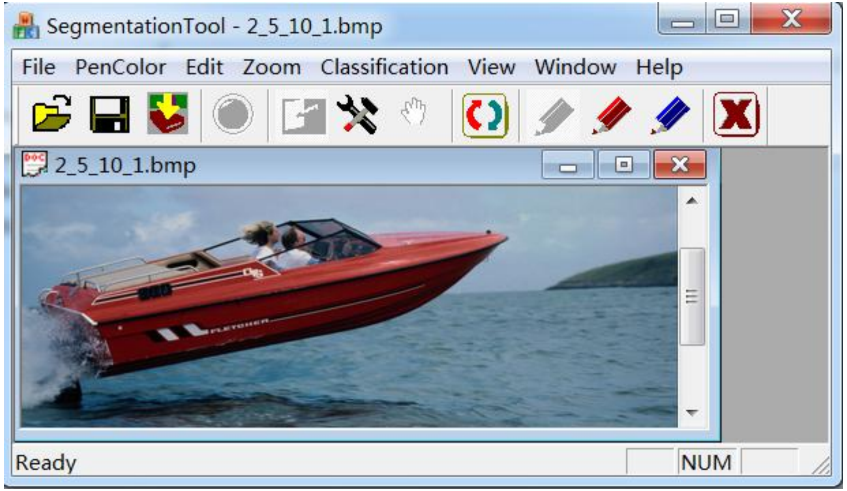
$$Q_{p1} = 0.81, Q_{p2} = 0.80, \overline{Q_1} = 0.73, \overline{Q_2} = 0.69$$



$$Q_{p1} = 0.52, Q_{p2} = 0.56, \overline{Q_1} = 0.49, \overline{Q_2} = 0.49$$

Datasets

- ➢ Image segmentation dataset

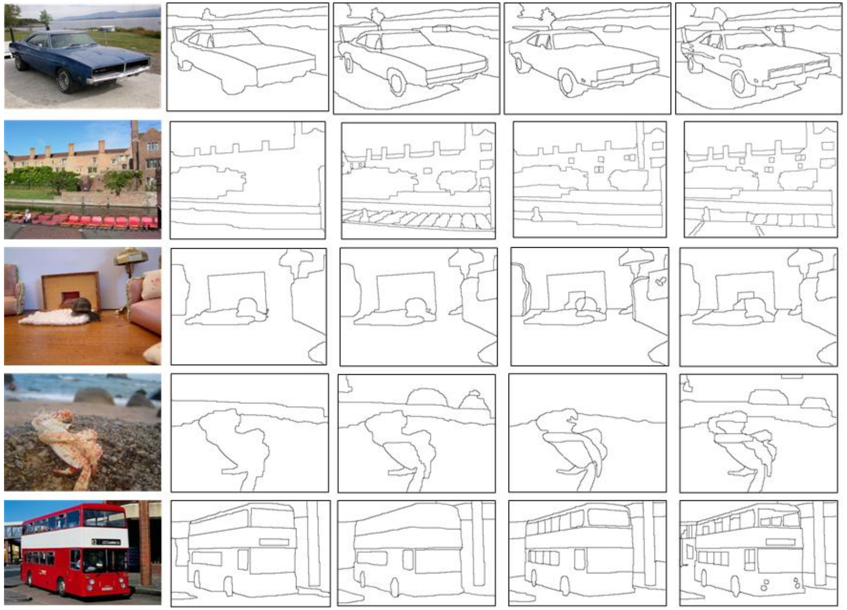

SegmentationTool - 2_5_10_1.bmp
File PenColor Edit Zoom Classification View Window Help
2_5_10_1.bmp
Ready
NUM

User interface of the developed image segmentation tool.

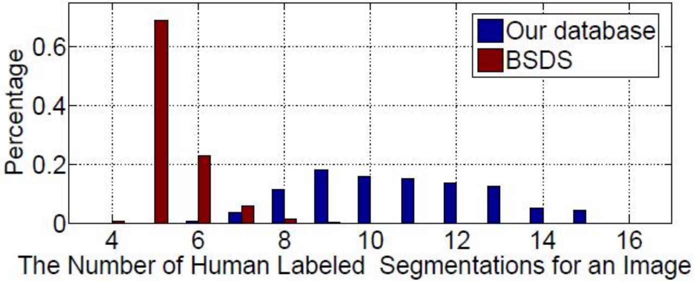
- ➢ Livewire

Datasets

- Image segmentation dataset



database	BDSD	Our Database
# images	500	200
# ground truths/image	4-9	6-15
Image type	Natural images	Natural images
Software supported	yes	yes
# subjects	30	45
Time/segmentation	5-30 min	2-4 min



Datasets

- Segmentation evaluation dataset
  - Compare the performance of a pair of segmentations based on a segmentation dataset with human labeled results.
  - Contains 500 pairs of segmentations and the corresponding evaluation results by human subjects.

Seg. Algorithms	Parameter values
EG	K = {600,800,1000,1400,1800}
MS	h_r ={7, 11, 15, 19, 23}, h_s=7, min_R = 150.
CTM	ε ={0.1 , 0.2, 0.3, 0.4, 0.5}
TBES	N_{sp} = 200, ε = {50, 100, 200,300, 400}

Datasets

- Segmentation evaluation dataset

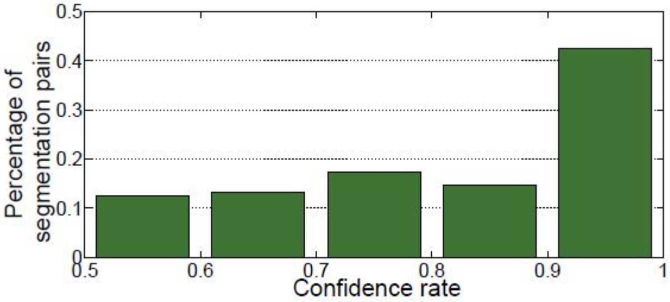
    - Compare the performance of a pair of segmentations based on a segmentation dataset with human labeled results.
    - Contains 500 pairs of segmentations and the corresponding evaluation results by human subjects.

Seg. pairs: with 10 human subjects, the the best 3 and the worst 3 segmentations randomly select one segmentation from the group of good/bad.

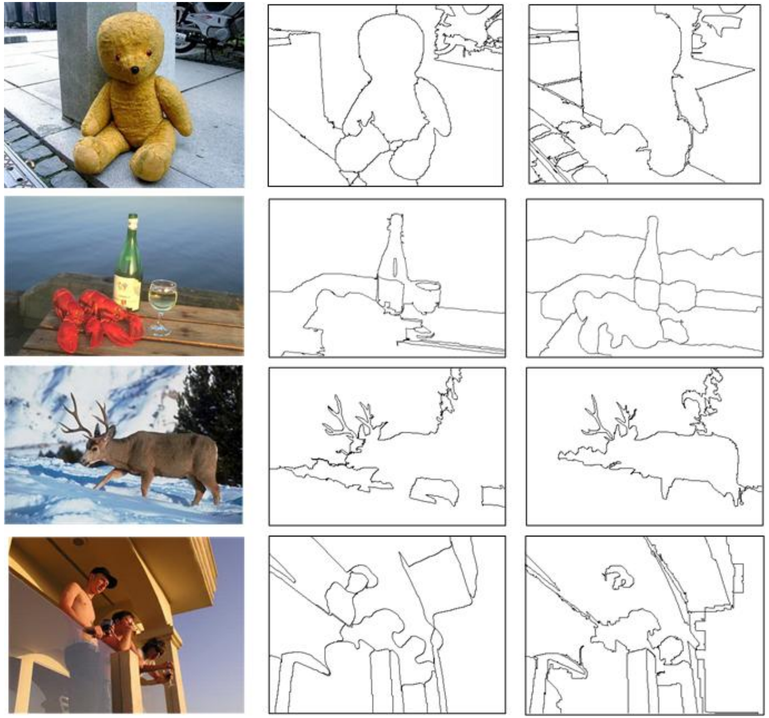
Subjective evaluation: 70 subjects with little or no research experience in image segmentation. 500 pairs of segmentations are evenly divided into 10 groups.

Datasets

- Segmentation evaluation dataset



Distribution of confidence rates on the proposed segmentation evaluation dataset.



Experimental Results

- Intel Core 2 Duo 3.00 GHz CPU and 4GB memory.
- The run time: 24.6 ± 6.0 seconds for composing the reference G^*
  10.7 ±1.1 seconds for computing the score Q_p.

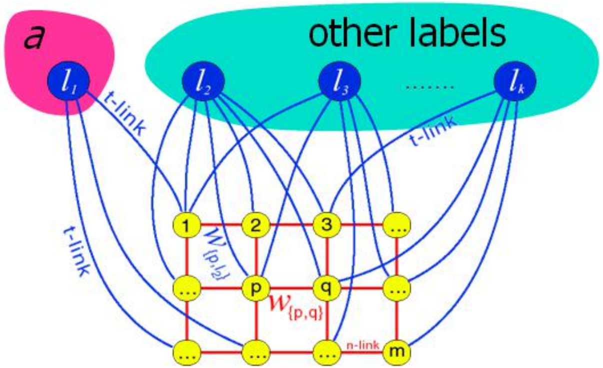
Experimental Results

- Sensitivity analysis

Test the effects of λ and initial labeling on the final evaluation score.

Alpha-expansion algorithm: break multi-way cut computation into a sequence of binary s-t cuts.

$$E(l)=\sum_{i} D(l_{gj})+\lambda\cdot\sum_{\{g_j,g_{j'}\}\in M} u_{\{g_j,g_{j'}\}}\cdot T(l_{g_j}\neq l_{g_{j'}})$$

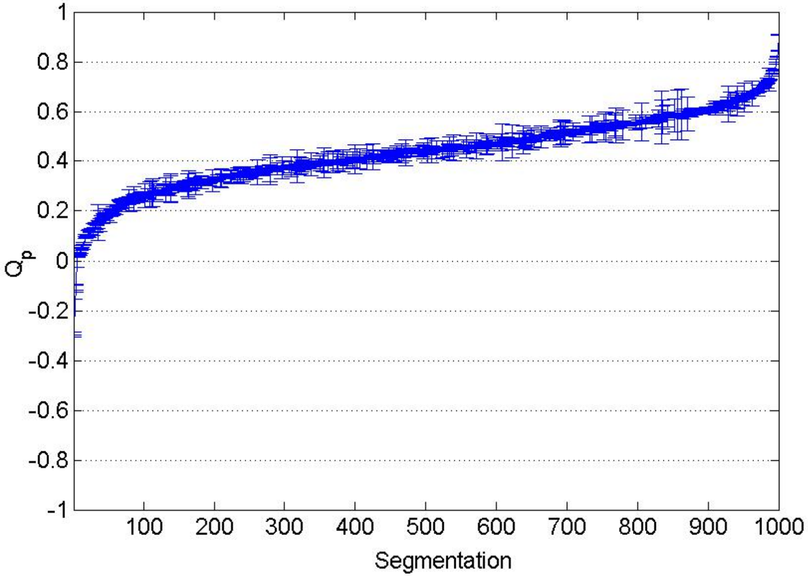


Experimental Results

- Sensitivity analysis

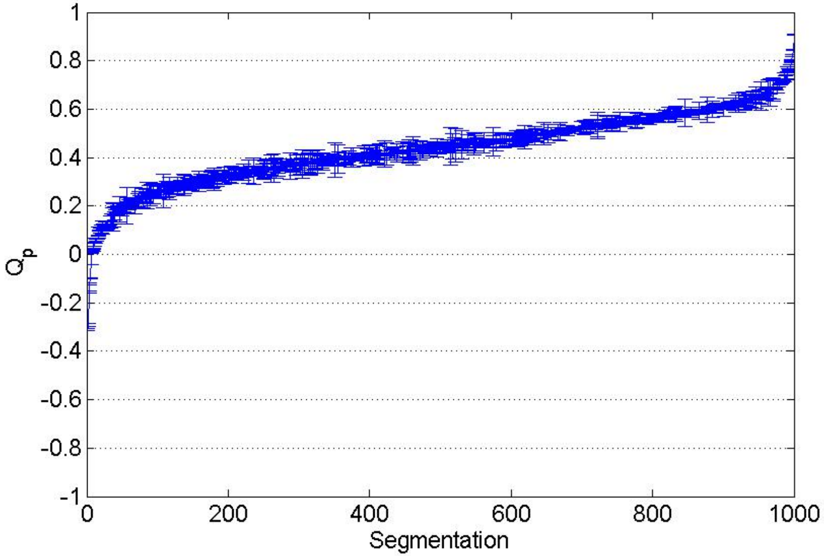
Fix λ to be [500,1200], with an interval of 50.

The initial labeling of graph cut is set randomly, then the mean values and standard deviations of Q_p



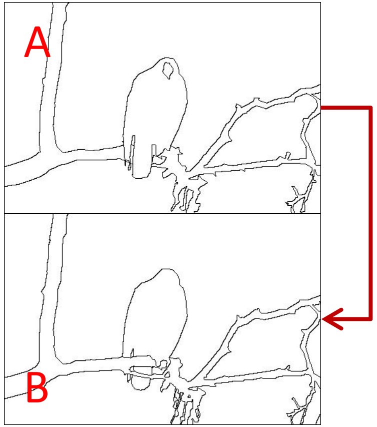
Experimental Results

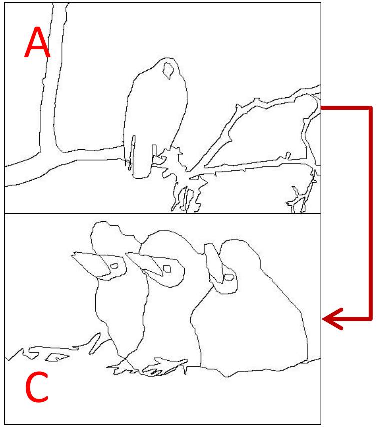
- Sensitivity analysis
  - Fix λ to be 800.
  - Carry out the proposed algorithm 50 times with random initialization of labeling.



Evaluation with Meta-Measure

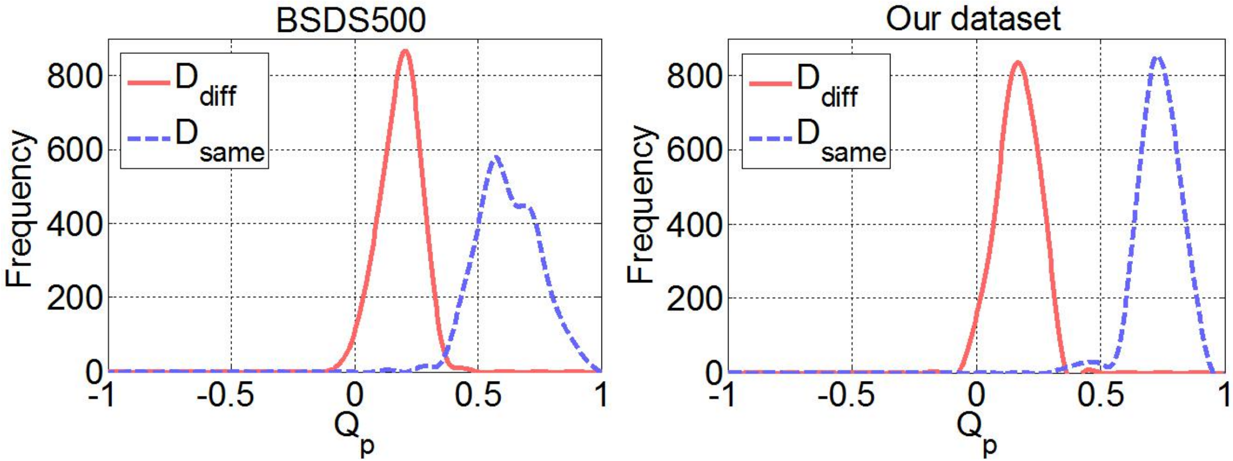
- The meta-measure
  - human labeled segmentation vs. human labeled segmentation of the same image
  - human labeled segmentation vs. machine segmentations of a different image





Evaluation with Meta-Measure

- The meta-measure
  - human labeled segmentation vs. human labeled segmentation of the same image
  - human labeled segmentation vs. machine segmentations of a different image
  - the percentage of comparisons that agree with this principle as the meta-measure result

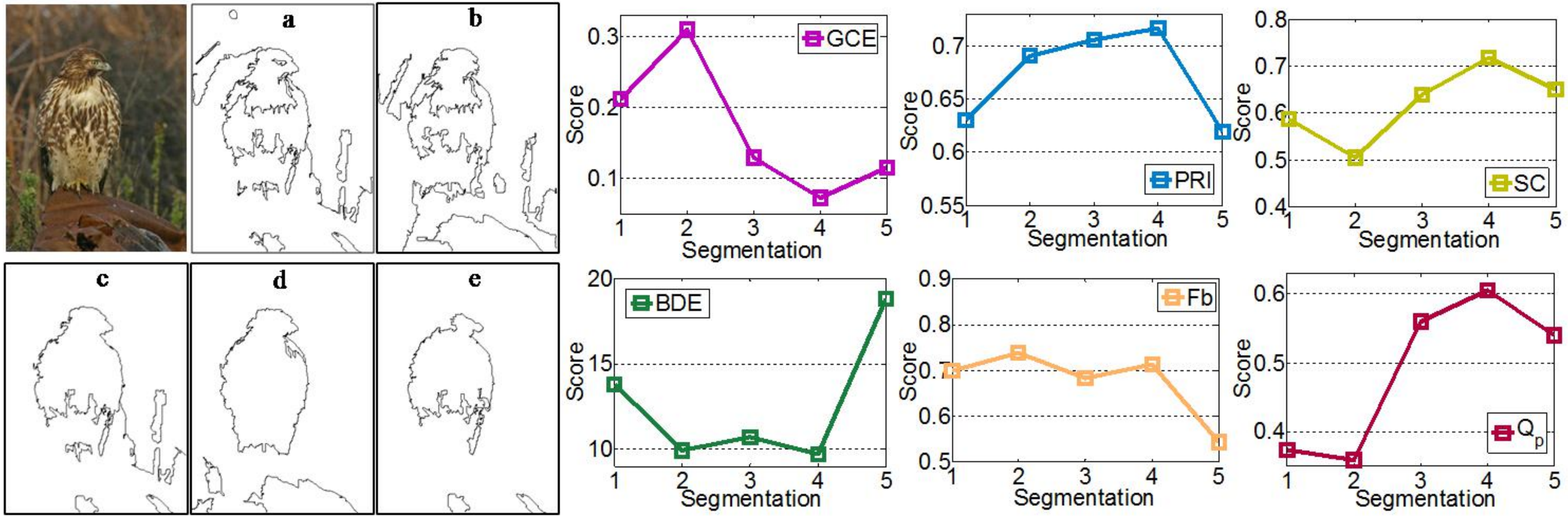


Evaluation with meta-measure

Evaluation results with the meta-measure on different measures

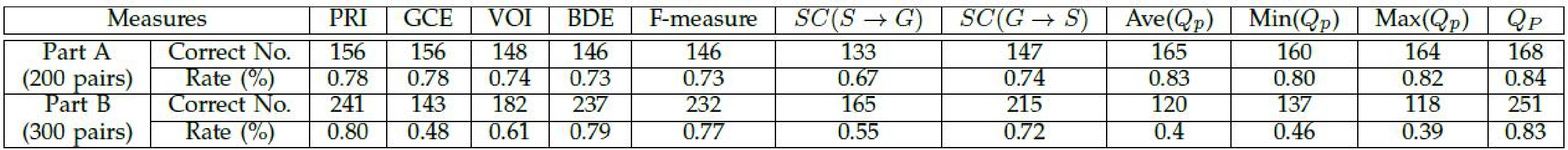
Measures	PRI	GCE	VOI	BDE	F-measure	SC(*S*->*G*)	SC(*G*->*S*)	Q_p
BSDS500	0.911	0.929	0.967	0.921	0.882	0.962	0.956	0.984
Proposed dataset	0.959	0.981	0.991	0.947	0.838	0.974	0.979	0.994

Evaluation with proposed segmentation dataset



Evaluation with proposed segmentation dataset

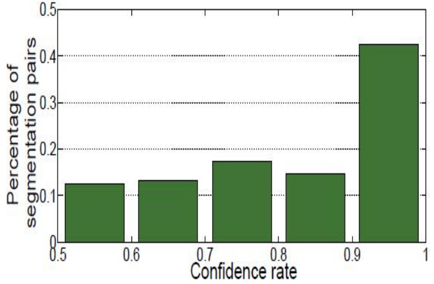
Evaluation results by different measures.

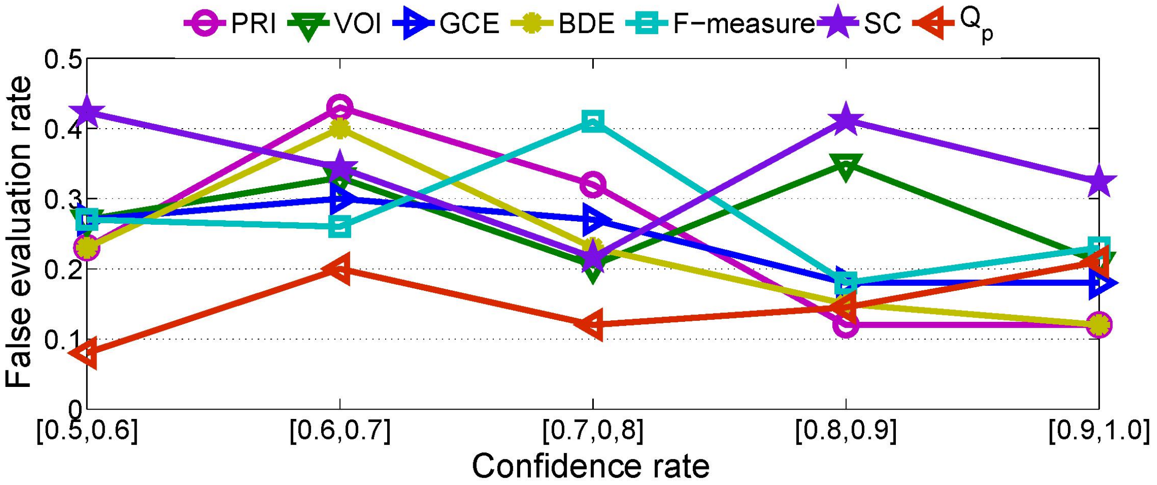


Measures		PRI	GCE	VOI	BDE	F-measure	$SC(S \to G)$	$SC(G \to S)$	Ave(Q_p)	Min(Q_p)	Max(Q_p)	Q_P
Part A (200 pairs)	Correct No.	156	156	148	146	146	133	147	165	160	164	168
	Rate (%)	0.78	0.78	0.74	0.73	0.73	0.67	0.74	0.83	0.80	0.82	0.84
Part B (300 pairs)	Correct No.	241	143	182	237	232	165	215	120	137	118	251
	Rate (%)	0.80	0.48	0.61	0.79	0.77	0.55	0.72	0.4	0.46	0.39	0.83

Evaluation with proposed segmentation dataset

The false evaluation rates with respect to the confidence rate of human subjects





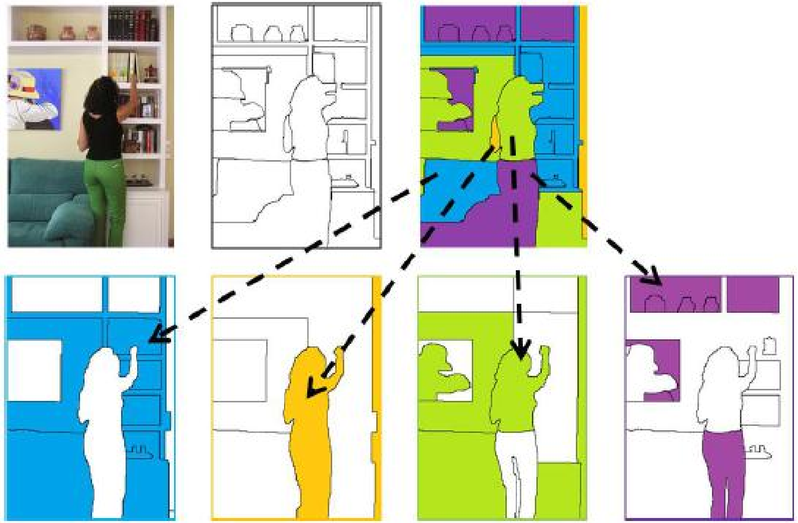
Further work

Composed exemplar reference image using region-based distance:

$$\Delta d(A,B)=1-\frac{M(A\cap B)}{M(A\cup B)}$$

Improved measure Q_{PRI}, Q_{GCE}, Q_{VOI}

$$Q=\sum_{R_j}\frac{N_{R_j}}{N}M_{R_j}$$



[1] Features of similarity.[A. Tversky. Psychological Review, 1977]

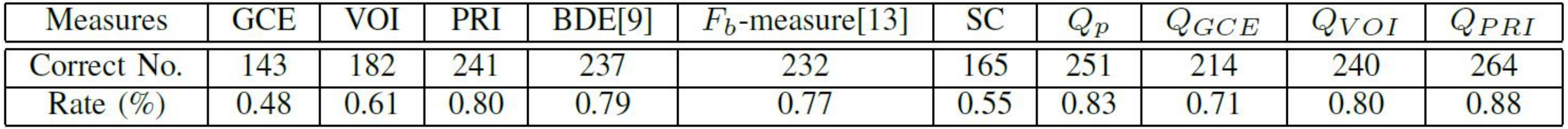
[2] Region based exemplar references for image segmentation evaluation. [B.Peng et al. SPL,2016]

Experimental Results

- Intel Core 2 Duo 3.00 GHz CPU and 4GB memory.
- The run time: 6:5 ±4.5 seconds for composing the reference G^*
- λ is set by line search within range [50, 500] for each input segmentation

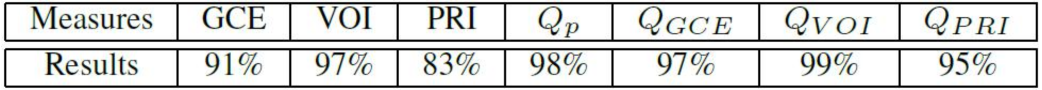
Experimental Results

Subjective evaluation resutls:



Measures	GCE	VOI	PRI	BDE[9]	F_b-measure[13]	SC	Q_p	Q_{GCE}	Q_{VOI}	Q_{PRI}
Correct No.	143	182	241	237	232	165	251	214	240	264
Rate (%)	0.48	0.61	0.80	0.79	0.77	0.55	0.83	0.71	0.80	0.88

Meta-measure resutls:



Measures	GCE	VOI	PRI	Q_p	Q_{GCE}	Q_{VOI}	Q_{PRI}
Results	91%	97%	83%	98%	97%	99%	95%

Conclusions

- Proposed a framework for evaluating segmentation quality with multiple human labeled segmentations.
- A reference segmentation was adaptively constructed.
- We presented a segmentation dataset and segmentation evaluation dataset to facilitate quantitative quality assessment.
- Extensive experiments demonstrate the effectiveness of our framework.