The Evolution of Evaluation for Image Segmentation

Yu-Jin ZHANG
Department of Electronic Engineering
Tsinghua University, Beijing 100084, CHINA

Outline

- **Introduction**
- **Brief Summary for Image Segmentation**
- **A General Framework for Segmentation and Its Evaluation**
- **Empirical Evaluation Criteria and Methods**
- **A Recent Progress Overview of Segmentation Evaluation**
- **Comparison of Various Evaluation Methods**
- **Concluding Remarks**

Introduction

- **Image Segmentation**
  - A process consists of subdividing an image into its constituent parts and extracting these parts of interest (objects) from the image
  - A critical process for computer vision
  - A focused research topic for image technique
  - There is no general theory for image segmentation, yet. So *ad hoc* techniques are often developed

Introduction

- **Three Levels of Research**

  Research works on image segmentation are currently conducted in three levels:

  (0) Base level:
  Segmentation algorithm development

  (1) Middle level:
  Evaluation of segmentation techniques

  (2) Top level:
  Systematic study and use of evaluation methods

Brief Summary for Image Segmentation

- **Definition of Image Segmentation**

[Fu 1981]

(1) $\bigcup_{i=1}^{n} R_i = R$;

(2) for all i and j, $i \neq j$, there exits $R_i \cap R_j = \varnothing$;

(3) for $i = 1, 2, \ldots, n$, it must have $P(R_i) = TRUE$;

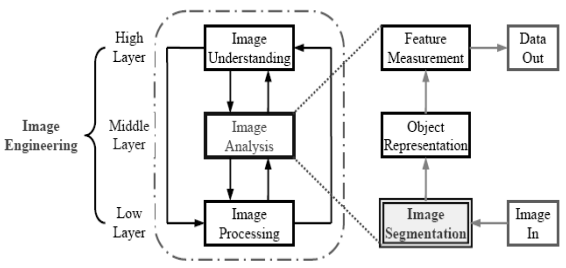
(4) for all $i \neq j$, there exits $P(R_i \cup R_j) = FALSE$;

(5) For all $i = 1, 2, \ldots, n$, R_i is a connected component.

Low Level Process — **High Level Process**

Brief Summary for Image Segmentation

- **Position of Image Segmentation**



Brief Summary for Image Segmentation

- **Number of Developed Algorithms (1)**
  - The history of segmentation of digital images using computers could be traced back to 40 years' ago [Roberts 1965]
  - Over the last 40 years, the research and development of segmentation techniques are going on steadily and have resulted a large number of developed algorithms
  - It is estimated 10 years' ago that the number of developed algorithm has attend 4 digits

Brief Summary for Image Segmentation

- **Number of Developed Algorithms (2)**

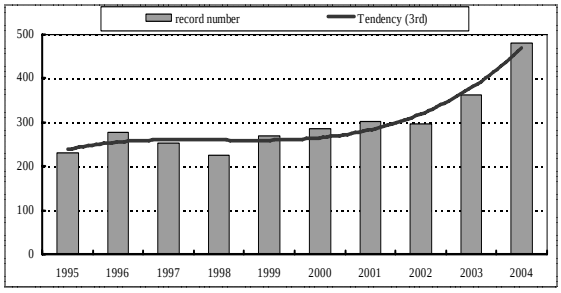
⇨ Search the number of records by using the term "image segmentation" only in the title field from "EI Compendex" gives the following results:

1965-1994	1995	1996	1997	1998	1999
965	232	278	253	226	268

2000	2001	2002	2003	2004	Total
287	303	297	364	481	4344

Brief Summary for Image Segmentation

- **Tendency of Development**



Brief Summary for Image Segmentation

- **Summary of Survey Papers (1)**

A number of survey papers for general image segmentation algorithms:

- **1975-1984**: [Davis, 1975]; [Zucker, 1976]; [Riseman, 1977]; [Zucker, 1977]; [Weszka, 1978]; [Fu, 1981]; [Rosenfeld, 1981]; [Peli, 1982];
- **1985-1994**: [Haralick, 1985]; [Nevatia, 1986]; [Pavlidis, 1986]; [Borisenko, 1987]; [Sahoo, 1988]; [Buf, 1990]; [Sarkar, 1993]; [Pal, 1993]

Brief Summary for Image Segmentation

- **Summary of Survey Papers (2)**
  - All these survey papers are dated in the second and third decades
  - The reason for no survey in the first decade is because the research results were just cumulating in that period
  - The reason for no survey in the last decade maybe attribute to the factor that so many techniques have already been presented, thus a comprehensive survey becomes less feasible

Brief Summary for Image Segmentation

- **Summary of Survey Papers (3)**

Various theories and models have been employed, for example:

Brownian string, Evolution theory, Expert system, Fractal, Fuzzy logic, Gabor filter, Gaussian mixture model, Genetic algorithm, Gibbs random field, Graph theory, Hidden Markov model, Level set, Markov random field, Neural network, Rough set, Simulated annealing, Wavelet, **……**

Brief Summary for Image Segmentation

- **Summary of Survey Papers (4)**

Some specialized / particular surveys have been published in the last 10 years

(1) Focused on particular group of segmentation algorithms: [Olabarriaga, 2001], [Freixenet, 2002], [Behiels, 2002], [Marcello, 2004]

(2) Focused on a particular application of image segmentation: [Pham, 2000], [Koprinska, 2001], [Lefèvre, 2003], [Kirbas, 2003], [Prati, 2003]

A General Framework for Segmentation and Its Evaluation

- **Segmentation Evaluation**

- None of the developed segmentation algorithms are generally applicable to all kinds of images and different algorithms are not equally suitable for a particular application
- Necessity of evaluation has been justified
- The history of segmentation evaluation could be traced back to 30 years' ago [Fram 1975]
- More than 100 major works have been reported

A General Framework for Segmentation and Its Evaluation

- **Segmentation Characterization**
  Intra-technique task
- **Segmentation Comparison**
  Inter-technique task
  - **Qualitative Evaluation**
    Ranking: good, acceptable, or unacceptable
  - **Quantitative Evaluation**
    Numeral score values: [0, 1]

A General Framework for Segmentation and Its Evaluation

- **Segmentation Characterization (1)**

The purpose of evaluation for a specific algorithm is to quantitatively recognize its behavior in treating various images and/or to help appropriately setting its parameters regarding different applications to achieve the best performance of this algorithm

This process could also help to improve the functioning of the algorithm under consideration

A General Framework for Segmentation and Its Evaluation

- **Segmentation Characterization (2)**

(1) Using the same parameter setting of the algorithm for segmenting multiple images. The ability and consistency of the algorithm in treating images with different contents and/or acquired under various conditions are evaluated

(2) Giving different values to the algorithm's parameters for segmenting some comparable images and then evaluating the influence of multiple settings of the algorithm over its performance. The adaptability and the best performance of this algorithm for given images are evaluated

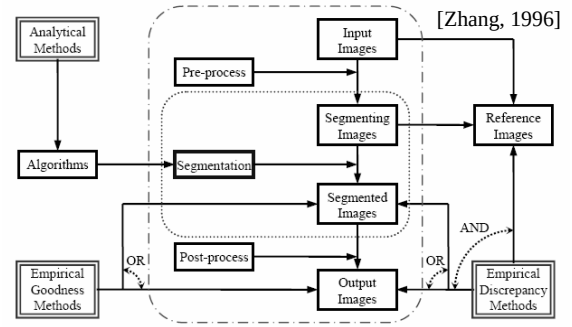
A General Framework for Segmentation and Its Evaluation

- **Segmentation Comparison**

The purposes of comparison for different segmentation algorithms are:

- to rank their performance under given conditions
- to provide guidelines in choosing suitable algorithms in facing to the desired applications
- to promote new development ideas by effectively taking into the strong points of several algorithms

A General Framework for Segmentation and Its Evaluation



A General Framework for Segmentation and Its Evaluation

- **Analytical Method**
- **Empirical Goodness Method**

  Un-supervised, Standalone
- **Empirical Discrepancy Method**

  Supervised, Relative objective

{ Analytical
{ Empirical { goodness
{ discrepancy

A General Framework for Segmentation and Its Evaluation

- **Analytical Evaluation**

(1) Amount of *a priori* knowledge that has been incorporated into segmentation algorithms

(2) Processing strategy: Parallel, Sequential, ……

(3) Processing complexity and efficiency

(4) Probability of correct detection / probability of false detection

(5) Resolution of segmented images: pixel, sub-pixel, group of pixels, ……

A General Framework for Segmentation and Its Evaluation

- **Evaluation Criteria (1)**

One essential element and critical factor

(Performance) metrics, measures, indices, etc.

Subjective criteria: reflect some desirable properties of segmented images, used in empirical goodness methods

Objective criteria: indicate the difference between the segmented images and reference images, used in empirical discrepancy methods

A General Framework for Segmentation and Its Evaluation

- **Evaluation Criteria (2)**

The behavior of an algorithm is dependent of many factors, a single metric for entire assessment can hardly reach an optimal solution

To better cover the various aspects of the algorithm, composite metric needs to be formed

(1) The combination of different metrics is often too empirical to be effective

(2) A final score is still needed

Empirical Evaluation Criteria and Methods

- **Methods ⇔ Criteria**

➢ The characteristics of evaluation methods are mainly determined by the criteria used

➢ Empirical evaluation could always provide quantitative evaluation score

➢ For empirical evaluation methods, suitable empirical evaluation criteria based on subjective (goodness) or objective (discrepancy) principle could be used

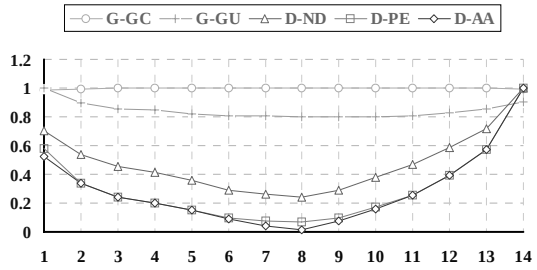
Empirical Evaluation Criteria and Methods

- **Compared Criteria** [Zhang, 1996]

Class	Criterion name	Method group
G-1	Intra-region uniformity	Goodness
G-2	Inter-region contrast	Goodness
G-3	Region shape	Goodness
D-1	Number of mis-segmented pixels	Discrepancy
D-2	Position of mis-segmented pixels	Discrepancy
D-3	Number of objects in the image	Discrepancy
D-4	Feature values of segmented objects	Discrepancy
D-5	Miscellaneous	Discrepancy

Empirical Evaluation Criteria and Methods

- **A Comparison Result**



Empirical Evaluation Criteria and Methods

- **Expanded Criteria**

  Listed but not compared [Zhang, 1996]

Class	Criterion name	Method group
G-4	Moderate number of regions	Goodness
D-5a	Region consistency	Discrepancy
D-5b	Grey level difference	Discrepancy
D-5c	Symmetric divergence (cross-entropy)	Discrepancy
S1	Amount of editing operations	Special
S2	Visual inspection	Discrepancy like
S3	Correlation between original image and bi-level image	Goodness like

Empirical Evaluation Criteria and Methods

Class	Criterion name	Method group
G-1	Intra-region uniformity	Goodness
G-2	Inter-region contrast	Goodness
G-3	Region shape	Goodness
G-4	Moderate number of regions	Goodness
D-1	Number of mis-segmented pixels	Discrepancy
D-2	Position of mis-segmented pixels	Discrepancy
D-3	Number of objects in the image	Discrepancy
D-4	Feature values of segmented objects	Discrepancy
D-5a	Region consistency	Discrepancy
D-5b	Grey level difference	Discrepancy
D-5c	Symmetric divergence (cross-entropy)	Discrepancy
S1	Amount of editing operations	Special
S2	Visual inspection	Discrepancy like
S3	Correlation between original image and bi-level image	Goodness like

A Recent Progress Overview of Segmentation Evaluation

- **New Progresses in Segmentation Evaluation**
- **Three Categories:**

  Recent evaluation works, mainly according to the criteria used, can be classified into three categories

  (1) Evaluation Works Based on Existing Criteria

  (2) Evaluation Works Made with Modifications / Improvements (of existing criteria)

  (3) Evaluation Works Supplying New Inspiration

A Recent Progress Overview of Segmentation Evaluation

- **Evaluation Works Based on Existing Criteria**

Method #	Source	Criteria used	Method #	Source	Criteria used
M-1	(Hoover, 1996)	D-5a	M-10	(Huo, 2000)	D-1, D-4
M-2	(Zhang, 1997)	D-4	M-11	(Cavallaro, 2002)	D-1, D-2
M-3	(Borsotti, 1998)	G-1, G-2, G-4	M-12	(Prati, 2003)	D-1
M-4	(Xu, 1998)	S-3	M-13	(Rosin, 2003)	D-1
M-5	(Chang, 1999)	D-5a	M-14	(Lievers, 2004)	G-1
M-6	(Yang,1999)	D-1	M-15	(Marcello, 2004)	S-2
M-7	(Mattana, 1999)	D-4	M-16	(Renno, 2004)	D-1, D-4
M-8	(Rosenberger, 2000)	G-1, G-2	M-17	(Carleer, 2004)	D-1, D-3
M-9	(Betanzos, 2000)	D-1	M-18	(Ladak, 2004)	D-1, S-1

A Recent Progress Overview of Segmentation Evaluation

- **Evaluation Works Made with Modifications / Improvements (of existing criteria)**

Method #	Source	Criteria used (modification)
M-19	(Oberti, 1999)	D-1 (ROC, curve of FP vs. FN)
M-20	(Gao, 2000)	D-1 (ROC, curve of FP vs. FN)
M-21	(Correia, 2000)	D-1 (with spatial and temporal extension)
M-22	(Udupa, 2002)	D-1, S-1 like (efficiency)
M-23	(Li, 2003)	D-1 and D-2, (contour matching, temporal consistency), S-1
M-24	(Zhang, 2004)	G-1, G-2, G-4 (using region entropy)
M-25	(Erdem, 2004)	G-1, G-2 (with extension to color, motion, color histograms)
M-26	(Niemeijer, 2004)	D-1 (ROC, curve of TP vs. FP)
M-27	(Udupa, 2004)	D-1 (DOC, curve of TP vs. FP)
M-28	(Kim, 2004)	D-1 (PDR, modified detection rate)

A Recent Progress Overview of Segmentation Evaluation

- **Evaluation Works Supplying New Inspiration**

Method #	Source	Novelty
M-29	(Everingham, 2002)	Finding out the Pareto front in a multi-dimensional fitness space
M-30	(Li, 2003)	Finding out the Pareto front in a 4-D fitness space
M-31	(Correia, 2003)	Using contextual relevance metric to match human visual system (HVS)
M-32	(Zhang, 2005)	Using weighted majority (WM), Bayesian and support vector machine (SVM)
M-33	(Desurmont, 2005)	Performing evaluation in different semantic levels

A Recent Progress Overview of Segmentation Evaluation

- Some **Observations**
  - Most new works based on existing criteria use empirical discrepancy criteria
  - Many new works made with modifications / improvements on existing criteria use ROC and its variations: DOC, PDR
  - New inspirations are mainly on how to combine several criteria into a composite criterion

A Comparison of Various Evaluation Methods

- **Works at Top level**
  - The history of systematic comparison of segmentation evaluation methods can only be traced back to about 10 years' ago [Zhang 1993]
  - Thèse présentée par Sébastien CHABRIER "*Évaluation de la segmentation d'images*"
  - An edited book "Advances in Image and Video Segmentation" will be published in 2006 by Idea Group, Inc.

A Comparison of Various Evaluation Methods

- **Four Factors**

  (Considering both the techniques and measures used in evaluation [Zhang 1993, Zhang 1996]):

  (1) Generality for evaluation

  (2) Subjective versus objective and qualitative versus quantitative

  (3) Complexity for evaluation

  (4) Consideration of segmentation applications

A Comparison of Various Evaluation Methods

Method #	Generality	Complexity	Method #	Generality	Complexity
M-1	General	Medium	M-15	General	High (Human)
M-2	General	Medium	M-16	General	Med./High
M-3	Numerous objects	Medium/High	M-17	Numerous objects	Low/Medium
M-4	Tree structure	High	M-18	General	High (Human)
M-5	Particular	Medium	M-19	General	Medium
M-6	General	Medium	M-20	Video	High
M-7	General	Low/Medium	M-21	General	Medium/High
M-8	General	Medium/High	M-22	General	Medium
M-9	General	Medium	M-23	General	High (Human)
M-10	General	Medium	M-24	Numerous objects	Medium
M-11	Video	Medium/High	M-25	General	Medium
M-12	General	High	M-26	General	Medium
M-13	Video	Medium	M-27	General	Medium
M-14	Thresholding	Medium	M-28	Video	Medium

Concluding Remarks

- **Numbers of Works Made**

Level	Description	Publication #
0 $f(x)$	Segmentation Algorithms	$O(10^3)$
1 $f'(x)$	Technique Evaluation	$O(10^2)$
2 $f''(x)$	Comparison of Evaluation Methods	$O(10^1)$

Concluding Remarks

- **Some Points about Evaluation**
  - More efforts have been put on evaluation recently
  - However, no many really redical changes / improvements have been widely reported
  - Some criteria are educed from the existing ones
  - No single evaluation method can be used in all circumstance (algorithms, images, ……)
  - No single evaluation criterion can cover all aspects of segmentation algorithms

Concluding Remarks

- **Limiting Factors for Evaluation**
  - (1) There is no common mathematical model or general strategy for evaluation
  - (2) It is difficult to define wide-ranging performance metrics and statistics
  - (3) The testing data used in evaluation are often not representative enough for actual application
  - (4) Appropriate ground truths are hard to determine objectively
  - (5) Often large costs (both time and effort) are involved in performing comprehensive evaluation

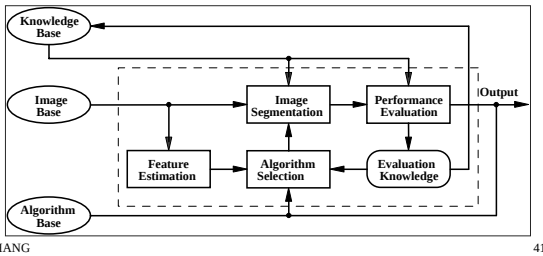
Concluding Remarks

- **Potential Research Directions**
  - (1) Combine multiple metrics efficiently
  - (2) Make evaluation in considering the final goal of segmentation
  - (3) Construct common databases for segmentation evaluation
  - (4) Characterize and compare various evaluation methods
  - (5) Real use of evaluation results for segmentation

Concluding Remarks

- **One Example Utilization of Evaluation**

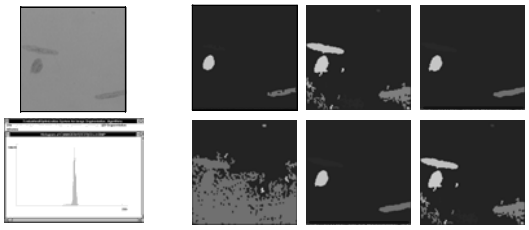
Optimal selection of segmentation algorithms based on evaluation with the help of expert system



Concluding Remarks

- **One Example Utilization of Evaluation**

40 tests: 75% investigating two algorithms
25% investigating three algorithms



References (Early-Day)

- Roberts L G. (1965). Machine perception of three-dimensional solids. In: Optical and Electro-Optical Information Processing, Tippett J, *et al.*, eds., 159-197
- Fram J R, Deutsch E S. (1975). On the quantitative evaluation of edge detection schemes and their comparison with human performance. IEEE Trans. C-24, 616-628
- Fu K S, Mui J K. (1981). A survey on image segmentation. Pattern Recognition, 13, 3-16
- Zhang Y J. (1993). Comparison of segmentation evaluation criteria. Proc. 2ICSP, 870-873

Thanks for Your Attention !

- **Department of Electronic Engineering**
- **Tsinghua University, Beijing 100084, China**
- **Tel: +86-10-62781430**
- **Fax: +86-10-62770317**
- **E-mail: zhang-yj@mail.tsinghua.edu.cn**
- **H-page: www.ee.tsinghua.edu.cn/~zhangyujin/**
- **L-Web: image.ee.tsinghua.edu.cn**