ELSEVIER

Computerized Medical Imaging and Graphics 30 (2006) 75–87

Computerized Medical Imaging and Graphics

www.elsevier.com/locate/compmedimag

A framework for evaluating image segmentation algorithms

Jayaram K. Udupa [a,*], Vicki R. LeBlanc [b,c], Ying Zhuge [a], Celina Imielinska [b,d,e], Hilary Schmidt [b,c], Leanne M. Currie [f], Bruce E. Hirsch [g], James Woodburn [h]

[a] *Medical Image Processing Group, Department of Radiology, University of Pennsylvania, Philadelphia, PA, USA*
[b] *Office of Scholarly Resources, Columbia University College of Physicians and Surgeons, New York, NY, USA*
[c] *Center for Education Research and Evaluation, Columbia University College of Physicians and Surgeons, New York, NY, USA*
[d] *Department of Biomedical Informatics, Columbia University College of Physicians and Surgeons, New York, NY, USA*
[e] *Department of Computer Science, Columbia University, New York, NY, USA*
[f] *School of Nursing, Columbia University, New York, NY, USA*
[g] *Department of Neurobiology and Anatomy, Drexel University College of Medicine, Philadelphia, PA, USA*
[h] *Rheumatology and Rehabilitation Research Unit, University of Leeds, Leeds, UK*

Received 27 August 2005; accepted 12 December 2005

Abstract

The purpose of this paper is to describe a framework for evaluating image segmentation algorithms. Image segmentation consists of object recognition and delineation. For evaluating segmentation methods, three factors—precision (reliability), accuracy (validity), and efficiency (viability)—need to be considered for both recognition and delineation. To assess precision, we need to choose a figure of merit, repeat segmentation considering all sources of variation, and determine variations in figure of merit via statistical analysis. It is impossible usually to establish true segmentation. Hence, to assess accuracy, we need to choose a surrogate of true segmentation and proceed as for precision. In determining accuracy, it may be important to consider different 'landmark' areas of the structure to be segmented depending on the application. To assess efficiency, both the computational and the user time required for algorithm training and for algorithm execution should be measured and analyzed. Precision, accuracy, and efficiency factors have an influence on one another. It is difficult to improve one factor without affecting others. Segmentation methods must be compared based on all three factors, as illustrated in an example wherein two methods are compared in a particular application domain. The weight given to each factor depends on application.

Keywords: Image segmentation; Evaluation of segmentation; Image analysis; Segmentation efficacy

1. Introduction

1.1. Background

Image segmentation is the process of identifying and delineating objects in images. It is the most crucial among all computerized operations done on acquired images. Even seemingly unrelated operations like image (gray-scale/color) display, 3D visualization, interpolation, filtering, and registration depend to some extent on image segmentation since they all would need some object information for their optimum performance. Ironically, segmentation is needed for segmentation itself since object knowledge facilitates segmentation. In spite of several decades of research [1,2], segmentation remains a challenging problem in image processing and computer vision.

Image segmentation may be thought of as consisting of two related processes—*recognition* and *delineation*. *Recognition* is the high-level process of determining roughly the whereabouts of an object of interest in the image. *Delineation* is the low-level process of determining the precise spatial extent and point-by-point composition (material membership percentage) of the object in the image. Humans are more qualitative and less quantitative, whereas, computerized algorithms are more quantitative and less qualitative. Incorporation of high-level expert

* Corresponding author. Address: Medical Image Processing Group, Department of Radiology, University of Pennsylvania, 423 Guardian Drive, Fourth Floor, Blockley Hall, Philadelphia, PA 19104-6021, USA. Tel.: +1 215 662 6780; fax: +1 215 898 9145.

E-mail address: jay@mipg.upenn.edu (J.K. Udupa).

0895-6111/$ - see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.compmedimag.2005.12.001

human knowledge algorithmically into the computer has remained a challenge. Most of the drawbacks of current segmentation methods may thus be attributed to the latter weakness of computers in the recognition process. We envisage, therefore, that the assistance of humans, knowledgeable in the application domain, will remain essential in any practical image segmentation method. The challenge and goal for image scientists are to develop methods that minimize the degree of this required help as much as possible.

While algorithms for image segmentation have been in development for several decades [1,2], the development of systematic evaluation frameworks for these algorithms has been lagging, particularly in medical imaging which is the focus of this paper. The lag is perhaps the result of problems such as limits in common data sets with which to compare methods, difficulty in defining the performance metrics and statistics, and the difficulty in establishing true segmentation. As early as 1977, the need for effective evaluation of the segmentation of biological images has been outlined [3]. More recently, this need has been echoed by many researchers [2,4–6]. In [4,5], the authors stress the need for an objective evaluation of medical image segmentation on large sets of common clinical data, arguing that this is a critical step towards establishing the validity and the clinical applicability of an algorithm. Similarly, [6] claims that the development of an objective approach will provide consistency in evaluation methods by removing biases due to human factors. Many attempts at evaluation do not address the important components that should be present in any evaluation methodology, thus limiting their validity and clinical applicability. Claims about the performance of segmentation algorithms are limited by problems such as (a) the data sets are too small, (b) different data sets are used for different estimations of performance, (c) the data sets are not representative of a clinical problem, (d) appropriate ground truths (or surrogates) are difficult to determine, (e) the performance metrics are poorly defined, (f) there is poor methodology for training and testing the algorithms, (g) large costs of time and effort are involved in collecting and hand-segmenting data, and (h) the algorithms are not compared against other algorithms [5,7].

In light of such difficulties, it is not surprising that many researchers develop complex applications (e.g. virtual colonoscopy systems) that make use of 3D visualizations of anatomical images derived from 3D segmentation methods that have not been formally evaluated by a consistent evaluation strategy (e.g. [8,9]). Many of the researchers who do evaluate their segmentation algorithms do so only on a limited number of components, such as cost analysis [10], inter-rater reliability [11], overall volume [12], or the Hausdorff distance [13]. These efforts, despite representing a valid attempt at evaluation, exemplify the difficulty in devising comprehensive and effective segmentation evaluation methodologies in this domain.

Few researchers [4–7] have made attempts to develop evaluation frameworks that incorporate many of the performance metrics necessary for a practical and informative evaluation of a segmentation algorithm. In [7], the authors discuss the variety of metrics that would result in a valid estimation of the performance of an algorithm. When comparing a segmentation method to a ground truth segmentation of the image, [7] argues that there are five possible outcomes that need to be identified. The computer algorithm can either (a) correctly segment a region, (b) over-segment a region, (c) under-segment a region, (d) miss a region, or (e) incorrectly segment a noise region. Hoover et al. [5] also developed a rigorous framework for the evaluation of segmentation algorithms. This involved the use of pixel-level ground truths in 30 real images. The ground truth consisted of the hand-segmentation, which was reviewed by a second human operator to catch obvious errors. Each pixel in the region segmented by the computer algorithm was classified as either a correct detection, an over-segmentation, an under-segmentation, a missed pixel, or noise. Four algorithms were then compared and described on the basis of these metrics, as well as on the basis of processing time. Zhang [6] approaches evaluation of segmentation methods by proposing analytical and empirical methods, where the empirical methods are divided into goodness and discrepancy measurements. The analytical methods examine and assess the segmentation algorithms themselves by analyzing their principles and properties. The empirical methods indirectly judge the algorithms by testing the images and evaluating the quality of segmentation results. The weakness of this approach is that it is intended for ‘all images’. Because of the lack of a general theory for image segmentation, not all characteristics of segmentation can be obtained and described by analytical studies.

We argue that a primary reason for the lack of activity in evaluation, commensurate with the level of investigation in segmentation algorithm development, is the lack of a framework which algorithm developers can readily utilize, without having to spend a great deal of time, to assess the efficacy of their methods. Such a framework, we believe, should consist of: (F1) a specification of readily computable, effective, and meaningful metrics of efficacy, (F2) real life image data, (F3) reference segmentations that can be used as surrogates of true segmentations (ground truth), (F4) a few standard segmentation algorithms, and (F5) a software system that incorporates the evaluation methods and the standard segmentation algorithms. We shall use the phrase *evaluation framework* to refer to this quintuple of components (F1)–(F5). It is clear from the above description that a comprehensive framework for the evaluation of segmentation algorithms in the sense of including the five components is lacking. Even the metrics of efficacy have not considered all important