Image and Vision Computing 31 (2013) 877–886

Contents lists available at ScienceDirect

Image and Vision Computing

journal homepage: www.elsevier.com/locate/imavis

A new evaluation measure for color image segmentation based on genetic programming approach☆

Hakime Vojodi *, Ali Fakhari, Amir Masoud Eftekhari Moghadam

Department of IT and Computer Engineering, Qazvin Branch, Islamic Azad University, Qazvin, Iran

ARTICLE INFO

Article history:
Received 29 June 2012
Received in revised form 14 June 2013
Accepted 12 August 2013

Keywords:
Image segmentation
Evaluation measure
Genetic programming
Intra-region
Inter-region
Over-segmentation

ABSTRACT

One of the greatest challenges while working on image segmentation algorithms is a comprehensive measure to evaluate their accuracy. Although there are some measures for doing this task, but they can consider only one aspect of segmentation in evaluation process. The performance of evaluation measures can be improved using a combination of single measures. However, combination of single measures does not always lead to an appropriate criterion. Besides its effectiveness, the efficiency of the new measure should be considered. In this paper, a new and combined evaluation measure based on genetic programming (GP) has been sought. Because of the nature of evolutionary approaches, the proposed approach allows nonlinear and linear combinations of other single evaluation measures and can search within many and different combinations of basic operators to find a good enough one. We have also proposed a new fitness function to make GP enable to search within search space effectively and efficiently. To test the method, Berkeley and Weizmann datasets besides several different experiments have been used. Experimental results demonstrate that the GP based approach is suitable for effective combination of single evaluation measures.

1. Introduction

Image segmentation is an essential problem in image processing, computer vision, video and image retrieval applications. It is often used to partition an image into separate regions for content analysis and image understanding. Because segmentation algorithms are often used as a preprocessing step for larger systems, we have to say that their improper performance can highly impress on the final results. To address this challenge, a variety of segmentation algorithms have been proposed [1–4]. Because defining a unique algorithm applicable in all domains is almost impossible, most of these algorithms are domain-specific. Hence, looking for some measures for assessing segmentation results is needed. Defining a good measure to evaluate the segmentation is a known challenging problem in machine vision and image processing applications.

Generally, supervised and unsupervised techniques are two well-known approaches used for obtaining a good evaluation measure. Generally, supervised algorithms have been used more in researches [5–8]. Supervised evaluations are also called relative evaluation or empirical discrepancy methods and need some prior knowledge such as a ground truth image. Unsupervised evaluation techniques are also known as stand-alone algorithms or empirical goodness methods. They are quantitative and objective. Unsupervised methods compute some statistical data in the segmentation results, without any prior knowledge. They compare segmentation methods without requiring human intervention or a manually segmented or pre-processed reference image. Unsupervised methods are important for real-time segmentation evaluation and can also be used for self-tuning algorithm [9].

From segmentation point of view, different images can be divided into several categories including color, gray-level, textural images, etc. Each of these categories can have its own offsprings like natural, medical, satellite, submarine images, etc. Both segmentation approaches and their evaluation measures are selected according to the images' features. For example we can refer to evaluation measures for medical images [10–12], for natural images [6,8,9], etc.

One of the important characteristics of evaluation measures is their accuracy. Almost all of the existing measures can evaluate only one or two features in a segmented image (such as error color, entropy, homogeneity regions, variance, etc.). Most of them are not sensitive to under and over-segmentation and cannot detect them. To improve the accuracy of evaluation, one approach is combining them. However, combining various measures can be too complicated to do. Furthermore, they may not be effective for all images, even for image in one class.

In this paper, a genetic programming based approach has been considered for proposing a new, effective and efficient evaluation measure. GP has been employed to combine different and unrelated evaluation measures. Each evaluation measure can be used either without any or with a little modification. For doing so, some proper single measures for a particular image type, according to the database, should be selected. After selecting the single measures, GP combines them so that the new measure works for all images in that particular category.

☆ This paper has been recommended for acceptance by Sinisa Todorovic.

* Corresponding author. Tel.: +98 911 9442764; fax: +98 182 4222391.
E-mail addresses: h.vojodi@qiau.ac.ir (H. Vojodi), a.fakhari@qiau.ac.ir (A. Fakhari), eftekhari@qiau.ac.ir (A.M. Eftekhari Moghadam).

0262-8856/$ – see front matter © 2013 Published by Elsevier B.V.
http://dx.doi.org/10.1016/j.imavis.2013.08.002

GP [13] belongs to a set of artificial intelligence problem-solving techniques which are based on the principles of biological inheritance and evolution. Each potential solution is called an individual (i.e. a chromosome) in the population. The main difference between GA [14] and GP is their internal representation (or data structure) for an individual. Collectively, GA-based applications usually represent each individual as a fixed-length bit string. In GP, on the other hand, more complex data structures like trees, linked lists, or stacks are used [15].

GP iteratively transforms a population of computer programs into a new generation by applying naturally occurring genetic operations. These operations are applied to individuals selected from the previous population [16]. The individuals are probabilistically selected to be applied into the genetic operations based on their fitness. The genetic operators include crossover (sexual recombination), mutation, reproduction, gene duplication, and gene deletion. In short, GP becomes a powerful method to solve the NP hard problems by using the capabilities of evolutionary search [13] and has been applied successfully in several problems [17,18].

In this paper, four unsupervised evaluation measures based on intra-region similarity and inter-regions disparity have been employed in combination process. The evaluation measures include extended version of Zeboudj [19] and Rosenberger evaluation [20], an evaluation measure based on entropy [21], and a new measure based on intra-region similarity and inter-region disparity. The proposed combined measure is able to take all features of these four evaluation criteria including color error, squared color error, entropy on each region, average color difference between regions, local color difference along boundaries, and layout entropy and then combine them in an effective way. Also, a new fitness function for GP has been proposed.

The rest of this paper is organized as follows. In Section 2, some proposed evaluation measures will be reviewed. Section 3 presents the GP framework for combining evaluation measures. Section 4 discusses about the experiments and the results. In Section 5, besides a brief conclusion, some opportunities for further studies will be discussed.

2. The related works

A variety of techniques have been proposed to evaluate image segmentation algorithms. In this section, the previously proposed measures for addressing image segmentation evaluation task will be reviewed.

Levine and Nazif [9] defined a criterion that calculates the uniformity of a region based on the variance of its characteristic. Complementary to the intra-region uniformity, Levine and Nazif defined a disparity measurement between two regions to evaluate the dissimilarity of regions in segmentation results [22].

Liu and Yang [9] proposed an evaluation function based on empirical studies. The parameters of their approach do not need to be set by a user and are defined automatically. Also, they are independent of the type of image. Borsotti et al. [22] improved Liu and Yang's method by some modified quantitative measures. All of these evaluation functions were generated from the results of empirical analysis and had no theoretical foundations.

Zeboudj [23] proposed a measure based on the combination of maximum inter-region similarity and minimal intra-region disparity in a neighborhood. Rosenberger [20] presented a criterion that estimated the intra region homogeneity and the inter regions disparity. This criterion quantifies the quality of a segmentation result.

Zhang et al. [24] proposed a novel segmentation evaluation method based on information theory. This method uses entropy to measure the uniformity of pixel characteristics within a segmentation region. Instead of using squared color error, they used the Shannon entropy of the luminance of all pixels in a region to measure its uniformity. Pal and Bhandari [25] proposed an entropy-based segmentation evaluation measure for intra-region uniformity based on the second order local entropy.

Chen and Wang [26] proposed a composite evaluation method for color images. Their method uses intra-region visual error to evaluate the degree of under-segmentation, and uses inter-region visual error to evaluate the degree of over-segmentation.

Zhang et al. [27] proposed a co-evaluation framework in which different measures judge the performance of the segmentation in different ways. Then, their results are combined by using a machine learning approach. After determining different evaluation measures as well or poor, their approach gains the appropriate measures to achieve a reliable result [27]. However, each single basic evaluator typically performs well in some types of images/segmentations, whereas works poorly for some others. Therefore, Zhang et al. [28] proposed a meta-evaluation method to resolve limitations of co-evaluation frameworks.

Meta-evaluation methods combine the results of a set of basic measures using machine learning algorithms (Naive Bayes, support vector machine, or the weighted majority algorithm) in order to obtain an overall acceptable evaluation measure. The training data used by the machine learning algorithm can be labeled by a user, based on either similarity to reference segmentation or system-level performance.

Several evaluation metrics that are designed for video frames can be easily modified and be used for image segmentation evaluation [29,30]. These methods use measures similar to image segmentation evaluation.

Correia and Pereira [31] proposed a method which consists of a set of metrics for both intra-object measures (e.g. shape regularity, spatial uniformity) and inter-object measures (such as contrast). Furthermore, each object in the image is weighted according to its relevance, which is an estimation of how much the reviewer's attention is attracted to that particular object.

Erdem et al. [32] proposed an evaluation measure based on spatial color contrast within the boundary of each object. The key component is the difference between the average colors of pixels and their neighbors in a region, averaged by the total number of lines drawn on object boundaries.

In [33] a Bayesian network framework for unsupervised evaluation of image segmentation has been proposed. This image understanding algorithm utilizes a set of given Segmentation Maps (SMs) ranging from results of under to over-segmentation for a target image. This method identifies some semantically meaningful objects in an image and ranks the SMs according to their applicability.

3. The GP approach for combining evaluation measures

In this section, details of the proposed GP framework will be discussed in details. The details include the different aspects of GP learning systems such as selection of terminal set, function set, fitness measure, input parameters, and determining the termination conditions.

3.1. The GP approach

The so far existing measures cannot have a comprehensive evaluation for image segmentation algorithm because they do not consider all the parameters which can impress on the performance. So, they have to be combined, of course in a proper manner which must be effective and efficient. According to the number of evaluation measures and different possible arithmetic operations, there is a large search space in which, we have to find the optimum, or at least a good enough solution.

The proposed approach uses GP to combine evaluation measures. This method relies on the creation of a composite measure, which is the combination of pre-defined evaluation measures. It also produces meaningful combination of evaluation measures.

GP is guided by the fitness function to search for the most effective computer program for solving a given problem. A fitness function assigns a fitness value to each individual in order to select the best ones. Then, genetic operators such as reproduction, crossover, and mutation are applied to create more diverse and efficient ones.

The reproduction operator simply involves the selection of an individual based on the fitness function and making a copy of it in the next generation. Mutation and crossover are two primary genetic operators, which are used to create new programs based on existing ones.