A SURVEY ON EVALUATION METHODS FOR IMAGE SEGMENTATION

Y.J.ZHANG

Department of Electronic Engineering, Tsinghua University, 100084 Beijing, China

ABSTRACT

This paper studies different methods proposed so far for segmentation evaluation. Most methods can be classified into three groups: the analytical, the empirical goodness and the empirical discrepancy groups. Each group has its own characteristics. After a brief description of each method in every group, some comparative discussions about different method groups are first carried out. An experimental comparison for some empirical (goodness and discrepancy) methods commonly used is then performed to provide a rank of their evaluation abilities. In addition, some special methods are also discussed. This study is helpful for an appropriate use of existing evaluation methods and for improving their performance as well as for systematically designing new evaluation methods.

Image analysis, Image segmentation, Segmentation evaluation, Analytical and empirical study, Performance assessment, Criteria function, Algorithm comparison, Image quality measure, Method characterization.

1. INTRODUCTION

Image analysis usually refers to processing of images by computer with the goal of finding what objects are presented in the image.[1] Image segmentation is one of the most critical tasks in automatic image analysis. It consists of subdividing an image into its constituent parts and extracting these parts of interest (objects). A great variety of segmentation algorithms have been developed in the last decades and this number continually increases each year.[2] Several survey papers for segmentation techniques have been presented in the literature.[3-9] Since none of the proposed segmentation algorithms are generally applicable to all images and different algorithms are not equally suitable for a particular application,[9] the performance evaluation of segmentation algorithms is indispensable and thus an important subject in the study of segmentation. More generally, performance evaluation is critical for all computer vision algorithms from research to application,[10] while image segmentation is an essential and important step of low-level vision.

While development of segmentation algorithms has attracted significant attention, relatively fewer efforts have been spent on their evaluation, although many newly developed algorithms are (most often subjectively) compared with some particular algorithms with few particular images. Moreover, most efforts spent on evaluation are just for designing new evaluation methods and only very few authors have attempted to characterize the different evaluation methods existed.[11] The present paper will review different existing methods for segmentation evaluation, as well as discuss and compare their applicability, advantages and limitations.

Segmentation algorithms can be evaluated analytically or empirically, so the evaluation methods can be divided into two categories: the analytical methods and the empirical methods. The analytical methods directly examine and assess the segmentation algorithms themselves by analysing their principles and properties. The empirical methods indirectly judge the segmentation algorithms by applying them to test images and measuring the quality of segmentation results. Various empirical methods have been proposed. Most of them can still be classified into two types: goodness methods and discrepancy methods. In the first category some desirable properties of segmented images, often established according to human intuition, are measured by "goodness" parameters. The performances of segmentation algorithms under investigation are judged by the values of goodness measures. In the second category some references that present the ideal or expected segmentation results are first found. The actual segmentation results obtained by applying a segmentation algorithm, sometimes preceded by preprocessing and/or followed by postprocessing processes, are compared with the references by counting their differences. The performances of segmentation algorithms under investigation are then assessed according to the discrepancy measures. Following this discussion, three groups of methods can be distinguished.

The above classification for evaluation methods can be seen more clearly in Fig.1, where a general scheme for segmentation and its evaluation is presented. The input image obtained by sensing is first (optionally) preprocessed to produce the segmenting image for the segmentation (in its strict sense) procedure. The segmented image can then be (optionally) postprocessed to produce the output image. Further processes, such as feature extraction and measurement, will be based on these output images. In Fig.1 the part enclosed by the rounded square with thin line corresponds to the segmentation procedure in its narrow-minded sense, while the part enclosed by the rounded square with point line corresponds to the segmentation procedure in its general form. The black arrows indicate the processing directions of segmentation. The access points for the three groups of evaluation methods are depicted with gray arrows in Fig.1. Note that there is an *or* condition between both arrows leading to the boxes containing segmented image" and output image" both from the empirical goodness method" and empirical discrepancy method". Moreover, there is an *and* condition between the arrow from empirical discrepancy method" to reference image" and the two (*or*) arrows going to eegmented image" and output image". The analysis methods treat the algorithms for segmentation directly. The empirical goodness methods judge the segmented image or output image so as to indirectly assess the performance of algorithms.

For applying empirical discrepancy methods, the reference image is necessary. It can be obtained manually or automatically from the input image or segmenting image. The empirical discrepancy methods compare the segmented image or output image to the reference image and use their difference to assess the performance of algorithms.

Each method group has its own particularities so as to be distinguished from other groups. Each method has also its own characteristics so as to be identified. In the following three sections a brief description of the methods belonging to the three groups will be provided. They are arranged according to the above method classification. The justification of the classification of methods into analytical and empirical ones as well as the separation of empirical methods into goodness and discrepancy groups will be made clear by the comparative discussion of different method groups in Section 5. In addition, an experimental comparison of several commonly used empirical methods will be carried out in Section 6. These representative methods are compared according to their ability and behavior in evaluating the same series of segmented images. A rank among them is then obtained. In Section 7 several special evaluation methods that do not fall clearly into the above three groups and some common problems for most existing evaluation methods are discussed. Finally, some concluding remarks are given in Section 8.

2. ANALYTICAL METHODS

The analytical methods directly treat the segmentation algorithms themselves by considering the principles, requirements, utilities, complexity, *etc.*, of algorithms. Using the analytical methods to evaluate segmentation algorithms avoids the concrete implementation of these algorithms and the results could be exempted from the influence caused by the arrangement of evaluation experiments as the empirical methods do. However, not all properties of segmentation algorithms can be obtained by analytical studies. The difficulty, up to now, is the lack of general theory for image segmentation.[12] Although some initial attempts in the direction of a unified theory about segmentation were reported, for example, in the relation of image models and segmentation,[13] no formal solution has been found yet. Until now, the analytical methods work only with some particular models or desirable properties of algorithms.

One analytical method has been proposed by Liedtke *et al.*[14] They presented an evaluation study of several algorithms by taking into account the type and amount of *a priori* knowledge that has been incorporated into different segmentation algorithms. Such knowledge for certain segmentation algorithms is ready to be analysed, which is mainly determined by the nature of the algorithms. However, such knowledge is usually heuristic information and different types of *a priori* knowledge are hardly comparable. The information provided by this method is then rough and qualitative. On the other side, not only "the amount of relevant *a priori* knowledge that can be incorporated into the segmentation algorithm is decisive for the reliability of the segmentation methods",[14] but it is also very important for the performance of the algorithm how such *a priori* knowledge has been incorporated into the algorithm.[15]

The analytical methods can in certain cases provide quantitative information about segmentation algorithms. Abdou and Pratt[16] analysed the performance of several edge detectors with a detection probability ratio in a statistical design procedure. Let T be the edge decision threshold, P_C the probability of correct detection and P_f the probability of false detection:

$$P_c = \int_T^\infty p(t|edge)dt \quad (1)$$

$$P_f = \int_T^\infty p(t|no-edge)dt \quad (2)$$

the plot of P_C versus P_f in terms of T can provide a performance index of detectors. Such an index should be useful for evaluating the segmentation algorithms based on edge detection [for example, see reference (9)]. In contrast to the *a priori* knowledge discussed above, this index can be precisely defined and calculated for simple edge detectors.[16]

Other properties of segmentation algorithms that can be obtained by analysis include the processing strategy, processing complexity and efficiency and segmentation resolution of algorithm.[17-18] These properties could be helpful for selecting suitable algorithms in particular applications. For example, the processing strategy of segmentation algorithms can be parallel, sequential, iterative or mixed. The parallel algorithms are suitable for fast implementation. However, for images that are severely contaminated by noise, the performance of parallel algorithms is often poorer than that of sequential methods.[19]

3. EMPIRICAL GOODNESS METHODS

The methods in this group evaluate the performance of algorithms by judging the quality of segmented images. To carry out this work certain quality measures should be defined. Most measures are established according to human intuition about what conditions should be satisfied by an "ideal" segmentation (for example, a pretty picture). In other words, the quality of segmented images is assessed by some "goodness" measures. These methods characterize different segmentation algorithms by simply computing the goodness measures based on the segmented image without the *a priori* knowledge of the correct segmentation.[20] The application of these evaluation methods exempts the requirement for references, so that they can be used for on-line evaluation.

Different types of goodness measures have been proposed.

3.1. Goodness based on intra-region uniformity

Weszka and Rosenfeld proposed a threshold evaluation method that uses a busyness measure as the criterion to judge thresholded images.[21] To apply the busyness measure they assume that the images are composed of objects and background of compact shapes and not strongly textured. Under these assumptions, the thresholded images should look smooth rather than busy. In practice, they compute the amount of busyness for a thresholded image by using the gray-level co-occurrence matrix of the image.[22] That is, those entries of the co-occurrence matrix representing the percentage of object-background adjacencies are summarized. The lower the busyness, the smoother the thresholded images and the better the segmentation result. In consequence, the better the segmentation results, the higher the performance of applied algorithms.

Similar to Weszka and Rosenfeld, Nazif and Levine also believe that an adequate segmentation should produce images having higher intra-region uniformity, which is related to the similarity of property about region element.[23] The uniformity of a feature over a region can be computed on the basis of the variance of that feature evaluated at every pixel belonging to that region.[20] In particular, for a gray-level image *f(x,y)*, let R_i be *i*th segmented region, A_i be the area of R_i, then the gray-level uniformity measure *(GU)* of *f(x,y)* is:

$$GU = \sum_i \sum_{(x,y)\in R_i} \left[f(x,y) - \frac{1}{A_i} \sum_{(x,y)\in R_i} f(x,y) \right]^2 \quad (3)$$

A normalized uniformity measure (*NU*) has been proposed by Sahoo *et al.*:[8]

$$NU = 1 - GU/C \quad (4)$$

where *C* is a normalization factor. Generally, other features can also be used.

The intra-region uniformity, as a desired property of segmented images, can also be measured by the higher-order local entropy based on information theory.[24] Pal and Pal proposed a thresholding method that maximizes the second-order local entropy of the object and background regions.[24] This entropy H^2, for an assumed threshold *T*, is computed by:

$$H^2(T) = -\sum_{i=0}^{T} \sum_{j=0}^{T} p_{ij} \ln p_{ij} \quad (5)$$

where p_{ij} is the probability of occurrence of the pair *(i,j)* within the object/background. This entropy is also used by Pal and Bhandari[25] as a measure of the region homogeneity in segmented images for the performance evaluation of segmentation results.

3.2. Goodness based on inter-region contrast

Except for intra-region uniformity, Levine and Nazif also believe that an adequate segmentation should in addition produce images having higher contrast across adjacent regions.[20] In a simple case that a gray-level image *f(x,y)* consists of the object with average gray-level f_o and the background with average gray-level f_b, a gray-level contrast measure *(GC)* can be computed by:

$$GC = \frac{|f_o - f_b|}{f_o + f_b} \quad (6)$$

Note that the similar idea has been already used by Otsu[26] for evaluating the "goodness" of threshold values in the development of a histogram based threshold selection algorithm. By maximizing the between-region variance, a threshold value producing the highest region separability can be obtained.

3.3. Goodness based on region shape

Not only the gray level but also the form of a segmented region can be taken into account to design goodness measures for satisfying the human intuition on an "ideal" segmentation. Sahoo *et al.*[8] proposed a shape measure (*SM*) for evaluating several threshold selection algorithms, which is defined as:

$$SM = \frac{1}{C} \left\{ \sum_{(x,y)} Sgn[f(x,y) - f_{N(x,y)}]\, g(x,y)\, Sgn[f(x,y) - T] \right\} \quad (7)$$

where $f_{N(x,y)}$ is the average gray value of the neighborhood *N(x,y)* of a pixel located at *(x,y)* with gray level *f(x,y)* and gradient value *g(x,y)*, *T* is the threshold value selected for segmentation, *C* is a normalization factor and *Sgn(.)* is the unit step function.

4. EMPIRICAL DISCREPANCY METHODS

In practical segmentation applications, some errors in the segmented image can be tolerated. On the other side, if the segmenting image is complex and the algorithm used is fully automatic, the error is inevitable.[27] The disparity between an actually segmented image and a correctly/ideally segmented image (reference image) that is the best expected result can be used to assess the performance of algorithms. Both (actually segmented and reference) images are obtained from the same input image. The reference image is sometimes called gold standard [*e.g.*, reference (27)]. In cases that the test images are synthetic images, the reference images can be simply obtained from image generation procedure,[28] while in cases that the test images are real images,

manually (with the help of visual inspection) segmented images are often used as references. The methods in this group take into account the difference (measured by various discrepancy parameters) between the actually segmented and reference images, *i.e.* these methods try to determine how far the actually segmented image is from the reference image. A higher value of the discrepancy measure would imply a bigger error in the actually segmented image relative to the reference image and this indicates the lower performance of applied segmentation algorithms.

In image encoding, the disparity between the original image and the decoded image has often been used to objectively assess the performance of coding algorithms. A commonly used discrepancy measure is the mean-square signal-to-noise ratio [see, *e.g.*, reference (29)]. However, in contrast to image encoding, image segmentation is a process that changes the image unit.[10] In other words, image encoding is an image processing process, while image segmentation is an image analysis process, in which the input and output are different matters. So many other discrepancy measures have been proposed and used.

4.1. Discrepancy based on the number of mis-segmented pixels

Considering image segmentation as a pixel classification process, the percentage of pixel mis-classified is the discrepancy measure that comes most readily to mind [30]. Suppose an image consists of N pixel classes, a confusion matrix C of dimension N can be constructed, with each entry C_{ij} represents the number of class j pixels classified as class i by the segmentation algorithms. Two error types can thus be computed for each pixel class k, which can both be used to describe the class-by-class performance of these algorithms [30]. The multi-class Type I error is defined as:

$$M_I^{(k)} = 100 \times \left[\left(\sum_{i=1}^{N} C_{ik} \right) - C_{kk} \right] \Big/ \left[\sum_{i=1}^{N} C_{ik} \right] \quad (8)$$

where the numerator represents the number of pixels of class k not classified as k and the denominator is the total number of pixels of class k.

The multi-class Type II error is defined as:

$$M_{II}^{(k)} = 100 \times \left[\left(\sum_{i=1}^{N} C_{ki} \right) - C_{kk} \right] \Big/ \left[\left(\sum_{i=1}^{N} \sum_{j=1}^{N} C_{ij} \right) - \sum_{i=1}^{N} C_{ik} \right] \quad (9)$$

where the numerator represents the number of pixels of other classes called class k. The denominator is the total number of pixels of other classes. In equations (8) and (9), each pixel class is weighted equally.

Weszka and Rosenfeld[21] used a similar approach to measure the difference between an "ideal" (correct) image and a thresholded image. Under the assumption that the image consists of objects and background each having a specified distribution of gray level, they compute, for any given threshold value, the probability of mis-classifying an object pixel as background, or *vice versa*. This probability in turn provides an index of segmentation results, which can be used for evaluating threshold selection algorithms. In their work, such a probability is minimized in the process of selecting an appropriate threshold.

Recently, a discrepancy measure based on the same principle has been defined. It is termed the probability of error (*PE*). For a two-class problem *PE* can be calculated by:[31]

$$PE = P(O) \times P(B|O) + P(B) \times P(O|B) \quad (10)$$

where $P(B|O)$ is the probability of error in classifying objects as background, $P(O|B)$ is the probability of error in classifying background as objects, $P(O)$ and $P(B)$ are *a priori* probabilities of objects and background in images. For multi-class problem, a general definition of *PE* can be found in reference (32).

The idea of computing discrepancy based on the number of error pixels is also reflected in some edge detection evaluation schemes. For example, a maximum likelihood estimate of the fraction of correctly detected edges has been used by Fram and Deutsch.[33] Such a measure could be readily extended to measure what fractions of the segmented object pixels were actually object pixels so as to be applied for segmentation evaluation.

4.2. Discrepancy based on the position of mis-segmented pixels

The discrepancy measures based only on the number of mis-segmented pixels do not take into account the spatial information of these pixels. It is thus possible that images segmented differently can have the same discrepancy measure values if these measures only count the number of mis-segmented pixels. To address this problem, some discrepancy measures based on pixel position error have been proposed.

One way is to use the distance between the mis-segmented pixel and the nearest pixel that actually belongs to the mis-segmented class. Let N be the number of mis-segmented pixels for the whole image and $d(i)$ be a distance metric from the ith mis-segmented pixel and the nearest pixel that actually is of the mis-classified class; a discrepancy measure (D) based on this distance is defined by Yasnoff *et al.* as:[30]

$$D = \sum_{i=1}^{N} d^2(i) \quad (11)$$

In equation (11), each distance is squared. This measure is further normalized (ND), to exempt the influence of image size and to give it a suitable value range by:[30]

$$ND = 100 \times \sqrt{D} \Big/ A \quad (12)$$

where A is the total number of pixels in the image (*i.e.* a measure of area).

In the evaluation of edge detectors a commonly used discrepancy measure is the mean-square distance figure of merit (*FOM*) proposed by Pratt:[34]

$$FOM = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{1+p\times d^2(i)} \tag{13}$$

where $N=max(N_i,N_a)$ and N_i and N_a denote the number of ideal and actually detected edge pixels, respectively, *d(i)* denotes the distance between the *i*th detected edge pixel and its correct position and *p* is a scaling parameter. This measure has been shown insensitive to correlation in false alarms and missed edges.[35] Strasters and Gerbrands used *FOM* for evaluating segmentation results with *N* denoting the number of pixel in image and *d(i)* denoting the distance between the *i*th pixel and its correct class.[36] In addition, they defined a modified version of *FOM* named FOM_e to expand the *FOM* value range in the near perfect segmentation:

$$FOM_e = \begin{cases} \frac{1}{N_e}\sum_{i=1}^{N_e}\frac{1}{1+p\times d^2(i)} & if \quad N_e > 0 \\ 1 & if \quad N_e = 0 \end{cases} \tag{14}$$

where N_e denotes the number of mis-segmented pixels.

4.3. Discrepancy based on the number of objects in the image

For perfect segmentation a necessary condition is that an equal number of objects of each class among a reference image and a segmented image should be met. A substantial disagreement of the object number indicates a large discrepancy between the reference and segmented images. Yasnoff and Bacus[37] proposed to compute the object-count-agreement *(OCA)* based on probability theory. Let R_i be the number of objects of class *i* in the reference image and S_i be the number of objects of class *i* in the segmented image, they use the probability F_{OCA} that the two numbers R_i and S_i represent samples from the same distribution for measuring the *OCA*:

$$F_{OCA} = \int_{L}^{\infty}\frac{1}{2^{M/2}\Gamma(M/2)}z^{(M-2)/2}e^{-z/2}dz \tag{15}$$

In equation (15), *M=N-1* denotes the number of degrees of freedom, *Γ(.)* denotes the Gamma function and *L* can be computed by:

$$L = \sum_{i=1}^{N}\frac{S_i - R_i}{p\times R_i} \tag{16}$$

where *N* is the number of object classes and *p* is a correction parameter.

On the basis of the similar idea, another weighting scheme called fragmentation (*FRAG*) is defined as:[36]

$$FRAG = \frac{1}{1+p\times\left|T_N - A_N\right|^q} \tag{17}$$

where T_N is the true object number in the reference image and A_N is the actual object number in the segmented image, *p* and *q* are scaling parameters.

4.4. Discrepancy based on the feature values of segmented objects

Image analysis is concerned with the extraction of information from an image, an image in yields data out.[38] Here the data are the measurement values of object features obtained from segmented images. One fundamental question in image analysis is whether a measurement made on the objects from segmented images is as accurate as one made on the original images. According to this measure, a segmented image has the highest quality if the object features extracted from it precisely match the features in the original. In practice, an image has high quality if the decision made on it is unchanged from that made on the original image.[39] The ultimate goal of image segmentation in the context of image analysis is to obtain measurements of object features.[11] The accuracy of these measurements obtained from the segmented image with respect to the reference image provides useful discrepancy measures. This accuracy can be termed "ultimate measurement accuracy" (*UMA*) to reflect the ultimate goal of segmentation. The *UMA* is feature dependent and so can be denoted as UMA_f. Let R_f denote the feature value obtained from the reference image and S_f denotes the feature value measured from the segmented image, the absolute UMA_f ($AUMA_f$) and relative UMA_f ($RUMA_f$) are defined as:[2]

$$AUMA_f = \left|R_f - S_f\right| \tag{18}$$

$$RUMA_f = \frac{\left|R_f - S_f\right|}{R_f}\times 100\% \tag{19}$$

Both $AUMA_f$ and $RUMA_f$ can represent a number of discrepancy measures when different object features are used. The features can be densitometric, statistic or geometric features. Some examples of geometric features are the area, bending energy, eccentricity, form factor, normalized mean absolute curvature, perimeter and sphericity of objects.[40] Among them, the area of objects is more suitable than others to appraise the quality of differently segmented images.[2,40]

4.5. Discrepancy based on miscellaneous quantities

There are other discrepancy measures that can describe the difference between the reference image and the segmented image. The discrepancy measure proposed by Levine and Nazif[41] is a 2-D (two-dimensional) distance measure based on two components. One is an under merging error measure and another is an over merging error measure. The former component is proportional to the amount by which the regions in the segmented image overlap the regions in the reference image. The latter component signifies the amount by which the segmented regions partition the reference regions.

Not only the spatial information, but also the gray-level information can be used to describe the difference between the segmented image and the reference image. Strasters and Gerbrands[36] defined a figure of certainty (*FOC*) for taking into account this information. Let f_i be the gray level of the *i*th pixel in the reference image and g_i be the representative gray level of a region comprising the *i*th pixel in the segmented image (note that both images are taken as masks here to extract f_i and g_i from the image to be segmented), the *FOC* is defined as:

$$FOC = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{1+p\times\left|f_i-g_i\right|^q} \tag{20}$$

where *N* denotes the total number of pixels in the image and *p* and *q* are scaling parameters.

If we consider both the segmented image and the reference image as probability distributions, the difference between them would be reflected by their divergence. Suppose that the segmented image has *N* regions and p'_i represents the *a posterior* probability of a pixel to be in the *i*th region, while p''_i is that in the reference image, Pal and Bhandari[25] proposed to use the symmetric divergence (*SD*):

$$SD = \sum_{i=1}^{N}\left(p'_i - p''_i\right) ln\frac{p'_i}{p''_i} \tag{21}$$

as a measure of performance for the segmentation algorithms.

5. COMPARISON OF METHOD GROUPS

The three method groups for segmentation evaluation described in the above sections have their own characteristics. In the following, their advantages and limitations are discussed.

5.1. Generality for evaluation

One desirable property of an evaluation method is its generality to be applied for studying various properties of different segmentation algorithms. To apply analytical methods some formal models of an image should be first defined. The behavior of the algorithm on such an image can then be analysed (mathematically) in terms of the parameters of the image and the algorithm.[42] Certain properties of segmentation algorithms can be easily obtained just by analysis, such as the processing strategy of algorithms and the resolution of segmentation results.[18] However, some other properties can not be precisely analysed since no formal model exists. For example, there is no quantitative measure for *a priori* knowledge about images that can be incorporated into segmentation algorithms,[14] so various types of knowledge are hardly to be compared. In addition, there are methods that can only be applicable to certain segmentation algorithms. For instance, the method based on detection probability ratio is merely suitable for studying simple edge detectors.[16]

Empirical methods, as described in Sections 3 and 4, are mainly used to study the correctness of segmentation algorithms by taking into account the accuracy of segmentation results. One reason is that other properties of algorithms, such as computation cost, have been partially overcome by the progress of technology. Another reason is that the accuracy of segmentation is often the primary concern in real applications and is difficult to be studied by analytical methods. From the point of view that only one property is studied, the empirical methods can be thought of as somewhat limited. However, most of them can be considered as relatively general, because they can evaluate different types of segmentation algorithms. The studies presented in reference (9, 23, 43-44) are some examples in which quite different types of algorithms are treated. In most empirical studies, only the images to segment and segmented are needed and no matter which type of algorithms is used. A few exceptions are the methods based on busyness[21] and shape measure.[8] Since the threshold value is necessary for calculating these measures other types of algorithms can not be evaluated.

5.2. Qualitative versus quantitative and subjective versus objective

Two more desirable properties of an evaluation method are the abilities to evaluate segmentation algorithms in a quantitative way and on an objective basis. Quantitative study can provide precise results reflecting the exactness of evaluation.[2] Objective study will exempt the influence of human factor and provide

consistency and no bias results.[38] Generally, analytical methods are more ready to apply, but they often provide only qualitative properties of algorithms. Empirical methods are normally quantitative as the values of quality measures can be numerically computed. Among them, goodness methods based on subjective measures of image quality are less suitable for an objective evaluation of segmented algorithms. Discrepancy methods can be both objective (the gold standard available yields objective results[27]) and quantitative.

5.3. Complexity for evaluation

The complexities for applying the above three groups of methods in segmentation evaluation increase progressively. Applying empirical methods for evaluation is usually more complicated than just algorithm analysis, because the algorithms are necessary to be concretely implemented and some extra efforts are needed to segment test images and to calculate the values of quality measure parameters. The computational cost of different empirical methods is first determined by the quality measures they used. For example, the object count agreement can be easily obtained, while the uniformity measure and shape measure need much more computation.

Among empirical methods, goodness methods are less complicated for applying than discrepancy methods and they can be used for on-line evaluation.[20] One particular requirement associated with the application of discrepancy methods is the reference image. Many studies use real images as test images and manually segment them to obtain the references [for example, see reference (31)]. This process greatly increases the complexity of applying discrepancy methods. In addition, since only real images from particular task domains were used in these studies, the evaluation results may be not appropriate for other applications. One possible and effective alternation is to use synthetic images.[10] The two problems associated with real images, as discussed above, can be overcome by using well-designed synthetic images. Other advantages of synthetic images include that they can be easily controlled and they can be reproduced by all users.[2, 28]

5.4. Consideration of segmentation applications

The effective use of domain-dependent knowledge in computer vision can help to make different processes reliable and efficient [see, for example, reference (45)]. To effectively evaluate segmentation algorithms, the consideration of segmentation applications in which algorithms are applied is also important.

The above three method groups are different in the extent to which they explicitly consider the applications for which the segmentation algorithms are used. At one extreme are the analytical studies that do not consider the nature and goal of application. The evaluation results depend only on the analysis of algorithms themselves. The empirical goodness methods in which some desirable properties of segmented images are quantified by goodness measures begin to address the application issue as the choice of which goodness measure should be used is related to the application goal. The empirical discrepancy methods, which take both the reference and segmented images into consideration, attempt to capture the application through the discrepancy measures. The need to have a reference forces the evaluation to be connected to the actual applications.[27]

6. COMPARISON OF SOME EMPIRICAL METHODS

In empirical studies the segmentation algorithm is applied to test images and statistics of its performance are gathered with the help of some measurements from segmentation results. Most empirical evaluation methods are developed independently and no comparison of performance or behavior with other methods has been made. Since a number of methods have been proposed, as described in the above sections, their comparison becomes important and necessary.

The performance of different empirical methods can be compared according to their behavior in judging the same sequences of segmented image. This sequence of images can be obtained by thresholding an image with a number of ordered threshold values.[2] As we know, the quality of thresholded images would be better if an appropriate threshold value is used and the quality of thresholded images would be worse if the selected threshold values are too high or too low. In other words, if the threshold value increases or decreases in one direction, the probability of erroneously classifying the background pixels as object pixels goes down, but the probability of erroneously classifying the object pixels as background pixels goes up, or *vice versa*. Since different evaluation methods using different measures to assess this quality, they will behave differently for the same sequence of images. By comparing the behavior of different methods in such a case, the performance of different methods can be revealed and ranked.

On the basis of this idea, a comparative study of different empirical methods has been carried out. The five methods studied (and the measures they based on) are the following:

(1) G-GU: Goodness based on Gray level Uniformity [see equation (3) in Subsection 3.1];
(2) G-GC: Goodness based on Gray level Contrast [see equation (6) in Subsection 3.2];
(3) D-PE: Discrepancy based on Probability of Error [see equation (10) in Subsection 4.1];
(4) D-ND: Discrepancy based on Normalized Distance [see equation (12) in Subsection 4.2];
(5) D-AA: Discrepancy based on Absolute UMA_f with Area as feature [see equation (18) in Subsection 4.4].

These five methods belong to five different method subgroups. They are considered for the comparative study mainly because the measures these methods based on are quite general for use and so are comparable. The methods in other subgroups and the measures they based on are less general. For example, the method based on

shape measure defined in equation (7) of Subsection 3.3 can only count the local smoothness of region boundary and cannot even distinguish a circle from a square.[2] On the other side, the measure based on the number of objects in the image is only meaningful when the segmentation results are quite poor. In near perfect segmentation, the number of objects in the reference image and segmented image are often the same and the discrimination power of this measure will be lost.

The whole experiment can be divided into several steps: define test images, segment test images, apply evaluation methods, measure quality parameters and compare evaluation results. It is arranged similar to the study of object features in the context of image segmentation evaluation.[11] A similar process has also been discussed by Haralick[10] for characterizing computer vision algorithms.

Test images are synthetically generated with the system described in reference (28). Since our main concern is to compare different evaluation methods with the same segmented images so some simple images are synthesized. They are 256x256 with 256 gray levels. The objects are centered discs of various sizes with gray level 144. The background is homogeneous with gray level 112. These images are then added by independent zero-mean Gaussian noise with various standard deviations. To cope with the random nature of noise, for each standard deviation five noise samples are generated independently and added separately to noise free images in this study. Five test images thus generated form a test group. Fig.2 gives an example.

Test images are segmented by thresholding them as described above. A sequence of 14 threshold values are taken to segment each group of images. The five evaluation methods are then applied to the segmented images. The values of their corresponding measures are obtained by averaging the results of five measurements over each group. In Table 1 comparison results of the five methods for one experiment are presented as examples. The labels correspond to the sequence of segmented images. In other words each column in Table 1 indicates a different threshold applied to a group of images. The measure values have been normalized to the range [0,1] for easy comparison. In Fig. 3 the curves corresponding to different measure values listed in Table 1 are plotted. These curves can be analysed by comparing their forms. Firstly, as the worst segmentation results give the value one for all measures, the valley values that correspond to the best segmentation results determine the margin between the two extremes. The deeper the valley, the larger the dynamic range of measures for assessing the best and worst segmentation results. Comparing the depth of valleys, these methods can be ranked in the order D-AA, D-PE, D-ND, G-GU, G-GC. Note that G-GC curve is almost unity for all segmented images (can be seen more clearly from Table 1), so that different segmentation results can hardly be distinguished in such a case.

Second, for evaluation purposes, a good method should be capable of detecting very small variations in segmented images. The sharper the curves, the higher the measure's discrimination capability to distinguish small segmentation degradation. The ranking of these five methods according to this point is the same as above. Looking more closely, though D-AA and D-PE curves are parallel or even overlapped for most cases in Fig. 3, the form of the D-AA curve is much sharper than that of D-PE curve near the valley. This means that D-AA has more power than D-PE to distinguish those slightly different and near-best segmentation results, which is more interesting in practice.[46] It is clear that D-AA should not be confused with D-PE as made by Beghdadi *et al.*[47] On the other side, the flatness of G-GC and G-GU curves around valley show that the methods based on goodness measures such as GC and GU should be less appropriate in segmentation evaluation.

The effectiveness of evaluation methods is largely determined by their employed image quality measures. From this comparative study, it becomes evident that the evaluation methods using discrepancy measures such as that based on the feature values of segmented objects and that based on the number of mis-segmented pixels should be more powerful than the evaluation methods using other measures. Moreover, as the methods compared in this study are representative of various method subgroups, it seems that the empirical discrepancy methods surpass the empirical goodness methods in evaluation.

7. FURTHER DISCUSSIONS

7.1. Special evaluation methods

There are also few particular evaluation methods that do not fall clearly in any one of the above three groups. The following is a critical review of them.

(1) For a general segmentation procedure, preprocessing and postprocessing are often needed (see Fig. 1). In practical applications, based on an automatically segmented image that is not perfect, some manually editing operations are often needed to bring the results to a certain level satisfying the desired quality.[27] The amount of such operations or the cost to do these operations can also provide an index of how the segmented image deviates from the desired quality. This index has been used by Graaf *et al.*[27,48] to validate segmentation results and to judge the performance of algorithms. Since the desired quality level of a segmentation is determined by the particular processing task, such a method makes a task-directed evaluation and depends on the tools available for image editing.[48] More generally, one tries to estimate the requirement for pre- and/or post-processing to obtain satisfactory segmentation results from the raw images.[18] In a sense, it is not the segmentation algorithms but the pre- and/or post-processing algorithms are studied.

(2) In image analysis the size of a region is obtained by counting the number of pixels belonging to this region.[38] The mis-segmented pixels modify the size of regions in segmented images. This size change can easily be observed by human eyes. Instead of defining numerical discrepancy measures MacAulay and Palcic proposed a qualitative evaluation method.[49] In their study for comparing four simple thresholding algorithms, a segmentation is determined to be acceptable if the area of segmented objects matches within a margin of 5% to the area of visually detected object. If a large number of images are processed, a statistic study of the results can help to compare the performance of the tested algorithms. This method is quite similar to the methods described in Subsection 4.1, except that the discrepancy is qualitatively and visually measured. Most subjective comparison studies are based on similar principles.

(3) To select an appropriate threshold value for segmentation Brink[50] proposed a thresholding technique that uses a gray-level correlation measure. An optimum threshold is selected by maximizing the correlation between the original image and the thresholded bi-level image. The value of correlation measure provides an index about the dissimilarity between these two images. This measure has been used in the evaluation of thresholding algorithms by Pal and Bhandari.[25] In contrast to discrepancy methods described in Section 4, this method takes the image to segment directly as the "reference" image. Although this correlation measure is seemed different in appearance than other measures, it has been proved[51] that the square of the correlation coefficient used in Brink's method is just the class separability quotient used by Otsu[26] in the "goodness" measure for threshold selection. This method should thus have a behavior similar to that based on inter-region contrast.

(4) Taking the image to segment as the reference is also followed by Beghdadi *et al.*[47] They proposed to use a measure termed the blurring effect for segmentation comparison. A noise-free synthetic image is generated and is then blurred with a Gaussian filter. The authors unusually set the blurred boundary pixels as object pixels and thus curiously take enlarged objects as references. The blurring effect is measured by the location difference between the detected boundary and the reference boundary. Such a use of synthetic images loses their advantages in evaluation. In addition, the noisy effect, a very important and common degradation factor influencing the performance of algorithms, cannot be studied by such a method.

(5) Different from all the above methods, Bryant and Bouldin[52] proposed another interesting evaluation procedure based on relative grading for edge detectors. The principle may be extended for evaluating segmentation algorithms. No precise quality measure or criterion is defined in this procedure. It consists of comparing the output of an algorithm to the consensus results of other algorithms. In other words, it compares the output of a number of algorithms and rates each algorithm by how often it agrees with the consensus of the others. This can be considered as an interesting idea, but it is unconscious to errors made by all algorithms and may even penalize a good algorithm that does not produce errors made by a majority of bad algorithms.[20]

7.2. Common problems for most existing methods

There are still two main problems associated with most of existing evaluation methods.

(1) Each evaluation method determines the performance of algorithms according to certain criteria. If the same criterion used for segmentation is also used for evaluation then some biased results will be produced.[2] For example, the second-order local entropy that was maximized for selecting threshold values in the new algorithm proposed by Pal and Pal[24] and was also computed for comparing the performance of this algorithm with that of other algorithms by Pal and Bhandari.[25] It is expected that the new algorithm should produce quite high entropy values. In many applications, images are modeled as a mosaic of regions of uniform intensity corrupted by additive Gaussian white noise [*e.g.* reference (53)]. Therefore the region homogeneity is a commonly used criterion for designing various segmentation algorithms [*e.g.*, Otsu algorithm [26]]. The method using the goodness measure based on uniformity takes the same criterion for evaluation. When this criterion is used to compare a number of thresholding algorithms,[8] it is not surprising that the Otsu[26] algorithm ranks at the first place. When other criteria were used, however, the ranking order becomes completely different.[8]

(2) To strengthen certain aspects in the quality measures, some scaling/weighting parameters are often used. For example, the parameter p in *FOM* [see equation (13)] provides a relative penalty between smeared edges and isolated but offset edges [34], while the parameters p and q in *FOC* [see equation (20)] determine the contribution of the large deviation relative to a small deviation.[36] There exists no suitable guideline or rule for choosing these parameters. In practice, they are often selected on the basis of human intuition or judgment. This makes an expected objective evaluation to be unpleasantly influenced by subjective factors.

8. CONCLUDING REMARKS

In this paper most methods proposed for segmentation evaluation and comparison so far are reviewed. A method classification scheme is introduced. Comparative studies for different method groups and for different methods are also carried out, both analytically and experimentally. Segmentation evaluation is indispensable for

improving the performance of existing segmentation algorithms and for developing new powerful segmentation algorithms. This study attempts to stimulate the work in this direction. To make segmentation get off trial-and-error status further studies and more efforts for segmentation evaluation are needed.

From this study some results concerning the performance of different evaluation methods are obtained. As there is currently no general segmentation theory, the empirical methods are more suitable and useful than the analytical methods for performance evaluation of segmentation algorithms. Among empirical methods, the discrepancy methods are better for objectively assessing segmentation algorithms than the goodness methods, although the former is somewhat complex in application than the latter due to the requirement for reference. According to the experimental comparison made in this paper, the method D-AA is more powerful for evaluation than other methods. More general studies are still carrying on.

Each method studied in this paper has advantages and limitations. From an application point of view, those that belong to different groups are more complementary than competitive. Besides, the performance of segmentation algorithms is influenced by many factors, so only one evaluation method would be not enough to judge all properties of an algorithm and different methods should be cooperated. One early work of this type is made by Yasnoff *et al.*,[54] who combined two error measures they proposed, namely pixel spatial distribution and pixel class proportion,[30] into one generalized measure. Later they incorporated another component, the object count agreement[37] together. Other evaluation studies using several measures can be found in references (23,25,36,43,44). Generally, for a complete evaluation and comparison of segmentation techniques, a set of performance measures should be necessary[9,18]. How to form such a set will be a promising research subject in segmentation evaluation.

ACKNOWLEDGEMENT

We are very grateful to the reviewer for his helpful comments and valuable suggestions to improve the presentation of this paper.

REFERENCES

1. T.Pavlidis, Image analysis, *Ann. Rev. Comput. Sci.* **3**, 121-146 (1988).
2. Y.J.Zhang and J.J.Gerbrands, Objective and quantitative segmentation evaluation and comparison, *Signal Processing* **39**, 43-54 (1994).
3. E.M.Riseman and M.A.Arbib, Survey: computational techniques in the visual segmentation of static scenes, *CGIP* **6**, 221-276 (1977).
4. J.S.Weszka, A survey of threshold selection techniques, *CGIP* **7**, 259-265 (1978).
5. K.S.Fu and J.K.Mui, A survey on image segmentation, *Pattern Recognition* **13**, 3-16 (1981).
6. R.M.Haralick and L.G.Shapiro, Survey: image segmentation techniques, *CVGIP* **29**, 100-132 (1985).
7. V.I.Borisenko, A.A.Zlatotol and I.B.Muchnik, Image segmentation (state of the art survey), *Automat. Remote Control* **48**, 837-879 (1987).
8. P.K.Sahoo, S.Soltani, A.K.C.Wong and Y.C.Chen, A survey of thresholding techniques, *CVGIP* **41**, 233-260 (1988).
9. N.R.Pal and S.K.Pal, A Review on image segmentation techniques, *Pattern Recognition* **26**, 1277-1294 (1993).
10. R.M.Haralick, Performance characterization in computer vision, *CVGIP-IU* **60**, 245-249 (1994).
11. Y.J.Zhang and J.J.Gerbrands, Segmentation evaluation using ultimate measurement accuracy, *SPIE* **1657**, 449-460 (1992).
12. R.M.Haralick and L.G.Shapiro, *Computer and Robot Vision*, Addison-Wesley, New York (1992).
13. A.Rosenfeld and L.S.Davis, Image segmentation and image models, *Proc. IEEE* **67**, 764-772 (1979).
14. C.E.Liedtke, T.Gahm, F.Kappei and B.Aeikens, Segmentation of microscopic cell scenes, *AQCH* **9**, 197-211 (1987).
15. Y.J.Zhang and J.J.Gerbrands, Transition region determination based thresholding, *Pattern Recognition Letters* **12**, 13-23 (1991).
16. I.E.Abdou and W.K.Pratt, Quantitative design and evaluation of enhancement/thresholding edge detectors, *Proc. IEEE* **67**, 753-763 (1979).
17. C.Garbay, Image structure representation and processing: a discussion of some segmentation methods in cytology, *IEEE Trans.* **PAMI-8**, 140-146 (1986).
18. Y.J.Zhang, Comparison of segmentation evaluation criteria, *Proc. 2ICSP*, 870-873 (1993).
19. J.J.Gerbrands, *Segmentation of noisy images*, Doctoral Thesis, Delft University of Technology, Delft, The Netherlands, (1988).
20. M.D.Levine and A.Nazif, Dynamic measurement of computer generated image segmentations, *IEEE Trans.* **PAMI-7**, 155-164 (1985).
21. J.S.Weszka and A.Rosenfeld, Threshold evaluation techniques, *IEEE Trans.* **SMC-8**, 622-629 (1978).
22. R.M.Haralick, K.Shanmugam and I.Dinstein, Textural features for image classification, *IEEE Trans.* **SMC-3**, 610-622 (1973).

23. A.M.Nazif and M.D.Levine, Low level image segmentation: an expert system, *IEEE Trans.* **PAMI-6**, 555-577 (1984).
24. N.R.Pal and S.K.Pal, Entropic thresholding, *Signal Processing* **16**, 97-108 (1989).
25. N.R.Pal and D.Bhandari, Image thresholding: some new techniques, *Signal Processing* **33**, 139-158 (1993).
26. N.Otsu, A threshold selection method from gray-level histogram, *IEEE, Trans.* **SMC-9**, 62-66 (1979).
27. C.N.Graaf, A.S.E.Koster, K.L.Vincken and M.A.Viergever, Validation of the interleaved pyramid for the segmentation of 3D vector images, *Pattern Recognition Letters* **15**, 467-475 (1994).
28. Y.J.Zhang and J.J.Gerbrands, On the design of test images for segmentation evaluation, *Proc. EUROSCO-92* **1**, 551-554 (1992).
29. R.C.Gonzalez and P.Wintz, *Digital Image Processing*, Addison-Wesley, New York (1987).
30. W.A.Yasnoff, J.K.Mui and J.W.Bacus, Error measures for scene segmentation, *Pattern Recognition* **9**, 217-231 (1977).
31. S.U.Lee, S.Y.Chung and R.H.Park, A comparative performance study of several global thresholding techniques for segmentation, *CVGIP* **52**, 171-190 (1990).
32. Y.W.Lim and S.U.Lee, On the color Image segmentation algorithms based on the thresholding and fuzzy c-means techniques, *Pattern Recognition* **23**, 935-952 (1990).
33. J.R.Fram and E.S.Deutsch, On the quantitative evaluation of edge detection schemes and their comparison with human performance, *IEEE Trans.* **C-24**, 616-628 (1975).
34. W.K.Pratt, *Digital Image Processing*, John Wiley and Sons, New York (1978).
35. F.Heyden, Evaluation of edge detection algorithms, *Proc. 3ICIPA*, 618-622 (1989).
36. K.Strasters and J.J.Gerbrands, Three-dimensional image segmentation using a split, merge and group approach, *Pattern Recognition Letters* **12**, 307-325 (1991).
37. W.A.Yasnoff and J.W.Bacus, Scene segmentation algorithm development using error measures, *AQCH* **6**, 45-58 (1984).
38. I.T.Young, Sampling density and quantitative microscopy, *AQCH* **10**, 269-275 (1988).
39. P.C.Cosman, R.M.Gray and R.A.Olshen, Evaluating quality of compressed medical images: SNR, subjective rating, and diagnostic accuracy, *Proc. IEEE* **82**, 919-932 (1994).
40. Y.J.Zhang, Influence of image segmentation over feature measurement, *Pattern Recognition Letters* **16**, 201-206 (1995).
41. M.D.Levine and A.Nazif, An experimental rule based system for testing low level segmentation strategies, in *Multi-Computers and Image Processing: Algorithms and Programs*, K.Preston and L.Uhr eds., pp.149-160. Academic Press, New York (1982).
42. L.J.Kitchen and J.A.Malin, The effect of spatial discretization on the magnitude and direction response of simple differential edge operators on a step edge, *CVGIP* **47**, 243-258 (1989).
43. Y.J.Zhang, Image synthesis and segmentation comparison, *Proc. 3ICYCS*, 8.21-8.24 (1993).
44. Y.J.Zhang, Segmentation evaluation and comparison: a study of various algorithms, *SPIE* **2094**, 801-812 (1993).
45. Y.Shirai, *Three-Dimensional Computer Vision*, Spinger-Verlag, Berlin (1987).
46. Y.J.Zhang and J.J.Gerbrands, Comparison of thresholding techniques using synthetic images and ultimate measurement accuracy, *Proc. 11ICPR* **3**, pp.209-213 (1992).
47. A.Beghdadi, A.Negrate and P.V.Lesegno, Entropic thresholding using a block source model, *GMIP* **57**, 197-205 (1995).
48. C.N.Graaf, A.S.E.Koster, K.L.Vincken and M.A.Viergever, Task-directed evaluation of image segmentation methods, *Proc. 11ICPR* **3**, 219-222 (1992).
49. C.MacAuley and B.Palcic, A comparison of some quick and simple threshold selection methods for stained cells, *AQCH* **10**, 155-164 (1988).
50. A.D.Brink, Gray-level thresholding of images using a correlation criterion, *Pattern Recognition Letters* **9**, 335-341 (1989).
51. I.Cseke and Z.Fazekas, Comments on gray-level thresholding of images using a correlation criteria, *Pattern Recognition Letters* **11**, 209-210 (1990).
52. D.J.Bryant and D.W.Bouldin, Evaluation of edge operators using relative and absolute grading, *Proc. IEEE PRIP*, 138-145 (1979).
53. P.C.Chen and T.Pavlidis, Image segmentation as an estimation problem, *CGIP* **12**, 153-172 (1980).
54. W.A.Yasnoff, W.Galbraith, J.W.Bacus, Error measures for objective assessment of scene segmentation algorithms, *AQC* **1**, 107-121 (1979).

ILLUSTRATIONS

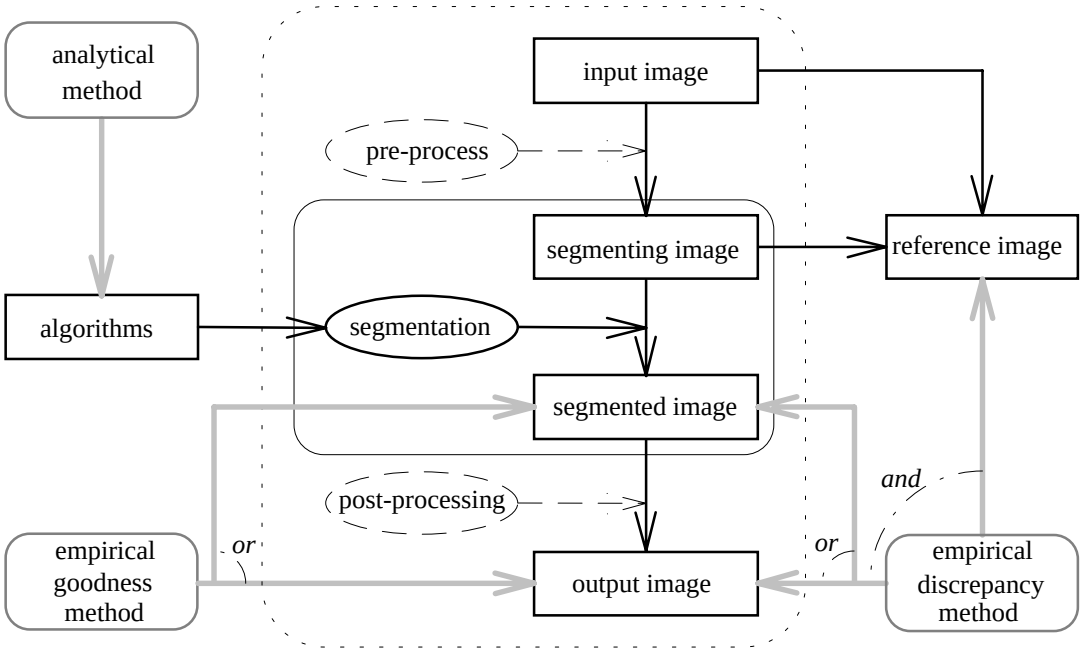

Fig.1: General scheme for segmentation and its evaluation.

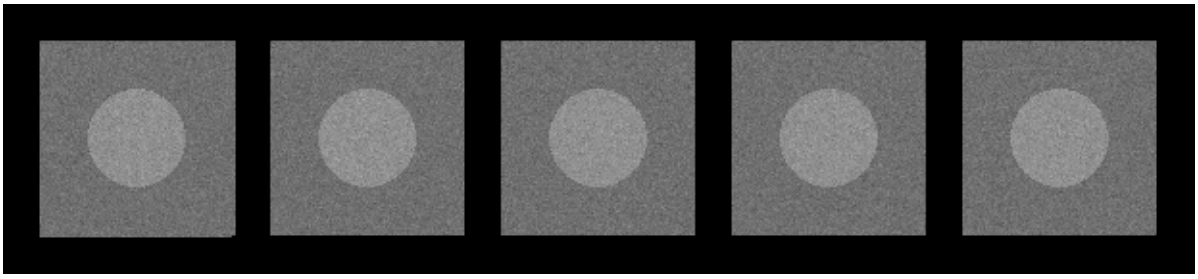

Fig.2: A group of test images.

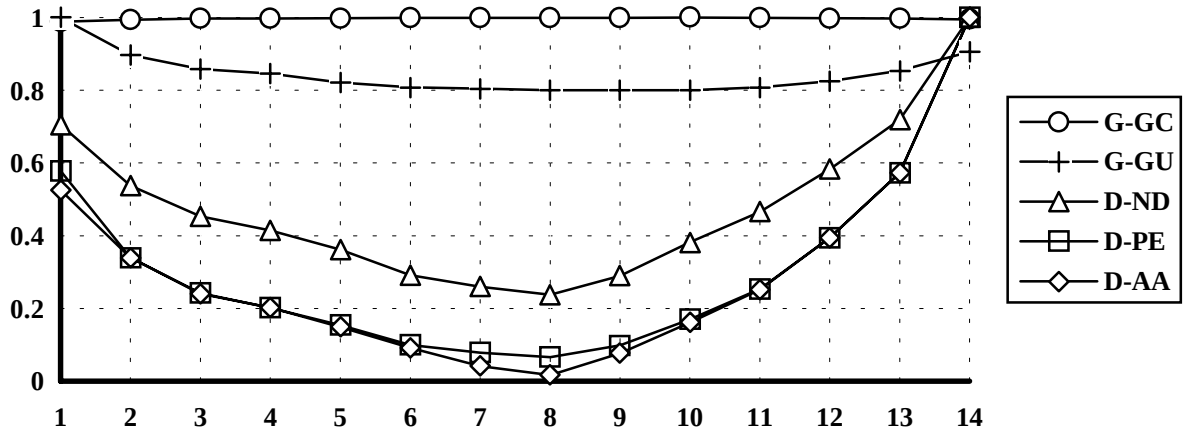

Fig.3: Plot of the comparison results listed in Table 1.

Table 1: Comparison results of different evaluation methods.

Label	1	2	3	4	5	6	7	8	9	10	11	12	13	14
G-GC	0.989	0.994	0.997	0.997	0.998	0.999	0.999	0.999	0.999	1.000	0.999	0.998	0.997	0.995
G-GU	1.000	0.897	0.858	0.846	0.821	0.808	0.804	0.800	0.800	0.800	0.808	0.825	0.854	0.906
D-ND	0.705	0.538	0.454	0.415	0.362	0.292	0.260	0.238	0.290	0.382	0.466	0.583	0.719	1.000
D-PE	0.578	0.340	0.242	0.202	0.154	0.100	0.079	0.066	0.099	0.170	0.254	0.395	0.573	1.000
D-AA	0.526	0.340	0.241	0.203	0.149	0.092	0.042	0.017	0.077	0.161	0.252	0.395	0.573	1.000