Artificial Intelligence Review (2020) 53:5637–5674
https://doi.org/10.1007/s10462-020-09830-9

Image segmentation evaluation: a survey of methods

Zhaobin Wang[1] · **E. Wang**[1] · **Ying Zhu**[2]

Published online: 18 April 2020

Abstract
Image segmentation is a prerequisite for image processing. There are many methods for image segmentation, and as a result, a great number of methods for evaluating segmentation results have also been proposed. How to effectively evaluate the quality of image segmentation is very important. In this paper, the existing image segmentation quality evaluation methods are summarized, mainly including unsupervised methods and supervised methods. Based on hot issues, the application of metrics in natural, medical and remote sensing image evaluation is further outlined. In addition, an experimental comparison for some methods were carried out and the effectiveness of these methods was ranked. At the same time, the effectiveness of classical metrics for remote sensing and medical image evaluation is also verified.

Keywords Image segmentation · Segmentation evaluation · Unsupervised evaluation · Supervised evaluation · Evaluation application

1 Introduction

Image segmentation is a very important and difficult problem in many fields such as image processing, pattern recognition and artificial intelligence. The primary and important key step in computer vision technology is image segmentation, which is also an important part of image semantic understanding. Correct image processing is impossible without proper segmentation, so image segmentation is an important image analysis technique in different areas and it is applied widely, especially in medical image analysis (Domingo et al. 2016; Zhou et al. 2020; Kaya et al. 2017; Li et al. 2020; Goceri et al. 2015; Goceri and Songul 2017b, 2018; Goceri 2018) for anomaly detection, disease diagnosis or monitoring. Similarly, segmentation technology is also indispensable in remote sensing image processing (Wang et al. 2020; Zhang et al. 2020; Peng et al. 2019; Zeng et al. 2019; Nogueira et al.

✉ Zhaobin Wang
zhaobin_wang@hotmail.com

✉ Ying Zhu
zhuying_365@126.com

[1] School of Information Science and Engineering, Lanzhou University, Lanzhou 730000, China

[2] Key Laboratory of Microbial Resources Exploitation and Application of Gansu Province, Institute of Biology, Gansu Academy of Sciences, Lanzhou, China

2019; Wu et al. 2019). In the research and application of images, people tend to be only interested in certain parts of the image (Göçeri 2013; Goceri 2016; Goceri and Songül 2017a; Goceri et al. 2017; Goceri 2019b). These parts are often referred to goals or prospects (other parts are called backgrounds). Goals or prospects generally refer to certain areas of an image with useful properties. In order to identify and analyze the targets in the image, these must be separated and extracted from the image. On this basis, the target can be further measured and the image can be further utilized. Image segmentation is the technique and process of dividing an image into regions with specific characteristics and extracting the target of interest, which is a key step from image processing to image analysis. Image segmentation is applied in many aspects. After the image is segmented, the segmentation result is evaluated so that the segmentation algorithm can be distinguished or the parameters of the algorithm can be adjusted according to the evaluation result. More importantly, the image can be processed in the next step only after the optimal segmentation image is selected by the evaluation algorithm.

The rest of this paper is organized as follows. Section 2 classifies and introduces the assessment methods. Section 3 introduces the application of assessment methods in three aspects of nature, medicine and remote sensing images. Section 4 introduces the metric selection for medical and remote sensing images segmentation evaluation. In Sect. 5, experiments are conducted on both supervised and unsupervised evaluation metrics. These experiments evaluate two different types of assessment indicators on different images. At the same time, the effectiveness of classical metrics for remote sensing and medical image evaluation is also verified. The experimental results are further analyzed and discussed in Sect. 6. Section 7 summarizes this paper.

2 Evaluation methods

In recent years, as the importance of segmentation assessment has increased, more and more segmentation evaluation algorithms have been proposed. The segmentation evaluation algorithm can be divided into the following categories according to Zhang (1996) and Chen et al. (2018), as shown in Fig. 1.

2.1 Subjective evaluation

Subjective methods can be also called observation method. The most convenient method of assessment is subjective assessment, where segmentation results are judged by human evaluators. The disadvantage of this approach is that visual or qualitative assessments

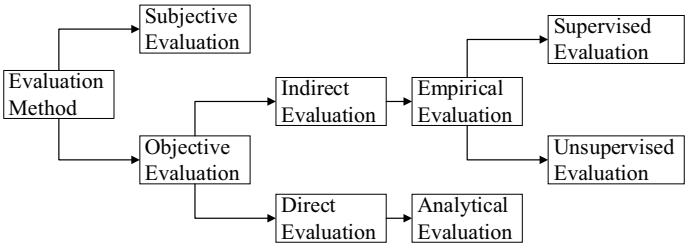


Fig. 1 The hierarchy of evaluation criteria

are subjective in nature. From one human evaluator to another, the subjective assessment scores can vary greatly, as each evaluator has his/her own unique criteria to assess the quality of the segmented image. This method requires a visual assessment study of a large number of subjects. The number of test images must be large enough to represent the image category for which the segmentation algorithm is directed. Similarly, human evaluators must be large enough to represent a typical human observer. Therefore, subjective assessment is a very cumbersome and time consuming process.

2.2 Direct evaluation and indirect evaluation

Direct evaluation is the evaluation of the segmentation algorithm itself. A segmentation problem can be solved by a variety of algorithms, and people prefer a simple and efficient algorithm. How to measure the quality of the segmentation algorithm is an important basis for image processing. Initially, the required computation time is used to measure the quality of an algorithm, but different machines cannot be compared with each other, so objective measurement independent of the specific computer is required. Nowadays, the size of the commonly used problem, the basic operation and the calculation function of the algorithm directly evaluate the pros and cons of the algorithm, that is, the time complexity and spatial complexity of the algorithm affect its performance.

Indirect evaluation is done by means other than the algorithm itself. For example, a segmentation algorithm segments any image to obtain a segmentation result graph. The performance of the segmentation result is judged by calculation or experiment, thereby indirectly obtaining the performance of the algorithm. It is worth noting that algorithms are complex and the same algorithm may have different results for different applications. Indirect evaluation is to judge the performance of an algorithm through a specific application, and is not universal.

2.3 Analytical evaluation and empirical evaluation

Analytical evaluation is a kind of direct evaluation. This method does not evaluate the algorithm through experiments or specific applications, but considers the performance of the algorithm based on theoretical knowledge. This method is usually used when the two algorithms have large differences in computational complexity or speed in operation. Analytical assessment can be divided into two categories: quantitative assessment and qualitative assessment (Chen et al. 2018).

Empirical evaluation is an indirect assessment that uses some specific operations to quantitatively evaluate the algorithm. Empirical assessment can be divided into supervised assessment and unsupervised assessment depending on whether a reference image is needed. Empirical evaluation, with high accuracy and precision, is currently the most widely used assessment method.

2.4 Supervised evaluation and unsupervised evaluation

Supervised evaluation methods are also called relative evaluation methods or empirical difference methods, they evaluate the segmentation algorithm by comparing the obtained segmented image with a manually segmented reference image, which is often referred to as a gold standard or ground truth. The similarity between the reference image and the

segmented image determines the quality of the segmented image. A potential benefit of the supervised approach is existing direct comparison between the segmented image and the reference image which could provide finer accuracy of the evaluation. However, manually generating reference images is a difficult, subjective and time consuming task. In addition, people generally cannot guarantee that reference images could be made manually for most images, especially for natural images, and there is also no guarantee that the reference images manually generated by one expert will outperform the other's. In this sense, the comparison based on such reference images is somewhat subjective.

In the field of supervised segmentation evaluation, understanding of True Positive(TP), False Positive(FP), True Negative(TN) and False Negative(FN) is indispensable. The understanding of classical evaluation algorithms is mainly based on TP, FP, TN and FN, and the interpretation of some formulas in this paper require these terms as the basis. We simply explain the meaning of these abbreviations through Fig. 2, definitions of *TP*, *FP*, *TN* and *FN* can be found in Taha and Hanbury (2015) and Chang et al. (2009).

As shown in Fig. 2, the ellipse area represents the ground truth, the triangle area represents the segmentation result. *TP* is the true positive, that is, the segmentation result is 1, the ground truth is 1, and it is represented by yellow. *TN* is the true negative, that is, the segmentation result is 0, and the ground truth is 0, which is represented by white. *FP* is the false positive, that is, the segmentation result is 1, and the ground true is 0, which is represented by green. *FN* is the false negative, that is, the segmentation result is 0, the ground truth is 1, and it is represented by red.

Unsupervised evaluation methods, also known as independent evaluation methods or empirically goodness methods, do not require reference images, and segmentation images are evaluated by calculating human-recognized criteria that represent good segmentation results. Unsupervised evaluation is quantitative and objective, it has obvious advantages, the most important one is that it does not require a reference image. Manually created reference images are subjective in nature. In general, creating reference images is complex and time consuming, and sometimes difficult or even impossible for many applications. Unsupervised evaluation methods can evaluate many different types of image segmentation without a reference image. However, due to the lack of comparison with reference images, one drawback of the unsupervised method over the supervised method is that it does not provide finer evaluation accuracy.

2.5 Evaluation methods based on neural networks

Deep learning or neural networks are widely used in image segmentation, but there are few evaluation methods using neural networks or deep learning knowledge. Convolutional Neural Networks (CNN) are important tool for most machine learning users

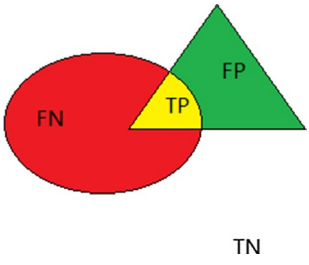


Fig. 2 The graphical representation of TP, FP, TN and FN

today (Vedaldi and Lenc 2015; Zagoruyko and Komodakis 2015; Goceri 2019a, c). Recently, the CNN-based segmentation method has shown superior performance (Yan et al. 2018; Goceri and Dura 2015a; Goceri and Goceri 2015b; Poudel et al. 2018). However, CNN-based effective segmentation evaluation models are few and still under development. Zhang et al. (2006) introduced a Meta-Segmentation Evaluation Technique (MSET). This technique uses a stand-alone method to evaluate the quality of the segmentation, it combines the basic evaluation according to a weighting function (depending on the image to be segmented) and any set of basic evaluation methods is combined by machine learning algorithms. Huang et al. (2016) proposed three types of deep neural networks to automatically learn deep features, thereby using deep neural networks for target segmentation evaluation. The first network uses segmented objects as input first, and then uses a general network. The second network uses a Full Convolutional Network (FCN) to predict the foreground mask of the input image and uses Intersection over Union (IoU) to evaluate the segmentation quality. The third network uses weighted mask layers to extract mask features to use both foreground and background information. Chen and Zhu (2019) proposed a new Relative Quality Prediction Network (RQPN) based on CNN as a new objective evaluation criterion and used a robust regression mapping model to find the relationship between subjective evaluation and objective distance. Shi et al. (2017) proposed a new multi-scale target segmentation quality assessment method based on CNN network, which combines global information and local information with multi-scale information of images to better evaluate segmentation quality.

CNN is currently mainly used for image classification (Zhao et al. 2020), object detection (Xia et al. 2016), and image segmentation. There are few applications of convolutional networks in segmentation evaluation. Traditional methods of object segmentation evaluation mainly adopt manual calculation methods, such as region-based, boundary-based and mixed measures. CNN can learn the relationship between input and output with sufficient data samples. So given the source image and the corresponding segmentation, you can try to evaluate the quality of the segmentation by training the network. We believe that more researchers will try to use the deep learning framework for object segmentation quality assessment in the future.

As can be seen from the above analysis, there are many types of segmentation evaluation methods, each of which has its own advantages and disadvantages. In the specific application, it should be considered from the characteristics of the evaluation method to obtain the best evaluation result. The most widely used assessment methods are empirical assessments, including supervised assessments and unsupervised assessments. This paper mainly summarizes the relevant indicators of supervised evaluation and unsupervised evaluation and introduces their application in natural, medical and remote sensing images. In this paper S represents segmentation result and GT represents ground truth.

3 Applications of image segmentation evaluation methods

In most areas of digital image processing, image segmentation has a wide range of applications, so segmentation evaluation will also be essential in various fields. Based on the wide range of applications, we mainly introduce the application of segmentation evaluation on natural, medical and remote sensing images.

3.1 Natural image segmentation evaluation

Natural images are closely related to human life, as the basis of natural image analysis, natural image segmentation plays an important role in the field of natural image processing. It can be said that good segmentation can lay a solid foundation for subsequent natural image processing. Therefore, it is important to evaluate the quality of the segmentation results. People initially rely on vision to evaluate the segmentation results, but this process requires a lot of manpower. Only when the number of people participating in the assessment experiment reaches a certain amount can a more accurate result be obtained. Human assessment has subjective bias to some extent, and assessment results are influenced by personal experiences, life and emotions. At the same time, people can't make accurate judgments on segmentation results that look very similar.

3.1.1 Supervised evaluation methods

The supervised segmentation evaluation method compares the gold standard image with the segmentation result, which makes the evaluation has a reference. This method overcomes limitations of visual evaluation and improves the accuracy of evaluation, so it has been widely used. Reviewing a lot of literatures, we find that researchers have proposed many effective metrics for evaluating image segmentation, among which the common methods include three categories: pixel-based evaluation, region-based evaluation and distance-based evaluation. Almost all metrics belong to above three categories. For example, region-based metrics include the following (Pont-Tuset and Marques 2016): Directional Hamming distance(DH), van Dongen distance(d_{vD}), Segmentation covering(SC), Bipartite Graph Matching(BGM), Bidirectional Consistency Error(BCE), Global Consistency Error (GCE) (Taha and Hanbury 2015; Garcia-Lamont et al. 2018; Dey et al. 2018; Mageswari and Mala 2014) and Variation of Information(VOI). Pixel-based metrics include the following: False Positive Rate (FPR), False Negative Rate(FNR) (Taha and Hanbury 2015; Dey et al. 2018), True Positive Rate (TPR)(=Sensitivity /Recall) (Taha and Hanbury 2015; Dey et al. 2018; Lukac et al. 2011; Chouhan et al. 2018), Boundary Displacement Error(BDE) (Garcia-Lamont et al. 2018), Matthews Correlation Coefficient(MCC), True Negative Rate (TNR)(=Specificity), F-measure (FMS) (Garcia-Lamont et al. 2018), Dice(=FMS), Jaccard index (JAC), Cohens kappa (KAP) (Taha and Hanbury 2015; Dey et al. 2018; Chouhan et al. 2018) and Precision (Lukac et al. 2011). Distance-based metrics include the following: Hausdorff Distance(HD), Average Hausdorff Distance(AVD), Mahalanobis Distance (MHD) (Taha and Hanbury 2015; Dey et al. 2018) and directional hamming distance(dhd) (Pont-Tuset and Marques 2016, 2013).

Of course there are other different classifications, Taha and Hanbury (2015) have divided the evaluation indicators into the following six categories. (1) Spatial overlap

based metrics: Dice, JAC, Recall = Sensitivity = TPR, Specificity = TNR, FPR, FNR, Precision = Positive Predictive Value (PPV), FMS, GCE. (2) Volume based metrics: Volume metric Similarity (VS). (3) Pair counting based metrics: Rand Index (RI), Adjusted Rand Index(ARI). (4) Information theoretic based metrics: Mutual Information (MI), VOI. (5) Probabilistic metrics: The Interclass Correlation (ICC), The Probabilistic Distance(PBD) (Taha and Hanbury 2015; Dey et al. 2018), KAP, The Area Under the ROC curve (AUC). (6) Spatial distance based metrics: HD, AVD and MHD.

As the research progresses, more evaluation indicators are proposed, for example, Regional Area Error(RAE) (Chouhan et al. 2018), Average Precision(AP) and IoU (Shan 2018). ARI, PBD, Probabilistic Random Index(PRI) (Garcia-Lamont et al. 2018), Mean Average Precision (MAP) (Henderson and Ferrari 2017) and Normalized Probabilistic Rand index (NPR) (Unnikrishnan et al. 2007) are improvements to previous indicators. MAP is the average AP value, which is the average AP value of multiple verification set individuals, as an index to measure the detection accuracy in object detection. The NPR has the following basic advantages: it does not favor special segmentation, and it does not assume the value of any data. It is used to quantitatively compare image segmentation algorithms using hand-marked ground truth sets. NPR allows for adaptive adjustment and normalization of the results to determine whether there is comparability between the segmentation algorithm and the image and gives a comparison score.

Recent panoramic segmentation metrics are also important. Panoptic segmentation assigns a category Label and instance ID to each pixel in the image to generate a global, uniform segmented image (Liu et al. 2019). The Facebook AI Research (FAIR) team set a new evaluation standard Panoptic Quality (PQ) (Kirillov et al. 2019) for panoramic segmentation. PQ is composed of Segmentation Quality (SQ) and Recognition Quality (RQ). Among them, RQ is a widely used F-measure in detection, which is used to calculate the accuracy of object recognition for each instance in panoramic segmentation. SQ represents the IOU between the predicted segment and the labeled after matching. Only when the IOU of the predicted segment and the labeled segment is strictly greater than 0.5, the two segments are considered to be matched.

In addition, there are some metrics that are proposed earlier but commonly used, which are expressed as follows. Symmetric partition Distance (d_{sym}), Asymmetric Partition Distance (d_{asy}) and Mutual partition distance (d_{mut}) (Cardoso and Corte-Real 2005) are measures of the distance between segmented partitions to overcome some limitations of existing methods, replacing symmetric and asymmetric distance metrics to meet the specificities of various applications. Roman-Roldan et al. (2001) proposed a quality measurement method(R), this technique combines the difference between the edge of the segmentation result and the theoretical edge with each error indicated by the human observer, thereby reducing the estimation error of the segmentation result. The contradiction between universality and particularity is one of the main challenges in segmentation assessment, and indicators that perform well on a particular application may not scale to other applications. In general, image segmentation is considered to identify some of the most important and useful structures from the image. Ge et al. (2006) proposed a benchmark for evaluating image segmentation, which collects various images, and constructs a simple and unambiguous ground truth for each image. This method solves two important problems: some of the most prominent structures in an image may not be unique and many image segmentation methods cannot directly extract a single protruding structure.

A variety of evaluation frameworks and metrics have also been proposed in recent years. Theories of these methods are diverse, including areal image matching, patch images,

sampling theory, segmentation data sets and targets, precision-recall curve, full-reference objective measure, discrepancy measure and hybrid ground truth fusion technique. Berezsky et al. (2016) analyzed the region image matching algorithm for evaluating the segmentation algorithm in the Gromov–Hausdorff metric. Ledig et al. (2014) introduced Patch-based Evaluation of Image Segmentation (PEIS), this method estimate segmentation bias based on finding patch correspondences and associated patch displacements. Compared with the Dice score, PEIS produces more comparable scores, increases sensitivity and accurately estimates segmentation bias. Arhid et al. (2016) proposed a new method based on sampling theory to calculate the difference between automatic segmentation and ground truth to evaluate the three-dimensional segmentation algorithm. Shi et al. (2014, 2015) proposed a novel full-reference objective measure named $Sim(G, S)$:

$$Sim(G,S) = \frac{A(G \cap S) + \sum_{i \in ab} C(i)}{A(G \cap S) - \sum_{j \in bh} C(j) + \sum_{j \in bh} P(j) + \sum_{k \in ih} P(k)} \tag{1}$$

This method compensates or penalizes the Gelasca's four basic errors types (see in Fig. 3) based on human visual tolerance, making the evaluation of image segmentation results more accurate. Yang et al. (2015) proposed a new discrepancy measure of Segmentation Evaluation Index (SEI), which uses a two-sided 50% overlap instead of the commonly used one-sided 50% overlap to redefine the corresponding segment. Unlike most existing measures, SEI emphasizes the importance of object recognition in segmentation assessment.

$$SEI_L(i) = \begin{cases} \sqrt{\frac{(1-\frac{area(g_i \cap s_i)}{area(g_i)})^2 + (1-\frac{area(g_i \cap s_i)}{area(s_i)})^2}{2}} & s_i \in S_{ds} \\ 1 & otherwise \end{cases} \tag{2}$$

$$SEI = \frac{1}{n} \sum_{i=1}^{n} SEI_L(i) \tag{3}$$

Pont-Tuset and Marques (2016, 2013) described new precision-recall frameworks: the precision-recall for regions (P_r, R_r, and F_r), the precision-recall for boundaries (P_b, R_b, and F_b) and the precision-recall for objects and parts (P_{op}, R_{op}, and F_{op}). P_r, R_r and P_b, R_b are metrics based on pairs-of-pixels and boundary respectively. The definition of the precision-recall for objects and parts is as follows:

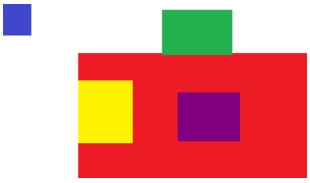


Fig. 3 Illustrations of one combination of Gelasca's four basic error types. The red rectangle represents the ground truth and the other colorful rectangles represent the segmentation errors: added region (blue), added background (green), border hole (yellow), inside hole (purple). (Color figure online)

$$P_{op} = \frac{on + fr + \alpha\ pn}{|S|} \tag{4}$$

$$R_{op} = \frac{on' + fr' + \alpha\ pn'}{|GT|} \tag{5}$$

Taha et al. (2014) proposed a formalization method based on the segmented data and the target of the segmentation task to select the most appropriate metric, the proposed method depends on measuring the metric bias of the target attribute being evaluated. Malladi et al. (2018) proposed a hybrid ground truth fusion technique for image segmentation evaluation and compared it with existing ground truth fusion methods on multiple ground truth datasets with different roughness levels.

A segmented image can have multiple reference images. In general, we select the best one in many reference images as the ground truth. Peng and Li (2013) proposed a probability measurement method to evaluate segmentation, the metric adaptively evaluates the structural information extracted from the segmented image. Thus, the local similarity score of each pixel in the segmentation result can be obtained, or the local similarity score of each point can be accumulated in the global similarity score of the entire segment by an information theory method. Based on previous work, Peng et al. (2016) constructed a region-based adaptive evaluation framework, and no longer compared the segmentation results with each reference segmentation. This framework adaptively builds a common indicator evaluation reference, that is, the new example reference is used with the general form of region-based measurement. This paper applies three commonly used region-based metrics (PRI, GCE, and VOI) and proposes an effective scheme for calculating each metric. The proposed new measurements are denoted as Q_{PRI}, Q_{GCE} and Q_{VOI}, respectively, the values of them are calculated using only exemplary references, the formula is as follows:

$$Q = \sum_{R_j} \frac{N_{R_j}}{N} M_{R_j} \tag{6}$$

where Q refers to Q_{PRI}, Q_{GCE} and Q_{VOI}. Combining methods of the previous two articles, Peng et al. (2017) then came up with another framework, in which the ground truth is no longer a certain optimal image, but a combination of multiple reference images. The best partial segmentation result in each reference image is labeled and then adaptively compose into the final ground truth, which not only maintains structural consistency but also locally matches the input segments. The quality of a given segmentation is then measured by its distance to the combined reference. The segmentation results of multi-person markers reflect different levels of perceptual detail. Therefore, using more marked segments as a reference will make the evaluation fairer. The author proposed a new algorithm (Q_p) based on the ground truth of multi-person markers to evaluate the segmentation quality.

Most of the existing evaluation metrics use pixel element analysis results. The larger the percentage of pixels that do not match the ground truth, the worse the performance of the segmentation results, but these methods are not always reasonable and robust. Feng et al. (2016) proposed a new objective evaluation metric based on the weighted-ROC graph, it overcomes the irrationality that the traditional objective evaluation metric always gives the same processing to the object pixels and the background pixels in the image.

$$wTPR = \frac{wTP}{wTP + wFN} \tag{7}$$

$$wFPR = \frac{wFP}{wFP + wTN} \tag{8}$$

A weighted ROC map is depicted with wFPR as the x-axis and $wTPR$ as the y-axis. Lower $wFPR$ and higher $wTPR$ indicate better segmentation (definitions of TP, FP, TN and FN see in Fig. 2). Dogra et al. (2012) combined two robust factors, namely segmentation region and boundary information, which greatly improves the robustness of the metric. A linear combination of boundary and area measurements is considered to estimate the impact of boundary and region matching information, the formula for this metric is as follows. The higher the combined value, the better the quality of the segmentation.

$$CMI(G, S) = \alpha \times SC(G, S) + (1 - \alpha) \times BS(G, S) \tag{9}$$

SC and BS are metrics based on segmentation region and boundary information, respectively. Where α and $1 - \alpha$ are weights of respective indices.

Some new methods apply a perceptual pooling strategy to segmentation assessment and use fuzzy set theory instead of classical evaluation methods to evaluate image segmentation. Peng et al. (2018) proposed two pooling strategies for segmentation quality assessment and an evaluation measure related to human perception of the quality of segmentation is designed by assigning meaningful weights to the quality map. Specifically, Quality-based Pooling strategies (QP) designed and tested some commonly used assessment metrics, while also designing and testing some popular assessments based on Visual Importance Pooling strategies (VIP). The advantage of the method is that this is the first work to apply a perceptual pooling strategy to segmentation assessment. Image segmentation evaluation measures proposed by Ziolko et al. (2018) use fuzzy set theory instead of classical evaluation methods, the fuzzy method (F-score) needs to consider the importance of regional differences relative to the area of the region. Methods for estimating regional similarity are proposed based on three different methods, namely the common number of pixels, the similarity of the contours, and the position of the centroid. The final measurement of the entire image is based on the recall and accuracy assessment of adaptive fuzzy set theory.

3.1.2 Unsupervised evaluation methods

Unsupervised evaluation method evaluates the segmentation by calculating the feature parameters of the segmentation result image directly, which has the advantage that the ideal segmentation reference image is not required. Feature parameters of the segmentation result image are also called indicators or measures. In practical applications, the construction of standard reference images varies from person to person, with great subjectivity, and is too time-consuming and laborious, some natural images even cannot construct accurate reference images. Therefore, the application of supervised evaluation methods is limited greatly, and in some cases, it even can't be used. On the contrary, the unsupervised method evaluates the performance of the segmentation algorithm objectively, which gets rid of the constraints of human subjective factors on the evaluation process through qualitative analysis and characteristic parameters. Although accuracy of unsupervised is not as good as the supervised evaluation method, the evaluation results are objective and stable, so it is a

reliable segmentation evaluation method. The unsupervised evaluation method is not limited by the application and can be used under any circumstances. Therefore, it is of great significance to study an unsupervised evaluation method that can replace the supervised evaluation to some extent and realize the quantitative determination of image segmentation quality. This may become an inevitable trend in the research of image segmentation evaluation methods.

Reviewing several literatures, we find that researchers try to use some simple indicators to evaluate the segmentation results at first. These indicators related to quantifying and analyzing, such as the amount of intervention, segmentation errors, robustness of segmentation, ease of segmenting objects in order, and the time required to complete a task (Flores and Lotufo 2008). Some mathematical indicators such as Peak Signal to Noise Ratio(PSNR) and Root Mean Square Error(RMSE) (Chouhan et al. 2018) are also used to evaluate the quality of the segmentation. Many more effective unsupervised indicators were subsequently proposed. Zeboudj's contrast(Zeb) (Garcia-Lamont et al. 2018; Zhang et al. 2008) measures the inside and outside contrast of the adjacent area of each pixel. D_{WR} (Zhang et al. 2008) measures the difference in gamma between the input image and the segmentation result image. η (Zhang et al. 2008) measures the foreground and background variances, respectively, and also measures the variance between the two. Performance Vector (PV) (Levine and Nazif 1985) is a set of optimization criteria used to evaluate performance. These measures jointly consider the factors involved in defining segmentation. They reflect the segmentation status at any point in time so they are dynamic. NU (Sahoo et al. 1988) uses a normalized region uniformity metric, which enhances the region uniformity metric in PV and improves PV, and Shape Measure (SM) (Sahoo et al. 1988) is defined as the sum of the gradients at each pixel where the feature value exceeds the segmentation threshold and the average value of its neighboring pixels. The calculation of Second-Order Entropy (SE) (Pal and Bhandari 1993) is an entropy-based evaluation method that evaluates segmentation through the intra-region uniformity of second-order local entropy. F (Liu and Yang 1994) was originally proposed to evaluate the segmentation results of real and synthetic images locally and globally, which uses the square root of the number of segments to be proportionally weighted to punish over- segmentation. E_{CW} (Chen et al. 2004) is a combination evaluation of over-segmentation and under-segmentation. V_{CP} (Correia and Pereira 2003) and V_{EST} (Erdem et al. 2004) are two metrics that evaluate the quality of video segmentation. When the Validation evaluation partition coefficient(V_{pc}) (Chouhan et al. 2018) is high and the Validation evaluation partition entropy(V_{pe}) (Chouhan et al. 2018) is low, it means that the segmentation is better. Levine and Nazif's inter-class contrast(Inter) (Chabrier et al. 2004) calculates the sum of regional contrasts, Levine and Nazif's intra-class uniformity(Intra) (Chabrier et al. 2004) calculates the sum of standard deviations of each region, and normalizes them. Combination of intra-class and inter-class disparity (Intra-inter) (Chabrier et al. 2004) calculates the similarity between Intra and Inter. The calculation of Borsotti criterion (Borsotti) (Borsotti et al. 1998) is based on the number of regions, surface and variance. Borsotti improved F by punishing the segmentation of many small areas with the same size and reducing the bias against over-segmentation and under-segmentation. Rosenberger's criterion (Rosenberger) (Rosenberger and Chehdi 2000) can adaptively calculate the segmentation result according to the uniformity or texture features of the region and takes into account intra -regional consistency and inter-regional differences. E (Garcia-Lamont et al. 2018; Zhang et al. 2008, 2004) is an objective segmentation evaluation method based on information theory. This method uses entropy as the basis for measuring the uniformity of pixel features within the segmentation region. Srubar (2012) proposed a threshold-based approach to evaluate the

effectiveness of some existing methods of measuring segmentation quality, such as Goodness Uniformity(GU) (Levine and Nazif 1985) and Figure of Certainty(FOC) (Strasters and Gerbrands 1991).

We hope that an evaluation method provides general measurement that can process synthetic images, natural images and medical images, and provides comprehensive information about segmentation. Kubassova et al. (2008) proposed a new unsupervised method for automatic segmentation assessment and compared its performance to existing unsupervised and supervised methods. This method can meet the above requirements. Philipp-Foliguet and Guigues (2006) proposed an evaluation criterion for segmenting a color image into regions when no basic facts are available, this criterion considers the complexity of the segmentation and the appropriateness of the extracted region to the original image. Ye et al. (2012) used the original image as a comparison object, instead of manually referencing image, thereby establishing a calculation model that can automatically predict the image quality without the reference image, making the calculation more efficient. Shi et al. (2017) proposed an unsupervised objective evaluation method for the quality assessment of image segmentation of a single object. This method combines the prior information of the object quantity and better conforms to subjective evaluation of the object segmentation quality, the formula is as follows:

$$Q_q = \frac{1}{1 + (1 - \frac{A_{RL}}{A_S})(N_R - 1)} \tag{10}$$

$$Q_o = \frac{1}{N} \sum_{n=1}^{N} \frac{\sum_{p \in B_{S_n}} exp(-\frac{D_p}{\alpha L_{d_n}})(1 - \frac{A_{e_n}}{W_n \cdot H_n})}{P_n} \tag{11}$$

$$Q_{overall} = \frac{1}{2}(Q_o + Q_q) \tag{12}$$

Q_q is used to estimate the object quantity. Q_o is used to evaluate objectness.

Image edges, used to measure the quality of the segmentation, are more robust relative to pixel values and other features. Cai et al. (2017) introduced an edge-based segmentation evaluation method, defined as:

$$E_S(S) = \sum_{R_i \in R} \frac{|R_i|}{N} [E_{r1}(R_i) - E_{r2}(R_i)] E_{r3}(R_i) E_A(R_i) \tag{13}$$

$E_{r1}(R_i)$ and $E_{r2}(R_i)$ measure the edge fitness of a region and the intra-region edge error, respectively. $E_A(R_i)$ is the expression of appearance evaluation. Habba et al. (2018) proposed a novel Gini-Entropy based evaluation criterion(called GE), this new approach combines the Gini index with entropy, overcomes the limitations of the Gini index, allows for the evaluation of the combination of regions arrangement in a segmented image, and measures pixel uniformity within each region. The formula for GE is as follows:

$$GE = G(I) + H_{lay}(I) \tag{14}$$

$$H_{lay}(I) = -\sum_{i=1}^{m} \frac{m_i}{N} log(\frac{m_i}{N}) \tag{15}$$

The implementation of many quantitative assessment methods relies on visual quality judgment, which means subjective. The commonly used subjective evaluation method compares different segmentation result images through the human eye, and its stability is poor, and it is difficult to describe quantitatively, which is greatly limited in application. Objective evaluation is more impartial and accurate without being influenced by human factors. Therefore, an objective evaluation method that can replace subjective evaluation to some extent and can be quantitatively described has important research significance and application value. Khan and Bhuiyan (2014, 2011) proposed a weighted self-entropy and weighted mutual entropy segmentation evaluation method, which measures the heterogeneity of pixels between regions and the homogeneity within the region objectively, it is easy to implement. Sharma et al. (2010) proposed an objective image segmentation evaluation method based on regional contrast and intensity uniformity over regions, the method considers the mean and variance parameters for each region and can be used to compare different segmentation methods. This method can also be used to evaluate compressed domain segmentation images. Eftekhari-Moghadam and Abdechiri (2010) proposed a new unsupervised method (MA) to evaluate image segmentation objectively, the method using a Gaussian distribution probability model to reduce the estimation error, using the pixel information (mean and variance) of each region to balance under and over-segmentation, extracting statistical information about the image from each region to provide adaptive evaluation.

Previous unsupervised evaluation methods often failed to accommodate multi-scale segmentation. Lu et al. (2016) believed that the quality of segmentation depends on the application, therefore, the target scale is introduced into the evaluation model to constrain the relative segmentation quality. A scale constraint assessment method (Q_t) based on the specified target scale to estimate segmentation quality is proposed. First, regional significance and merging necessity are used to describe intra-regional homogeneity and inter-regional heterogeneity, respectively. Then they are normalized to the equivalent spectral distance of the predefined area. Finally, an evaluation model is established by analyzing the relationship between image features and segmentation quality. Segmentation techniques are also commonly used in froth images. Jinping et al. (2013) proposed an unsupervised evaluation method for froth image segmentation based on gray features, because the ground truth of froth image segmentation is always difficult to obtain, this method can select the optimal froth image segmentation algorithm by setting appropriate algorithm parameters.

3.2 Medical image segmentation evaluation

Good segmentation results require effective evaluation methods, so the evaluation of medical image segmentation is valued worldwide highly. In order to cope with various threats and meet application requirements, segmentation algorithms are evolving constantly. However, existing evaluation methods cannot evaluate the segmentation algorithm accurately. Recent years, due to the wide application of image segmentation technology in medical image processing, indicators and frameworks for evaluating the quality of medical image segmentation have also been proposed continuously. After reading a lot of literature, it is found that the application of unsupervised evaluation in medical image segmentation is

scarce. Therefore, the application of unsupervised evaluation in medical image is not summarized in this paper.

The widely used local measurement for 3D medical image segmentation quality assessment is Surface Distance(SD) (Aspert et al. 2002). Other available measures have certain limitations for 3D medical segmentation assessment. However, SD also has obvious disadvantages, such as asymmetry and under-estimation errors in different regions during the evaluation process. Getto et al. (2015) proposed a more reliable distance measurement that finds and improves the shortcomings of the SD in areas of high dissimilarity. This method is used to analyze and evaluate the local differences between the segmentation results and the 3D segmentation ground truth. Karimi et al. (2014) used statistical techniques and information theory to develop two complementary assessment methods for measuring systematic errors such as over-segmentation and under-segmentation, outliers and overall errors.

Although there are many image segmentation algorithms, it is not clear which one is universal and best. In addition, microscopic images have various microscopic properties that may degrade the performance of image segmentation algorithms. Benes and Zitova (2015) evaluated the performance of microscopic image segmentation algorithm using several indicators of the image quality, and then analyzed the effects of various segmentation methods on the microscopic image test set. Bernard et al. (2016) constructed a standardized evaluation framework that evaluates and compares the performance of left ventricular intimal segmentation techniques Reliably in real-Time 3D Echocardiography (RT3DE). A database consisting of 45 multi-donor cardiac ultrasound records with corresponding reference values obtained at different centers is provided. Quantitative evaluation and comparison show that the framework is effective in the extraction and measurement of clinical indicators, and provides good segmentation accuracy in terms of average distance error.

Many methods for tissue segmentation in brain MRI scans have been proposed, but it is not possible to determine which segmentation method is optimal. Mendrik et al. (2015) presented the MRBrainS online assessment framework for direct and objective evaluation of automated and semi-automated methods. Researchers can apply their algorithms to the data provided in the framework, and segmentation methods are ranked according to the overall performance of the algorithm, helping to choose the best performing method for segmentation. Skalski et al. (2018) proposed a Locally-oriented Evaluation Framework for Medical Image Segmentation (LEFMIS) based on anisotropic Euclidean distance transform (EDT) and distance projection, the proposed method is robust and takes into account the anisotropy of medical data such as MRI or CT, with greater utility and flexibility. Laurent et al. (2016) presented an evaluation platform for evaluating segmentation algorithms of detecting anatomical structures in medical images, the proposed platform is divided into two main parts: the first part is the generation and description of the gold standard. The second part is the segmentation method evaluation.

3.3 Remote sensing image segmentation evaluation

In addition to medical images, remote sensing images are also the most widely used field of segmentation techniques in life. In the study of topography, wetland resource monitoring, land cover monitoring and prevention of geological disasters, it is necessary to acquire characteristics of certain places, which are often not accessible easily by humans. At this time, remote sensing images can be obtained by radar or satellite. After processing of the

obtained remote sensing image, the required information can be extracted, and image segmentation is an indispensable step in remote sensing image processing.

3.3.1 Supervised evaluation methods

Most of the remote sensing image segmentation quality evaluation indicators proposed in recent years are based on the comparison of global and local feature consistency and the improvement of over or under-segmentation, so more effective measures need to be developed. Marpu et al. (2010) defined the case of over-segmentation and under-segmentation, and proposed corresponding evaluation criteria according to the definition. The purpose of this method is to deal with the result of multi-level segmentation algorithm commonly used in Geographic Object-Based Image Analysis(GEOBIA). Zhang et al. (2015) proposed region-based precision and recall, these two metrics are combined in four different ways and these four ways are used to compare segmentation results in order to assess segmentation quality. The four combination strategies:

$$f = \frac{1}{\alpha \frac{1}{precision} + (1-\alpha)\frac{1}{recall}} \tag{16}$$

$$sum = precision + recall \tag{17}$$

$$ed = \sqrt{precision^2 + recall^2} \tag{18}$$

$$ed' = \sqrt{(1-precision)^2 + (1-recall)^2} \tag{19}$$

In practical applications, comparing the performance of different segmentation algorithms in a large number of image databases is a rather arduous task. In order to speed up the process of comparison and improve the efficiency of calculation, Cruz et al. (2017) realized the automatic detection system by using MATLAB parallel computing toolbox. The main part of the system has been parallelized, and the segmentation algorithm can be analyzed and executed simultaneously and the accuracy of the analysis results can be evaluated. CPU usage and processing time is reduced significantly compared with systems that execute each algorithm one by one.

Most of current segmentation assessment metrics only focus on globally valid assessments. However, these methods are ineffective for the case of two segmentation results with very similar overall performance and very different local errors. In order to solve this problem, Su and Zhang (2017) proposed a method for local and global quantization segmentation error, the over-segmentation and under-segmentation errors of each reference segmentation object are quantized using the degree of region overlap. The segmentation error map generated by these quantized error values can effectively depict local indicators. The global indicator can be obtained by region-weighted summation of the error values of all reference segmentation objects, the formula is as follows:

$$GOSE = \frac{1}{A_t} \sum_{i}^{N} A_i \cdot OSE_i \tag{20}$$

$$GUSE = \frac{1}{A_t} \sum_{i}^{N} A_i \cdot USE_i \tag{21}$$

where OSE_i and USE_i represent the over-segmentation and under-segmentation errors of the i-th reference segmentation object, respectively.

3.3.2 Unsupervised evaluation methods

Remote sensing images are difficult to obtain standard reference images due to the complexity of their composition, so the evaluation of segmentation performance more often uses unsupervised evaluation. In recent years, unsupervised evaluation methods for remote sensing image segmentation have mostly weighted combination the existing unsupervised metrics to obtain better evaluation results. Zhang et al. (2012) proposed an evaluation criterion that takes into account the measure of homogeneity and inter-segment heterogeneity within global segments, this improved unsupervised method can be used to compare the segmentation results produced by a single segmentation method, the formula is as follows:

$$Z = T + \lambda D \tag{22}$$

When Z reaches its minimum value, the best segmentation result can be obtained. The weight λ is determined based on the application of the segmentation result. T and D represent the homogeneity within the region and the heterogeneity between the regions, respectively. Wang et al. (2018) evaluated segmentation quality by combining the area-Weighted Variance (WV) and the Jeffries–Matusita (JM) distance using three different strategies. These three methods are F-measure, Z method and Local Peak (LP) method (Yang et al. 2014), respectively.

Global Scoring (GS) is one of the most commonly used method in the GEOBIA environment to automatically select parameters of the segmentation algorithm. It is a pity that the serious source of instability of this method makes the segmentation quality assessment not very accurate so this method has been ignored. Boeck et al. (2017) proposed a way to modify the global score to alleviate the problem. Global Score (GS) is a combination of area-weighted variance(v) and Moran's I(I).

$$GS = v_{norm} + I_{norm} \tag{23}$$

Gao et al. (2017) extracted multiple spectral and spatial features of an image simultaneously and integrated them into a feature set, the spatial stratification heterogeneity and spatial autocorrelation metrics are then combined into a global assessment metric as the final quality score. This method overcomes the difficulties of designing effective metrics that the unsupervised method currently faces in practice. Johnson and Xie (2011) used a multiscale approach to improve the segmentation quality of high spatial resolution color infrared images in residential areas.

Under and Over-segmentation Aware (UOA) can estimate the optimal scale and also clearly determine whether a segment is over-segmented or under-segmented, however, due to design flaws that may result in over-estimation of over-segmentation errors. Su (2018) improved the UOA method by overcoming the defect of estimating the excessive segmentation error, and proposed a UOA-based unsupervised remote sensing image segmentation evaluation scheme. Two cases of error-prone defects are listed, and edge strength is used to

Table 1 Supervised evaluation methods

Application field	Metric (Author)	Publication date
Natural image	R (Roman-Roldan et al.)	2001 (Roman-Roldan et al. 2001)
	$d_{sym}, d_{asy}, d_{mut}$ (Cardoso et al.)	2005 (Cardoso and Corte-Real 2005)
	Feng Ge et al.	2006 (Ge et al. 2006)
	CMI(G,S) (Dogra et al.)	2012 (Dogra et al. 2012)
	Peng et al.	2013 (Peng and Li 2013)
	PEIS (Ledig et al.)	2014 (Ledig et al. 2014)
	Taha et al.	2014 (Taha et al. 2014)
	Sim(G,S) (Shi et al.)	2015 (Shi et al. 2015, 2014)
	SEL (Yang et al.)	2015 (Yang et al. 2015)
	Berezsky et al.	2016 (Berezsky et al. 2016)
	Arhid et al.	2016 (Arhid et al. 2016)
	F_{op}, F_b, F_r (Pont-Tuset et al.)	2016 (Pont-Tuset and Marques 2013, 2016)
	$Q_{PRI}, Q_{GCE}, Q_{VOI}$ (Peng et al.)	2016 (Peng et al. 2016)
	Feng et al.	2016 (Feng et al. 2016)
	Q_p (Peng et al.)	2017 (Peng et al. 2017)
	Malladi et al.	2018 (Malladi et al. 2018)
	QP,VIP (Peng et al.)	2018 (Peng et al. 2018)
	F-score (Ziolko et al.)	2018 (Ziolko et al. 2018)
Medical image	WMI,FDR (Karimi et al.)	2014 (Karimi et al. 2014)
	Getto et al.	2015 (Getto et al. 2015)
	Benes et al.	2015 (Benes and Zitova 2015)
	Bernard et al.	2016 (Bernard et al. 2016)
	Mendrik et al.	2016 (Mendrik et al. 2015)
	Laurent et al.	2016 (Laurent et al. 2016)
	LEFMIS (Skalski et al.)	2018 (Skalski et al. 2018)
Remote sensing image	Marpu et al.	2010 (Marpu et al. 2010)
	Zhang et al.	2015 (Zhang et al. 2015)
	Cruz et al.	2017 (Cruz et al. 2017)
	GOSE,GUSE (Su et al.)	2017 (Su and Zhang 2017)

Here is only a summary for some of the latest segmentation methods, as well as a few of methods that have not been summarized before. Please refer to Taha and Hanbury (2015) for the rest of the methods mentioned in the article

In the second column, the metric name is presented outside the brackets, and the author's name is shown in parentheses. Some articles do not indicate the name of the metric explicitly so we only have written the author's name

avoid over-estimation of over-segmentation errors (OSEs). In terms of OSEs identification, the proposed method is superior to the original UOA.

We have made a brief summary of the supervised evaluation methods proposed in the literature to facilitate search. These methods are listed in Table 1. Similarly, we have made a brief summary of unsupervised evaluation methods proposed in the literature to facilitate the search. These methods are listed in Table 2.

Table 2 Unsupervised evaluation methods

Application field	Metric (Author)	Publication date
Natural image	Philipp-Foliguet et al.	2006 (Philipp-Foliguet and Guigues 2006)
	Kubassova et al.	2008 (Kubassova et al. 2008)
	MA (Eftekhari-Moghadam et al.)	2010 (Eftekhari-Moghadam and Abdechiri 2010)
	Sharma et al.	2010 (Sharma et al. 2010)
	Srubar et al.	2012 (Srubar 2012)
	Ye et al.	2012 (Ye et al. 2012)
	Jinping et al.	2013 (Jinping et al. 2013)
	Khan et al.	2014 (Khan and Bhuiyan 2011, 2014)
	Cai et al.	2017 (Cai et al. 2017)
	Shi et al.	2017 (Shi et al. 2017)
	GE (Habba et al.)	2018 (Habba et al. 2018)
Remote sensing image	Johnson et al.	2011 (Johnson and Xie 2011)
	Z (Zhang et al.)	2012 (Zhang et al. 2012)
	Gao et al.	2017 (Gao et al. 2017)
	GS (Böck et al.)	2017 (Boeck et al. 2017)
	UOA (Su et al.)	2018 (Su 2018)
	Wang et al.	2018 (Wang et al. 2018)

Here is only a summary for some of the latest segmentation methods, as well as a few of methods that have not been summarized before. Please refer to Zhang et al. (2008) for the rest of methods mentioned in the article

In the second column, the metric name is presented outside the brackets, and the author's name is shown in parentheses. Some articles do not indicate the name of the metric explicitly so we only have written the author's name

3.4 Other evaluation applications

Some assessment methods or frameworks not only include supervised theories but also involve unsupervised knowledge, we make a brief summary of these unsupervised and supervised assessment methods. Prabha and Kumar (2016) had studied different evaluation techniques based on subjective and objective methods and conducted experiments on different segmentation methods. Using boundary-based methods (such as Sobel, Canny and Susan) and region-based methods (such as region growing, thresholding and blending methods) and combining the two methods to identify which method performs better in the nature image and the real-time image. Monteiro et al. (2012) had studied evaluation measures which can evaluate the quality of image segmentation results quantitatively. The status of distance assessment measurements is first introduced in his paper, followed by several evaluation criteria.

There are also some assessment methods that do not accurately determine whether they are supervised or unsupervised. These methods are related to the evaluation of medical, general and high-resolution fingerprint images segmentation. Fernandez et al. (2015) evaluated the segmentation method from a higher level of task(detection or classification) perspective, this method regards the classification of plankton as a higher-level task, and evaluates the segmentation method from the perspective of the accuracy of plankton

classification. In addition, this more comprehensive form of segmentation assessment better meets the requirements of big data analysis. Peng and Varshney (2015) took the segmentation algorithm Mean Square Error (MSE) as an example and developed a systematic method for evaluating the lower bound in the statistical estimation framework. The proposed method is effective and robust in determining the performance of the segmentation algorithm and providing a benchmark for the segmentation problem.

The complex structure of blood vessels, boundary blurring and uneven blood vessel contrast in Computed Tomography Angiography (CTA) images makes the evaluation of vessel segmentation algorithms a well-known problem. Recent blood vessel segmentation studies include the evaluation of only a single bifurcation or a small portion of a vascular tree. Luu et al. (2015) proposed an assessment method using landmarks. Wiesmann et al. (2017) proposed and validated a new method for actual fluorescence cell image simulation to evaluate cell segmentation algorithms. Algorithms that attempt to segment overlapping images can also be objectively evaluated by this proposed method. High-resolution fingerprint images can provide finer features than standard fingerprint images to improve recognition accuracy. Existing fingerprint image quality assessment methods can be classified into three types: local features of fingerprint images, global features, and classifiers such as neural networks to predict fingerprint image quality. Zhao et al. (2010) presented a comparative study for the quality assessment of high-resolution fingerprint images.

4 Metric selection for segmentation evaluation

The above has summarized some evaluation indicators for natural, medical and remote sensing image segmentation. Natural images are most common in life, and the segmentation evaluation often chooses commonly used metrics. There are large differences between medical images and remote sensing images, so the selection of their evaluation metrics also differs. In medical imaging, qualitative or theoretical evaluation is not enough, quantitative evaluation is necessary. The evaluation of medical image segmentation quality requires a reliable detailed comparison between reference segmentation and automatic segmentation. Researchers have proposed various evaluation methods, such as Dice, Sensitivity, specificity (Sundara and Aarthi 2019); FPR, HD, Mean surface distance (Hoang et al. 2019); Jaccard, TPR, TNR, FNR and PPV based on the accuracy of the edges, good measurements in the marked area, feature recovery, or differences from GT. Cappabianco et al. (2019, 2017) have improved Jaccard and Dice indicators for more complex medical anatomy segmentation evaluation. The appropriateness of these methods depends on the application and often there is no optimal method. It is worth noting that the evaluation of medical image segmentation should be locally targeted, so that problematic areas are found, so local metrics are usually used, and more widely used local metrics such as SD (Aspert et al. 2002), LEFMIS (Skalski et al. 2018), etc.

Remote sensing image segmentation evaluation focuses on the problems of over-segmentation and under-segmentation in order to accurately determine areas of various geographic objects, such as farmland, cities, forests and oceans. There is no standard way of evaluating remote sensing image segmentation results. Commonly used evaluation indicators are precision, recall, entropy, Moran's I, etc. Single indicators can be combined by different weights to obtain better evaluation results. There are two types of images with special imaging methods, hyper-spectral and multispectral images. Hyper-spectral images usually have more than 100 spectral bands and can provide detailed spectral information.

Multispectral images contain information with low statistical intervals and large amounts of information. For the processing of hyper-spectral and multispectral images, good segmentation is the basis, so various segmentation methods have been proposed. As for the evaluation methods of hyper-spectral and multispectral image segmentation, there are no specific standards, and they need to be based on specific applications. Common segmentation evaluation methods for hyper-spectral images are detection percentage(DP), quality percentage(QP) (Angulo et al. 2009); average recall(AvgRec), overall accuracy(OA) (Saqui et al. 2019); Sensitivity, Precision, FPR, Accuracy, MCC (Gautam and Bhutiyani 2016) and Euclidean Distance(ED), Spectral Angle Distance(SAD), Correlation Coefficient(COR) (Tang et al. 2019), and common segmentation evaluation methods for multispectral images are OA (Li and Xiao 2004), entropy (Mantilla and Yari 2017), Precision, Recall (Jordan and Angelopoulou 2012) and Inter, Intra, Intra-inter, Borsotti, Rosenberger (Chabrier et al. 2004). All in all, there are no determined criteria for the selection of image segmentation evaluation indicators, and their selection depends on the particular application.

5 Experimental results

In this paper, different experiments are designed for some indicators of supervised and unsupervised evaluation, and the indicator with better performance in recent years are introduced in detail and separate experiments. We demonstrate the validity of the classic and commonly used supervised and unsupervised indicators on different types of images. It should be noted that the purpose of the experiment is not to compare the advantages and disadvantages between different segmentation methods, but to consider the application of the supervised and unsupervised evaluation methods from the perspective of utilization.

5.1 Supervised evaluation

First, experiments on several newer indicators such as Q_p, F-score and Precision-recall are introduced in detail, and then experiments of other supervised indicators are briefly described.

5.1.1 Experiment on Q_p

An advantage of the Q_p method is the construction of the ground truth that combines different reference images to form a final standard image rather than relies solely on a single image as a reference, as shown in Fig. 4. The figure h can be obtained by combining the human segmentation reference images d-g with the figure b by the Q_p method. Similarly, the figure i can be obtained by combining the figure d-g with the figure c. The general evaluation algorithm is to select the best picture from the picture d-g as the standard picture. However, in this method, figure h and figure i are the reference images for evaluating the performance of the Compression-based Texture Merging(CTM)method and the mean-shift method, respectively. This method combines the best part of the segmentation of many reference images, making the segmentation evaluation results more accurate.

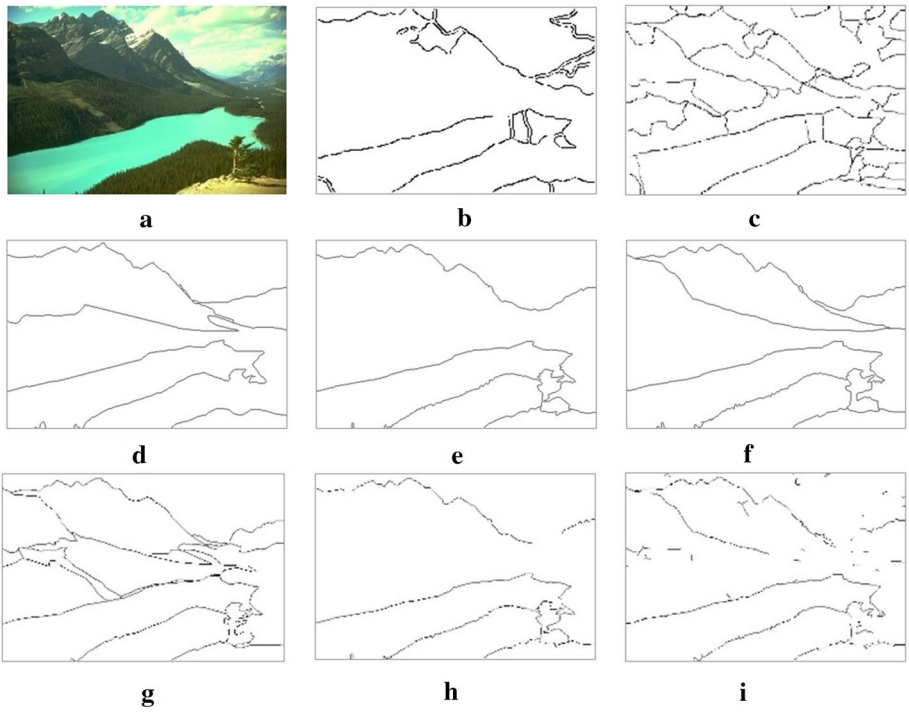


Fig. 4 An example of combining ground truth G^*. **a** Original image. **b** Segmentation result obtained by the CTM method. **c** The segmentation result obtained by the mean-shift method. **d–g** Human segmentation references. **h** The combined ground truth obtained by the CTM segmentation result. **i** The combined ground truth obtained by the mean-shift segmentation result

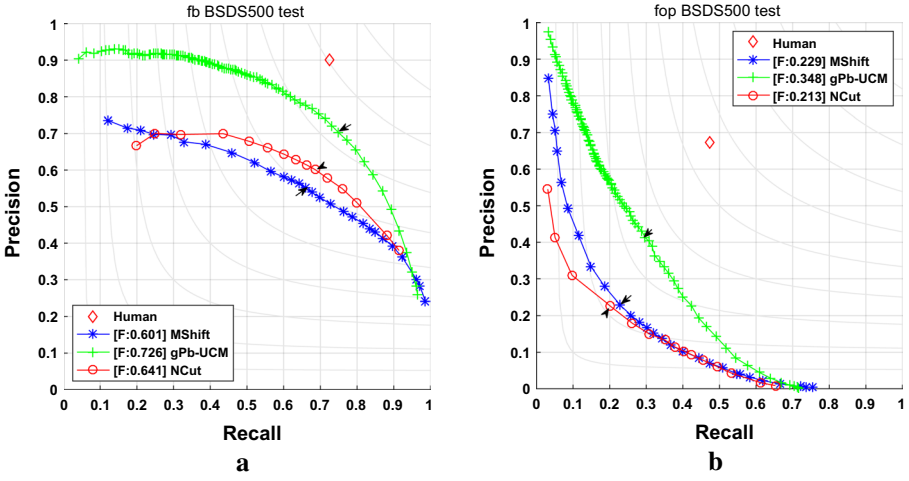


Fig. 5 Precision-Recall curves of the boundary (left) and the object and the part (right). The different curves represent three common segmentation methods (see legends). The isolated red symbol refers to the human performance evaluated on the same image. In the legend, the F-measure for the arrow mark on each curve is indicated in parentheses

5.1.2 Experiment on Precision-recall

F_b and F_{op} are two supervised metrics mentioned before. The two measurement methods are complementary, so we suggest them as the preferred tool for image segmentation evaluation. The average calculation time of F_b is only 0.5760 ± 0.200 s, while F_{op} is much faster than F_b, thus F_{op} greatly reduces the computational cost of the measurement. In the case of tight time constraints, F_{op} will be the tool of first choice.

Concatenating boundaries and objects-and-parts Precision-Recall curves is a good choice for evaluating segmentation algorithms, because in addition to obtaining the best meta-measurement results, they all have richer information and result knowledge in the form of precision-recall curves. The examples are shown in Fig. 5.

As can be seen in Fig. 5(f_{op} in the figure is F_{op}, and f_b is F_b), the two frameworks demonstrate that globalPb-Ultrametric Contour Map(gPb-UCM) (Pont-Tuset and Marques 2016) technology has outstanding results in the comparison of segmentation techniques, and gPb-UCM has the highest F-measure. For the other two techniques, the left framework considers Normalized Cuts(NCut) to be better than Mean-Shift(Mshift), while the right framework considers oppositely. From the value of F-measure we can think that the NCut and Mshift methods are not much different, and the very similar curves in the figure also show this. In summary, the two measurement methods are very complementary in terms of the attributes of the reflected partitions, so we also think they can be the preferred tool for image segmentation evaluation. The overall value of F_{op} is lower than F_b. In other words, F_{op} provides higher resolution so improvements can be made to the segmentation method.

5.1.3 Experiment on F-score

The evaluation plan should consider the diversity of the segmentation, the fuzzy *P&R* assessment measure (F-score) uses the fuzzy set theory to consider the importance of the segment difference relative to the segment size, more promising than classical methods to

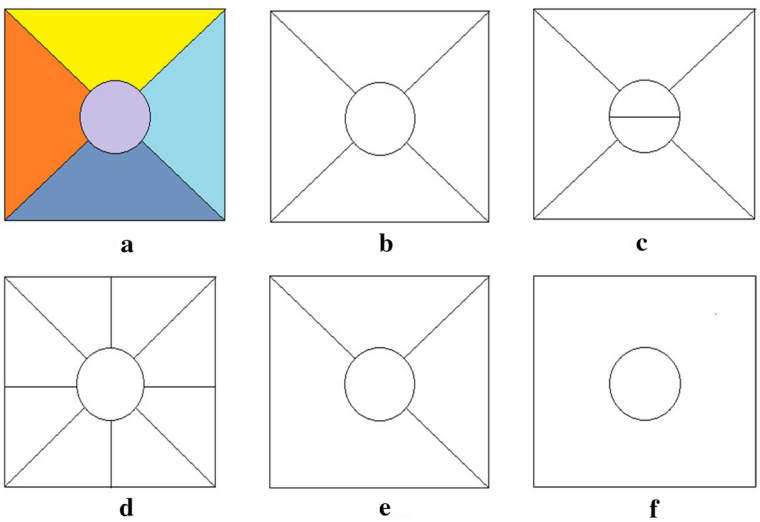


Fig. 6 Segmentation for synthetic image **a** original image, **b** ground-Truth, **c** minor over-segmentation, **d** serious over-segmentation, **e** minor under-segmentation, **f** serious under-segmentation

Table 3 Comparison of the results between fuzzy and classical evaluation

Metrics	Segmentation methods	c	d	e	f	b
Precision	Fuzzy	0.2730	0.1572	0.6149	0.6936	1.0000
	Classical	0.9006	0.9899	0.5465	0.5413	1.0000
Recall	Fuzzy	0.2940	0.3749	0.3784	0.2134	1.0000
	Classical	0.9006	0.9899	0.5465	0.5413	1.0000
F-score	Fuzzy	0.2831	0.2215	0.468	0.3264	1.0000
	Classical	0.9006	0.9899	0.5465	0.5413	1.0000

narrow the gap between human and machine evaluation. The classical membership function depends linearly on the number of common pixels. In general, due to the localization, size and shape of the segments, a lower F-score value is resulting. This method applies fuzzy logic method to the well-known "precision and recall" (*P&R*) method. The F-score method replaces the classical evaluation method with fuzzy set theory. The fuzzy evaluation F-score is often used to obtain a threshold to judge whether the segmentation result from same image or different image.

If all the pixels of the image are segmented, then $|S|$ and $|GT|$ equal the number of all pixels, the classic *P&R* method fails, as a result, Precision = Recall = F-score. In addition, the classical method is usually biased towards segmentation with a large number of segments, for the segmentation of a small number of segments reduces the number of common pixels. The fuzzy method does not have above disadvantages. We use Fig. 6 to test the validity of the F-score indicator. The values obtained from the test are listed in Table 3.

The value of F-score is obtained by combining Precision and Recall. In Fig. 6, c and d belong to over-segmentation, and c is better than d. Therefore, the F-score value of c is higher than d. e, f belong to under-segmentation, and e is better than f, so the F-score value of e is higher than f. The rightmost column in the Table 3 indicates that b and b are compared and evaluated, because b is the ground truth image, so the result is 1 definitely. It shows the rationality of the proposed method.

5.1.4 Experiments on other supervised indicators

As shown in Fig. 7, we use different types of images such as landscapes, animals, and characters to conduct experiments. Various types of images make the experiment more convincing. The experimental data is listed in Table 4, and the line chart is shown in Fig. 8.

Observing Table 4, the smaller the value of VOI is, the better the result of image segmentation will be. The rest of the metrics should be as large as possible. Comparing the ground truth of a and b with the two segmentation results, the results of the CTM should be better than the mean-shift results. The Q_p method uses the constructed Ground-Truth (the construction method is shown in Fig. 4) to get a correct judgment. Other methods make a slightly biased judgment by using a single reference image. For Figure a, F-score also has an accurate judgment. For figure c, it can be seen clearly that the segmentation result of the CTM method is better than mean-shift, because the body part of the second bear is not well recognized in the segmentation result of mean-shift. PRI, SC, dhd, bgm, precision, specificity, Q_p and F-score provide more accurate results, and other metrics are slightly biased. In addition to Specificity in Figure d, other metrics all agree that CTM results are better than mean-shift, which is also in line with human visual assessment criteria. Looking at the graph e, the results of the two segmentation methods are similar basically, and

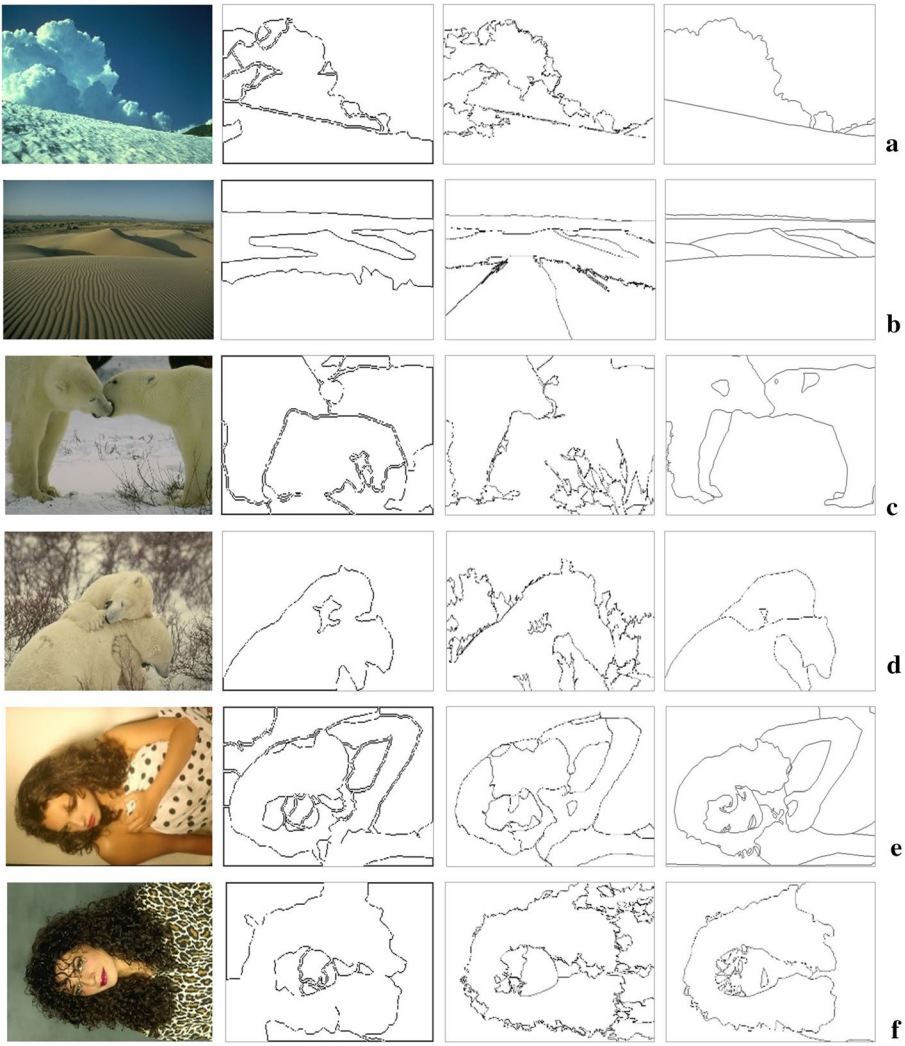


Fig. 7 Examples of segmentation results by different segmentation methods. The leftmost column represents the original images, the middle two columns represent the segmentation results obtained by the CTM and mean-shift methods, and the rightmost column represents the ground truth

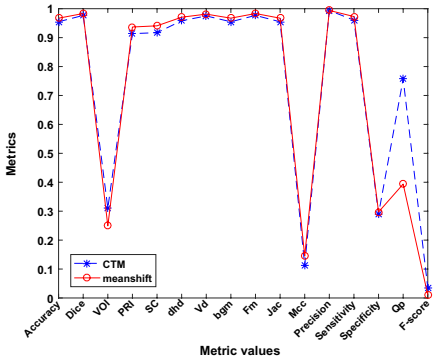
the result of the mean-shift segmentation is better because there is no line in the upper left corner. In addition to Q_p and F-score, other measurement methods successfully determine the result, and Q_p and F-score only have a small negligible error. Therefore, we think that the results of the two segmentation methods are comparable for this image. The segmentation result of Figure f cannot be accurately judged by the naked eye. VOI, Mcc, Precision, and Specificity consider that mean-shift is superior to CTM, while other measures consider the opposite.

The choice of evaluation metrics should consider different applications, different metrics have different evaluation results for the same image, and the same metric evaluates

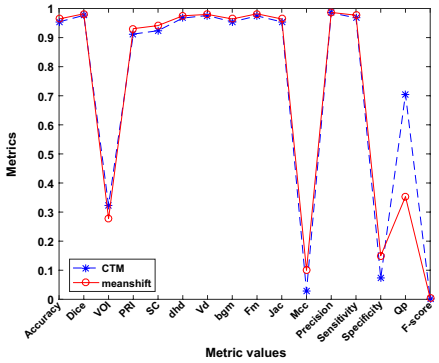
Table 4 Multiple indicator metrics for segmentation results of CTM and mean-shift methods on different images

Images methods	a		b		c		d		e		f	
	1	2	1	2	1	2	1	2	1	2	1	2
Accuracy	0.9548	0.9672	0.9539	0.9641	0.9364	0.9607	0.9785	0.9632	0.9150	0.9557	0.9414	0.9398
Dice	0.9768	0.9833	0.9764	0.9817	0.9671	0.9799	0.9891	0.9812	0.9554	0.9782	0.9698	0.9689
VOI	0.3107	0.2493	0.3227	0.2774	0.4057	0.2905	0.1790	0.2653	0.5242	0.3299	0.4049	0.3893
PRI	0.9138	0.9367	0.9121	0.9307	0.9246	0.8808	0.9579	0.9290	0.8445	0.9186	0.8898	0.8869
SC	0.9171	0.9408	0.9236	0.9413	0.9329	0.8869	0.9679	0.9336	0.8553	0.9335	0.9037	0.8985
dhd	0.9583	0.9706	0.9672	0.9749	0.9696	0.9442	0.9841	0.9681	0.9300	0.9724	0.9582	0.9541
Vd	0.9750	0.9812	0.9759	0.9797	0.9656	0.9784	0.9882	0.9802	0.9530	0.9741	0.9692	0.9672
bgm	0.9548	0.9673	0.9539	0.9641	0.9607	0.9364	0.9785	0.9631	0.9150	0.9575	0.9414	0.9398
F-measure	0.9768	0.9833	0.9746	0.9817	0.9671	0.9799	0.9891	0.9812	0.9554	0.9782	0.9698	0.9689
Jaccard	0.9547	0.9672	0.9539	0.9642	0.9362	0.9607	0.9785	0.9631	0.9146	0.9573	0.9413	0.9373
Mcc	0.1138	0.1451	0.0278	0.1007	0.0713	0.0837	0.0873	0.0746	0.0733	0.1567	0.0239	0.0634
Precision	0.9938	0.9940	0.9852	0.9865	0.9899	0.9888	0.9932	0.9934	0.9790	0.9800	0.9810	0.9822
Sensitivity	0.9604	0.9729	0.9678	0.9769	0.9460	0.9713	0.9851	0.9693	0.9329	0.9764	0.9589	0.9559
Specificity	0.2902	0.2972	0.0724	0.1498	0.1991	0.1560	0.1392	0.1800	0.1890	0.1912	0.0755	0.1356
Q_p	0.7576	0.3937	0.7057	0.3531	0.5710	0.2434	0.6580	0.2790	0.3029	0.2913	0.4123	0.1136
F-score	0.0359	0.0108	0.0023	0.0032	0.1300	0.0119	0.1663	0.0074	0.0561	0.0338	0.0146	0.0075

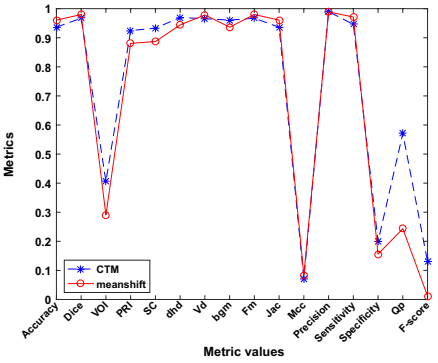
1 represents the CTM method and 2 represents the mean-shift method



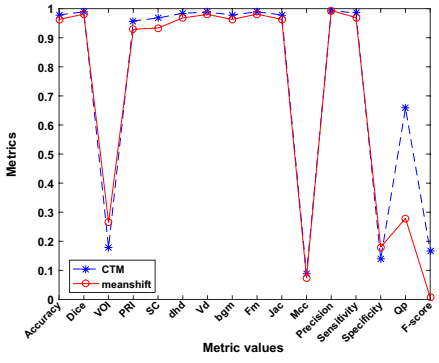
a Line chart of a



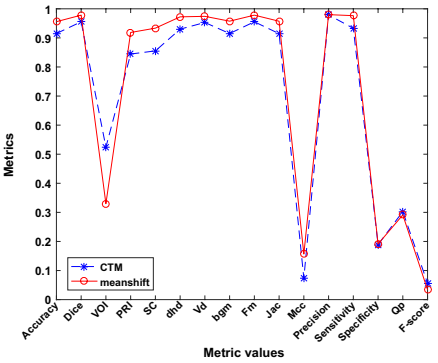
b Line chart of b



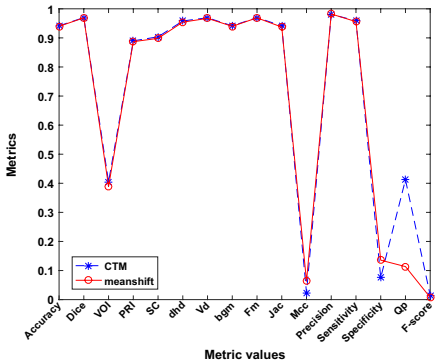
c Line chart of c



d Line chart of d



e Line chart of e



f Line chart of f

Fig. 8 Line chart of multiple metric values for **a**–**f**

different images differently. Based on the above analysis, for the judgment of the image segmentation result in Fig. 7, the order of selecting the metric should be Q_p firstly, because it can accurately judge the segmentation results of a and b. Followed by F-score, although

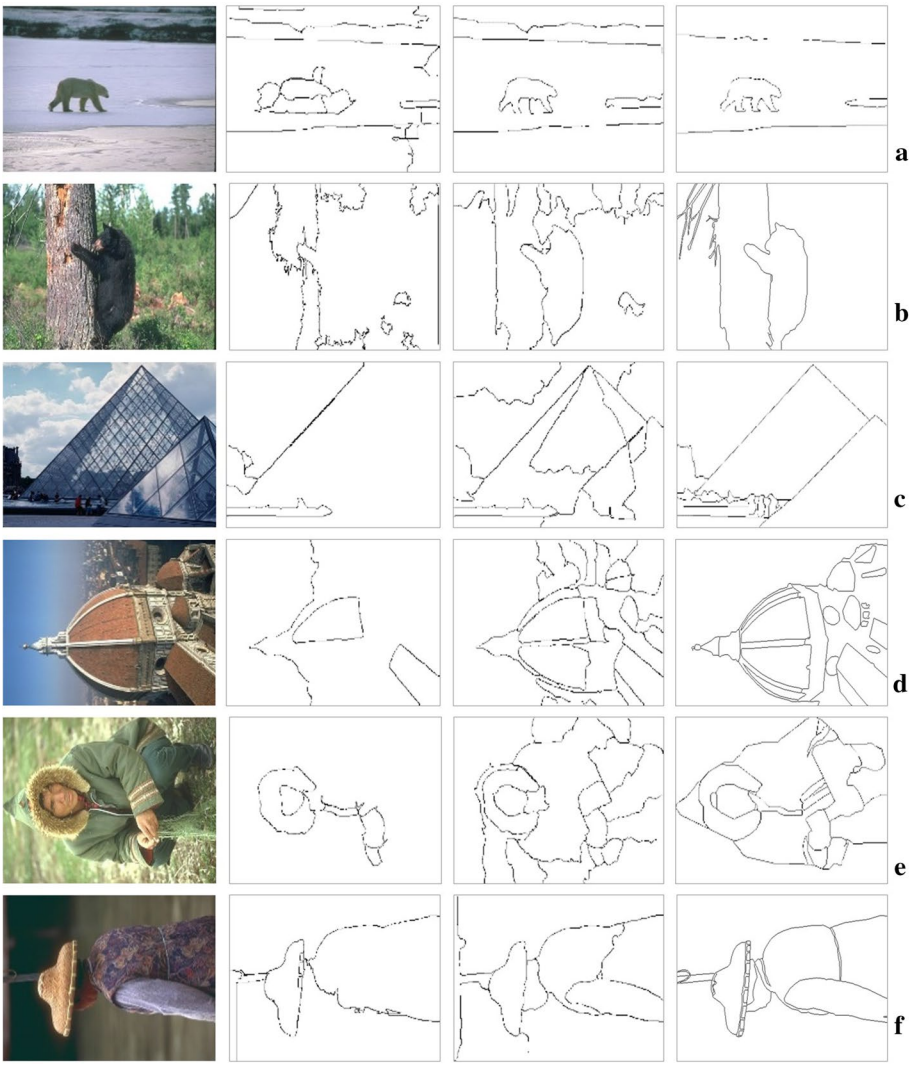


Fig. 9 Examples of segmentation results by the same segmentation method (mean-shift) with different parameter values. The leftmost column represents the original images, the rightmost column represents the ground truth

this metric fails to determine the exact result of b, it can better evaluate a, while the rest of the metrics make a wrong assessment. The following order is SC, PRI, Specificity, dhd, bgm and Precision. These six metrics judge that the CTM segmentation result of c is better than mean-shift, which is in line with visual observation. The numerical differences between CTM and mean-shift are 0.0460, 0.0438, 0.0431, 0.0254, 0.0243 and 0.0011, respectively. The larger the numerical differences are, the easier the evaluation results' distinctions will be, that is, the more accurate the measurement results will be. In the same way, the following order can be calculated as VOI, Sensitivity, Jaccard, Accuracy, Mcc, vd,

Table 5 Multiple indicator metrics for GT on different images and multiple indicator metrics for segmentation results for mean-shift methods with different parameters on different images

Images	Metrics	Zeb	Entropy	Intra-inter	Intra
a	Meanshift1	0.5223	0.0229	0.4503	0.1671
	Meanshift2	0.5163	0.0154	0.4640	0.1667
	GT	0.5014	0.0105	0.5016	0.1659
b	Meanshift1	0.3154	0.1065	0.4110	0.1986
	Meanshift2	0.4082	0.0209	0.4061	0.1986
	GT	0.4343	0.0140	0.5010	0.1988
c	Meanshift1	0.1971	0.0125	0.3480	0.2926
	Meanshift2	0.2135	0.0270	0.3660	0.2936
	GT	0.6181	0.0122	0.5004	0.2929
d	Meanshift1	0.5250	0.0088	0.4504	0.1612
	Meanshift2	0.3356	0.0295	0.4132	0.1597
	GT	0.5737	0.0247	0.5010	0.1589
e	Meanshift1	0.3428	0.0103	0.4192	0.1729
	Meanshift2	0.3691	0.0259	0.4154	0.1728
	GT	0.9187	0.0206	0.5014	0.1720
f	Meanshift1	0.1772	0.0133	0.4260	0.1629
	Meanshift2	0.4182	0.0161	0.4364	0.1648
	GT	0.5423	0.0175	0.5008	0.1642

Meanshift1 and Meanshift 2 represent mean-shift methods with different parameters

F-measure and Dice. Note that the ordering here is for the above picture, it is not universal, and it should be treated in a specific situation.

5.2 Unsupervised evaluation

Similarly, we use different types of images for unsupervised metric experiments, as shown in Fig. 9. The results obtained in the experiment are listed in Table 5. Figure 10 is a bar graph of the experimental results.

Looking at Fig. 9, it can be seen that the segmentation qualities of the pictures from the second column to the fourth column are sequentially increased, which is easily observed with the naked eye. To analyze the data in Table 5, the value of entropy should be as small as possible, while the larger Zeb, Intra-inter and intra values correspond to better segmentation results. For b, c, e, and f, Zeb has a good result, and accurately judges the quality of the segmentation result, which is in line with the human evaluation result. The judgment of d is slightly biased. It distinguishes that GT is the best, but it makes a mistake in the judgment of the results of Meanshitf1 and Meanshitf2. Zeb's judgment on the quality of a segmentation is exactly opposite to the human visual judgment, with a large error. Intra-inter made judgments on a, c, and f that meet human visual assessment criteria. Although it judges that GT is the best on b, d, e, it has the opposite judgment for the results of Meanshitf1 and Meanshitf2. The intra' judgments of the three segmentation results in a, b, c, d, e, and f are similar, and there is no clear judgments on the quality of the segmentation, which has a poor result. In summary, we can see from the segmentation result judgments

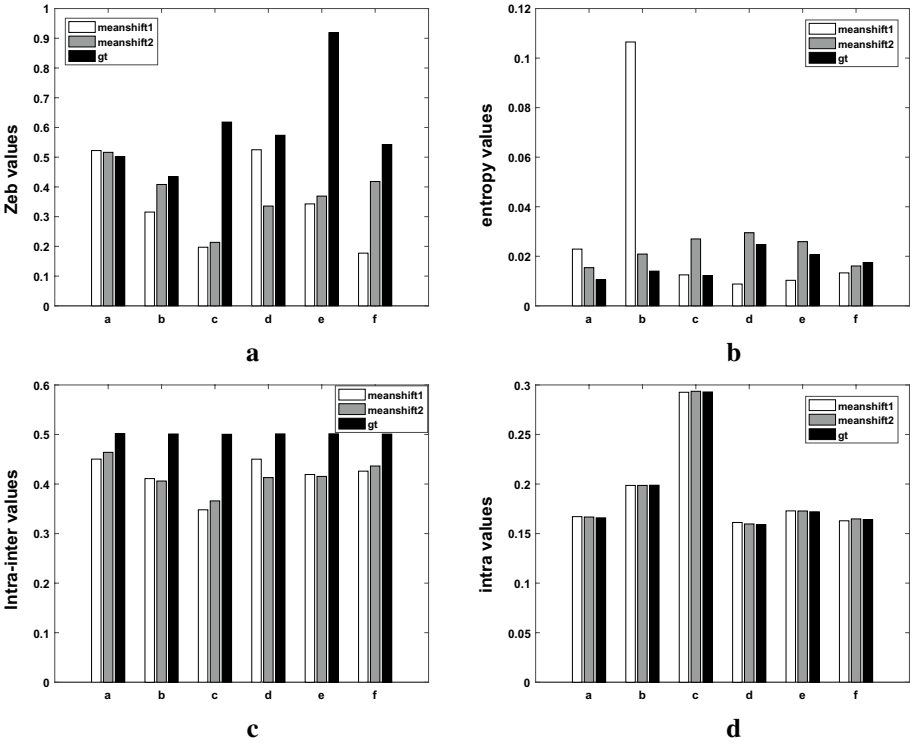


Fig. 10 Histogram of values of different metrics for **a–f**

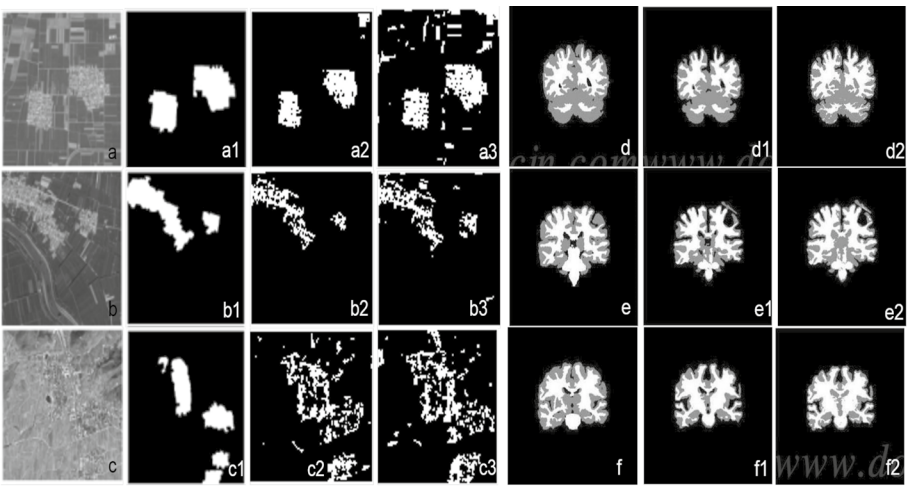


Fig. 11 Segmentation of remote sensing and medical images **a** The first column represents original image, the second column represents ground truth, and the third and fourth columns represent segmentation results obtained with GLSA (global and local saliency analysis model) (Zhang et al. 2016) and FDA (frequency domain analysis) (Zhang and Yang 2014), respectively. **b** The first column represents ground truth, and the second and third columns represent segmentation results obtained by the clustering (Qaddoura et al. 2020) and threshold (Pare et al. 2020) methods, respectively

Table 6 Evaluation of remote sensing and medical image segmentation results

Images methods	Remote sensing images						Medical images					
	a2	a3	b2	b3	c2	c3	d1	d2	e1	e2	f1	f2
Accuracy	0.9287	0.9085	0.9475	0.9253	0.8518	0.8270	0.9645	0.9783	0.9476	0.9506	0.9644	0.9619
Dice	0.7577	0.7280	0.7931	0.7096	0.5166	0.4440	0.9004	0.9407	0.8507	0.8650	0.8888	0.8831
Jaccard	0.6099	0.5723	0.6572	0.5499	0.3483	0.2853	0.8188	0.8881	0.7402	0.7621	0.7998	0.7907
Mcc	0.7170	0.6770	0.7792	0.6788	0.4294	0.3420	0.8822	0.9287	0.8253	0.8369	0.8719	0.8629
F-measure	0.7577	0.7280	0.7931	0.7096	0.5166	0.4440	0.9004	0.9407	0.8507	0.8650	0.8888	0.8831
Precision	0.7244	0.7956	0.6769	0.6140	0.5287	0.4613	0.8409	0.9033	0.7736	0.8193	0.8214	0.8306
Sensitivity	0.7941	0.6709	0.9575	0.8405	0.5051	0.4279	0.9689	0.9814	0.9448	0.9161	0.9681	0.9427
Specificity	0.9507	0.9615	0.9463	0.9356	0.9163	0.9038	0.9637	0.9776	0.9481	0.9578	0.9637	0.9654

of Fig. 9, the order of selecting the metrics should be Zeb, Intra-inter, entropy, and it is not recommended to select intra.

5.3 Experiments on medical and remote sensing image evaluation

Through the analysis in Sect. 4, we know that there are certain differences between medical and remote sensing images, but there are no specific criteria for the selection of these two image segmentation evaluation indicators. The selection of its metrics should be applied on a case-by-case basis, and some classic metrics are usually used. As shown in Fig. 11, we use classical supervised evaluation methods to evaluate medical and remote sensing images. The results obtained in this experiment are listed in Table 6.

Observing Fig. 11 and analyzing the data in Table 6 at the same time, we can see that in the remote sensing image segmentation results, a2 is better than a3, and other metrics except Precision and Specificity have made accurate judgments. All metrics can correctly judge that b2 is better than b3, and the evaluation results for all metrics of c2 and c3 are consistent with the visual evaluation. Also we can see that in the medical image, all judgments of d1 and d2 are correct. For e1 and e2, only Sensitivity did not make an accurate judgment. Regarding the evaluation of f1 and f2, Precision and Specificity have made some mistakes. Through the above analysis, we can prove the effectiveness of the classical evaluation method for medical and remote sensing image evaluation. At the same time, we can also understand that the selection of metrics should be based on specific applications. For example, in the application above, Precision and Specificity should not be used to evaluate b2, b3 and f1, f2. Evaluation of e1 and e2 does not recommend the use of Sensitivity.

6 Discussion

Segmentation assessment is essential if you want to improve the performance of existing segmentation algorithms and develop new efficient segmentation algorithms. So far, image segmentation technology has made significant progress, but the evaluation of these technologies is largely subjective. In general, the validity of a new algorithm is only demonstrated

on a small number of segmented images, or subjectively evaluated by the reader. The evaluation criteria can be used in different applications: comparison of segmentation results, automatic selection of best-fit parameters for a given image segmentation method, or definition of a new segmentation method by optimization.

There are many kinds of images in real life, including natural images, synthetic images, medical images and remote sensing images, so the comparison of segmentation results is also different. It is possible that an evaluation criterion has not achieved good results in the evaluation of natural image segmentation, but it is extremely accurate in the evaluation of medical image segmentation results. Therefore, the same evaluation criteria may have different effectiveness in different applications. Researchers will prove their assessment methods on some images and point out that the results are good. But we have never known from these studies that the technique is only valid in the examples listed in the literature or is still valid in all applications, whether this technique only works for images without textures, and so on. We can see this phenomenon from Fig. 8. Q_p and F-score are better for the measurement of graph a, and Q_p is better for graph b. Most of the metrics have good judgments for graph c, such as PRI, SC, dhd, bgm, Precision, specificity, Q_p, and F-score. In addition to specificity, all metrics apply to graph d. For the metric of graph e we should choose all indicators except Q_p and F-score, and for the segmentation metric of graph f, we should try not to choose VOI, MCC, precision and specificity. Again, Fig. 10 can prove this. For a and f, entropy and intra-inter have better segmentation evaluation results, while entropy and Zeb perform better on b. As for c, Zeb and intra-inter have good assessment. For d, the performance of the four metrics is not very prominent, and the slightly better one is intra-inter. Zeb's evaluation for e performs better, followed by the performance of intra-inter.

In summary, the image segmentation evaluation method has no absolute good or bad, and its use depends on the specific application. Although the supervised evaluation method can obtain more accurate and reliable evaluation results, the evaluation process relies heavily on standard segmentation images, while it is not easy or even impossible to obtain standard reference images in practical applications. Therefore, for applications that have standard segmentation images and require high accuracy evaluation results, we should use a supervised approach. Although the unsupervised evaluation method does not require standard segmentation, so far, it is a low-level data-driven evaluation method based on image features, and it is difficult to have both accuracy and versatility. It can be used for assessments where accuracy requirements are not very high and standard images are difficult to obtain. Analytical methods are used directly for simple evaluation results that can be judged with the naked eye. Thus, for a segmentation result to be evaluated, we can evaluate its quality in order from simple to complex. First, using the analytical method, that is, the naked eye to judge. The second is the unsupervised evaluation method, and the last is the supervised evaluation method.

The future development trend for the performance evaluation of the segmentation algorithm should be able to use the advanced information of the image such as semantic information. For example, the overall frame of an image is a comprehensive description of the image content, so an evaluation method combining the overall frame information may obtain a more accurate understanding of the segmentation result image. In addition, if the evaluation process can be combined with the segmentation effect that humans hope to achieve, the evaluation method will represent the subjective will of human beings to the greatest extent, so the evaluation results will be more accurate and reliable.

7 Conclusion

This paper classifies the image segmentation quality evaluation standards, and mainly summarizes the two types of supervised evaluation and unsupervised evaluation. In addition to natural images, segmentation methods are most commonly used in the fields of medical and remote sensing images. Therefore, many evaluation criteria for medical and remote sensing image segmentation results have been proposed. This article outlines the application of metrics in natural, medical, and remote sensing image evaluation, and further introduces supervised and unsupervised evaluation methods in the application. Experiments are carried out on some of the evaluation criteria, and the priorities of their evaluations were sorted. Through comparison of experiments, we found that different metrics have different evaluation results for the same image, and the same metrics have different evaluation results for different images. Therefore, the choice of image segmentation quality assessment criteria depends on the specific application.

Funding This study was funded by National Natural Science Foundation of China (Grant No. 61201421).

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

Angulo J, Velasco-Forero S, Chanussot J (2009) Multiscale stochastic watershed for unsupervised hyperspectral image segmentation. In: 2009 IEEE international geoscience and remote sensing symposium, vol 3, pp III-93–III-96

Arhid K, Bouksim M, Zakani FR, Aboulfatah M, Gadi T (2016) New evaluation method using sampling theory to evaluate 3D segmentation algorithms. In: ElMohajir M, Chahhou M, AlAchhab M, ElMohajir BE (eds) 2016 4th IEEE international colloquium on information science and technology (CIST), pp 410–415

Aspert N, Santa-Cruz D, Ebrahimi T (2002) Mesh: Measuring errors between surfaces using the Hausdorff distance. In: Proceedings of the IEEE international conference on multimedia and expo, vol I and II, pp 705–708. https://doi.org/10.1109/ICME.2002.1035879

Benes M, Zitova B (2015) Performance evaluation of image segmentation algorithms on microscopic image data. J Microsc 257(1):65–85. https://doi.org/10.1111/jmi.12186

Berezsky O, Melnyk G, Batko Y, Pitsun O (2016) Regions matching algorithms analysis to quantify the image segmentation results. In: 2016 XITH international scientific and technical conference computer sciences and information technologies (CSIT), pp 33–36

Bernard O, Bosch JG, Heyde B (2016) Standardized evaluation system for left ventricular segmentation algorithms in 3D echocardiography. IEEE Trans Med Imaging 35(4):967–977. https://doi.org/10.1109/TMI.2015.2503890

Boeck S, Immitzer M, Atzberger C (2017) On the objectivity of the objective function-problems with unsupervised segmentation evaluation based on global score and a possible remedy. Remote Sens 9(8):2017. https://doi.org/10.3390/rs9080769

Borsotti M, Campadelli P, Schettini R (1998) Quantitative evaluation of color image segmentation results. Pattern Recognit Lett 19(8):741–747. https://doi.org/10.1016/S0167-8655(98)00052-X

Cai Z, Liang Y, Huang H (2017) Unsupervised segmentation evaluation: an edge-based method. Multimed Tools Appl 76(8):11097–11110. https://doi.org/10.1007/s11042-016-3542-8

Cappabianco FAM, de Miranda PAV, Udupa JK (2017) A critical analysis of the methods of evaluating MRI brain segmentation algorithms. In: 2017 IEEE international conference on image processing (ICIP), pp 3894–3898

Cappabianco FAM, Ribeiro PFO, de Miranda PAV, Udupa JK (2019) A general and balanced region-based metric for evaluating medical image segmentation algorithms. In: 2019 IEEE international conference on image processing (ICIP), pp 1525–1529

Cardoso J, Corte-Real L (2005) Toward a generic evaluation of image segmentation. IEEE Trans Image Process 14(11):1773–1782. https://doi.org/10.1109/TIP.2005.854491

Chabrier S, Emile B, Laurent H, Rosenberger C, Marche P (2004) Unsupervised evaluation of image segmentation application to multi-spectral images. In: Proceedings of the 17th international conference on pattern recognition, vol 1, pp 576–579. https://doi.org/10.1109/ICPR.2004.1334206

Chang H-H, Zhuang AH, Valentino DJ, Chu W-C (2009) Performance measure characterization for evaluating neuroimage segmentation algorithms. Neuroimage 47(1):122–135. https://doi.org/10.1016/j.neuroimage.2009.03.068

Chen Z, Zhu H (2019) Visual quality evaluation for semantic segmentation: subjective assessment database and objective assessment measure. IEEE Trans Image Process 28(12):5785–5796

Chen Y, Ming D, Zhao L, Lv B, Zhou K, Qing Y (2018) Review on high spatial resolution remote sensing image segmentation evaluation. Photogramm Eng Remote Sens 84(10):629–646. https://doi.org/10.14358/PERS.84.10.629

Chen H, Wang S (2004) The use of visible color difference in the quantitative evaluation of color image segmentation. In: 2004 IEEE international conference on acoustics, speech, and signal processing, vol III, pp 593–596

Chouhan SS, Kaul A, Singh UP (2018) Soft computing approaches for image segmentation: a survey. Multimed Tools Appl 77(21):28483–28537. https://doi.org/10.1007/s11042-018-6005-6

Correia P, Pereira F (2003) Objective evaluation of video segmentation quality. IEEE Trans Image Process 12(2):186–200. https://doi.org/10.1109/TIP.2002.807355

Cruz H, Eckert M, Meneses JM, Martinez JF (2017) Fast evaluation of segmentation quality with parallel computing. Sci Program. https://doi.org/10.1155/2017/5767521

Dey N, Rajinikanth V, Ashour AS (2018) Tavares JMRS social group optimization supported segmentation and evaluation of skin melanoma images. Symmetry-Basel. https://doi.org/10.3390/sym10020051

Dogra DP, Majumdar AK, Sural S (2012) Evaluation of segmentation techniques using region area and boundary matching information. J Vis Commun Image Represent 23(1):150–160. https://doi.org/10.1016/j.jvcir.2011.09.005

Domingo J, Dura E, Goceri E (2016) Iteratively learning a liver segmentation using probabilistic atlases: preliminary results. In: 2016 15th IEEE international conference on machine learning and applications (ICMLA 2016), pp 593–598. https://doi.org/10.1109/ICMLA.2016.194

Eftekhari-Moghadam A-M, Abdechiri M (2010) An unsupervised evaluation method based on probability density function. In: IEEE international symposium on industrial electronics (ISIE 2010), pp 1573–1578

Erdem C, Sankur B, Tekalp A (2004) Performance measures for video object segmentation and tracking. IEEE Trans Image Process 13(7):937–951. https://doi.org/10.1109/TIP.2004.828427

Feng Y, Shen X, Chen H, Zhang X (2016) A weighted-ROC graph based metric for image segmentation evaluation. Signal Process 119:43–55. https://doi.org/10.1016/j.sigpro.2015.07.010

Fernandez MA, Lopes RM, Hirata NST (2015) Image segmentation assessment from the perspective of a higher level task. In: 2015 28th SIBGRAPI conference on graphics, patterns and images, pp 111–118. https://doi.org/10.1109/SIBGRAPI.2015.46

Flores FC, Lotufo RdA (2008) Benchmark for quantitative evaluation of assisted object segmentation methods to image sequences. In: SIBGRAPI 2008: XXI Brazilian symposium on computer graphics and image processing, pp 95–102. https://doi.org/10.1109/SIBGRAPI.2008.22

Gao H, Tang Y, Jing L, Li H, Ding H (2017) A novel unsupervised segmentation quality evaluation method for remote sensing images. Sensors. https://doi.org/10.3390/s17102427

Garcia-Lamont F, Cervantes J, Lopez A, Rodriguez L (2018) Segmentation of images by color features: a survey. Neurocomputing 292:1–27. https://doi.org/10.1016/j.neucom.2018.01.091

Gautam AK, Bhutiyani MR (2016) Performance evaluation of hyperspectral image segmentation implemented by recombination of pct and bilateral filter based fused images. In: 2016 3rd international conference on signal processing and integrated networks (SPIN), pp 152–156

Ge Feng, Wang Song, Liu Tiecheng (2006) Image-segmentation evaluation from the perspective of salient object extraction. In: 2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06), vol 1, pp 1146–1153

Getto R, Kuijper A, von Landesberger T (2015) Extended surface distance for local evaluation of 3D medical image segmentations. Vis Comput 31(6–8):989–999. https://doi.org/10.1007/s00371-015-1113-z

Göçeri E (2013) A comparative evaluation for liver segmentation from spir images and a novel level set method using signed pressure force function. Thesis (Doctoral)–Izmir Institute of Technology, Electronics and Communication Engineering

Goceri E (2016) Automatic labeling of portal and hepatic veins from MR images prior to liver transplantation. Int J Comput Assist Radiol Surg 11(12):2153–2161. https://doi.org/10.1007/s11548-016-1446-8

Goceri E (2018) A method for leukocyte segmentation using modified gram-schmidt orthogonalization and expectation-maximization. In: International conference on applied analysis and mathematical modeling ICAAMM18, Istanbul, Turkey

Goceri E (2019a) Analysis of deep networks with residual blocks and different activation functions: classification of skin diseases. In: 2019 ninth international conference on image processing theory, tools and applications (IPTA), pp 1–6

Goceri E (2019b) Challenges and recent solutions for image segmentation in the era of deep learning. In: 2019 ninth international conference on image processing theory, tools and applications (IPTA), pp 1–6

Goceri E (2019c) Diagnosis of Alzheimer's disease with Sobolev gradient-based optimization and 3D convolutional neural network. Int J Numer Methods Biomed Eng. https://doi.org/10.1002/cnm.3225

Goceri E, Dura E (2015a) Artificial neural network based abdominal organ segmentations: a review. In: 2015 IEEE 14th international conference on machine learning and applications (ICMLA), pp 1191–1194. https://doi.org/10.1109/ICMLA.2015.231

Goceri N, Goceri E (2015b) A neural network based kidney segmentation from MR images. In: 2015 IEEE 14th international conference on machine learning and applications (ICMLA), pp 1195–1198

Goceri E, Songül C (2017a) Automated detection and extraction of skull from mr head images: preliminary results. In: 2017 international conference on computer science and engineering (UBMK), pp 171–176

Goceri E, Songul C (2017b) Computer-based segmentation, change detection and quantification for lesions in multiple sclerosis. In: Adali E (ed) 2017 International conference on computer science and engineering (UBMK), pp 177–182

Goceri E, Songul C (2018) Biomedical information technology: image based computer aided diagnosis systems. In: International conference on advanced technologies, Antalya

Goceri E, Unlu MZ, Dicle O (2015a) A comparative performance evaluation of various approaches for liver segmentation from SPIR images. Turk J Electr Eng Comput Sci 23(3):741–768. https://doi.org/10.3906/elk-1304-36

Goceri E, Shah ZK, Gurcan MN (2017b) Vessel segmentation from abdominal magnetic resonance images: adaptive and reconstructive approach. Int J Numer Methods Biomed Eng. https://doi.org/10.1002/cnm.2811

Habba M, Ameur M, Jabrane Y (2018) A novel Gini index based evaluation criterion for image segmentation. Optik 168:446–457. https://doi.org/10.1016/j.ijleo.2018.04.045

Henderson P, Ferrari V (2017) End-to-end training of object class detectors for mean average precision. In: Computer vision—ACCV 2016 PT V, vol 10115, pp 198–213. https://doi.org/10.1007/978-3-319-54193-8_13

Hoang HS, Phuong Pham C, Franklin D, van Walsum T, Ha Luu M (2019) An evaluation of CNN-based liver segmentation methods using multi-types of ct abdominal images from multiple medical centers. In: 2019 19th international symposium on communications and information technologies (ISCIT), pp 20–25

Huang C, Wu Q, Meng F (2016) Qualitynet: Segmentation quality evaluation with deep convolutional networks. In: 2016 visual communications and image processing (VCIP), pp 1–4

Jianqing Liu, Yee-Hong Yang (1994) Multiresolution color image segmentation. IEEE Trans Pattern Anal Mach Intell 16:689–700

Jinping L, Weihua G, Qing C, Zhaohui T, Chunhua Y (2013) An unsupervised method for flotation froth image segmentation evaluation base on image gray-level distribution. In: 2013 32nd Chinese control conference (CCC), pp 4018–4022

Johnson B, Xie Z (2011) Unsupervised image segmentation evaluation and refinement using a multi-scale approach. ISPRS J Photogramm Remote Sens 66(4):473–483. https://doi.org/10.1016/j.isprsjprs.2011.02.006

Jordan J, Angelopoulou E (2012) Supervised multispectral image segmentation with power watersheds. In: 2012 19th IEEE international conference on image processing, pp 1585–1588

Karimi S, Jiang X, Cosman P, Martz H (2014) Flexible methods for segmentation evaluation: results from CT-based luggage screening. J X-Ray Sci Technol 22(2):175–195. https://doi.org/10.3233/XST-140418

Kaya B, Goceri E, Becker A, Elder B, Puduvalli V, Winter J, Gurcan M, Otero JJ (2017) Automated fluorescent miscroscopic image analysis of PTBP1 expression in glioma. PLoS ONE 12(3):e0170991. https://doi.org/10.1371/journal.pone.0170991

Khan JF, Bhuiyan SM (2014) Weighted entropy for segmentation evaluation. Opt Laser Technol 57(SI):236–242. https://doi.org/10.1016/j.optlastec.2013.07.012

Khan J, Bhuiyan S (2011) Evaluation of the number of segments using weighted entropy. In: Proceedings SSST 2011: 43rd IEEE southeastern symposium on system theory, pp 173–178

Kirillov A, He K, Girshick R, Rother C, Dollár P (2019) Panoptic segmentation. In: 2019 IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp 9396–9405

Kubassova O, Boesen M, Bliddal H (2008) General framework for unsupervised evaluation of quality of segmentation results. In: 2008 15th IEEE international conference on image processing, vol 1–5, pp 3036–3039. https://doi.org/10.1109/ICIP.2008.4712435

Laurent P, Cresson T, Vazquez C, Hagemeister N, de Guise JA (2016) A multi-criteria evaluation platform for segmentation algorithms. In: 2016 38th annual international conference of the IEEE engineering in medicine and biology society (EMBC), pp 6441–6444

Ledig C, Shi W, Bai W, Rueckert D (2014) Patch-based evaluation of image segmentation. In: 2014 IEEE conference on computer vision and pattern recognition (CVPR), pp 3065–3072. https://doi.org/10.1109/CVPR.2014.392

Levine M, Nazif A (1985) Dynamic measurement of computer generated image segmentations. IEEE Trans Pattern Anal Mach Intell 7(2):155–164. https://doi.org/10.1109/TPAMI.1985.4767640

Li Peijun, Xiao Xiaobai (2004) Evaluation of multiscale morphologicala segmentation of multispectral imagery for land cover classification. IGARSS 2004. In: 2004 IEEE international geoscience and remote sensing symposium, vol 4, pp 2676–2679

Li H, Zhao X, Su A, Zhang H, Liu J, Gu G (2020) Color space transformation and multi-class weighted loss for adhesive white blood cell segmentation. IEEE Access 8:24808–24818

Liu H, Peng C, Yu C, Wang J, Liu X, Yu G, Jiang W (2019) An end-to-end network for panoptic segmentation. In: 2019 IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp 6165–6174

Lukac P, Hudec R, Benco M, Kamencay P, Dubcova Z, Zacharasova M (2011) Simple comparison of image segmentation algorithms based on evaluation criterion. In: Proceedings of the 21st international conference—radioelektronika 2011, pp 233–236. https://doi.org/10.1109/RADIOELEK.2011.5936406

Luu HM, Klink C, Moelker A, Niessen W, van Walsum T (2015) Quantitative evaluation of noise reduction and vesselness filters for liver vessel segmentation on abdominal CTA images. Phys Med Biol 60(10):3905–3926. https://doi.org/10.1088/0031-9155/60/10/3905

Lu Y, Wan Y, Li G (2016) Notice of removal:scale-constrained unsupervised evaluation method for multiscale image segmentation. In: 2016 IEEE international conference on image processing (ICIP), pp 2559–2563

Mageswari SU, Mala C (2014) Analysis and performance evaluation of various image segmentation methods. In: 2014 international conference on contemporary computing and informatics (IC3I), pp 469–474

Malladi SRSP, Ram S, Rodriguez JJ (2018) A ground-truth fusion method for image segmentation evaluation. In: 2018 IEEE southwest symposium on image analysis and interpretation (SSIAI), pp 137–140

Mantilla SCL, Yari Y (2017) Multispectral images segmentation using fuzzy probabilistic local cluster for unsupervised clustering. In: 2017 IEEE Latin American conference on computational intelligence (LA-CCI), pp 1–5

Marpu PR, Neubert M, Herold H, Niemeyer I (2010) Enhanced evaluation of image segmentation results. J Spatial Sci 55(1):55–68. https://doi.org/10.1080/14498596.2010.487850

Mendrik AM, Vincken KL, Kuijf HJ (2015) MRBrainS challenge: online evaluation framework for brain image segmentation in 3T MRI scans. Comput Intell Neurosci. https://doi.org/10.1155/2015/813696

Monteiro FC, Campilho AC (2012) Distance measures for image segmentation evaluation. In: Numerical analysis and applied mathematics (ICNAAM 2012), volume A and B. American Institute of Physics, vol 1479, pp 794–797. https://doi.org/10.1063/1.4756257

Nogueira K, Dalla Mura M, Chanussot J, Schwartz WR, dos Santos JA (2019) Dynamic multicontext segmentation of remote sensing images based on convolutional networks. IEEE Trans Geosci Remote Sens 57(10):7503–7520

Pal N, Bhandari D (1993) Image thresholding: some new techniques. Signal Process 33(2):139–158. https://doi.org/10.1016/0165-1684(93)90107-L

Pare S, Kumar A, Singh GK, Bajaj V (2020) Image segmentation using multilevel thresholding: a research review. Iran J Sci Technol Trans Electr Eng 44(1):1–29. https://doi.org/10.1007/s40998-019-00251-1

Peng B, Li T (2013) A probabilistic measure for quantitative evaluation of image segmentation. IEEE Signal Process Lett 20(7):689–692. https://doi.org/10.1109/LSP.2013.2262938
Peng R, Varshney PK (2015) On performance limits of image segmentation algorithms. Comput Vis Image Underst 132:24–38. https://doi.org/10.1016/j.cviu.2014.11.004
Peng B, Wang X, Yang Y (2016) Region based exemplar references for image segmentation evaluation. IEEE Signal Process Lett 23(4):459–462. https://doi.org/10.1109/LSP.2016.2517101
Peng B, Zhang L, Mou X, Yang M-H (2017) Evaluation of segmentation quality via adaptive composition of reference segmentations. IEEE Trans Pattern Anal Mach Intell 39(10):1929–1941. https://doi.org/10.1109/TPAMI.2016.2622703
Peng B, Simfukwe M, Li T (2018) Region-based image segmentation evaluation via perceptual pooling strategies. Mach Vis Appl 29(3):477–488. https://doi.org/10.1007/s00138-017-0903-x
Peng C, Li Y, Jiao L, Chen Y, Shang R (2019) Densely based multi-scale and multi-modal fully convolutional networks for high-resolution remote-sensing image semantic segmentation. IEEE J Sel Top Appl Earth Observ Remote Sens 12(8):2612–2626
Philipp-Foliguet S, Guigues L (2006) New criteria for evaluating image segmentation results. In: 2006 IEEE international conference on acoustics, speech and signal processing, vol 1–13, pp 1357–1360
Pont-Tuset J, Marques F (2016) Supervised evaluation of image segmentation and object proposal techniques. IEEE Trans Pattern Anal Mach Intell 38(7):1465–1478. https://doi.org/10.1109/TPAMI.2015.2481406
Pont-Tuset J, Marques F (2013) Measures and meta-measures for the supervised evaluation of image segmentation. In: 2013 IEEE conference on computer vision and pattern recognition (CVPR), pp 2131–2138. https://doi.org/10.1109/CVPR.2013.277
Poudel P, Illanes A, Sheet D, Friebe M (2018) Evaluation of commonly used algorithms for thyroid ultrasound images segmentation and improvement using machine learning approaches. J Healthc Eng. https://doi.org/10.1155/2018/8087624
Prabha DS, Kumar JS (2016) Performance evaluation of image segmentation using objective methods. Indian J Sci Technol. https://doi.org/10.17485/ijst/2016/v9i8/87907
Qaddoura R, Faris H, Aljarah I (2020) An efficient clustering algorithm based on the k-nearest neighbors with an indexing ratio. Int J Mach Learn Cybern 11(3):675–714. https://doi.org/10.1007/s13042-019-01027-z
Roman-Roldan R, Gomez-Lopera J, Atae-Allah C, Martinez-Aroza J, Luque-Escamilla P (2001) A measure of quality for evaluating methods of segmentation and edge detection. Pattern Recognit 34(5):969–980. https://doi.org/10.1016/S0031-3203(00)00052-2
Rosenberger C, Chehdi K (2000) Genetic fusion: application to multi-components image segmentation. In: 2000 IEEE international conference on acoustics, speech, and signal processing, vol 4, pp 2223–2226
Sahoo P, Soltani S, Wong A, Chen Y (1988) A survey of thresholding techniques. Comput Vis Graph Image Process 41(2):233–260. https://doi.org/10.1016/0734-189X(88)90022-9
Saqui D, Saito JH, de Lima DC, Jorge LADC, Ferreira EJ, Ataky STM, Fambrini F (2019) Nsga2-based method for band selection for supervised segmentation in hyperspectral imaging. In: 2019 IEEE international conference on systems, man and cybernetics (SMC), pp 3580–3585
Shan P (2018) Image segmentation method based on K-mean algorithm. EURASIP J Image Video Process. https://doi.org/10.1186/s13640-018-0322-6
Sharma NK, Ronak S, Nema MK, Rakshit S (2010) Statistical evaluation of image segmentation. In: 2010 IEEE 2nd international advance computing conference, pp 101–105. https://doi.org/10.1109/IADCC.2010.5423030
Shi R, Ngan KN, Li S, Paramesran R, Li H (2015) Visual quality evaluation of image object segmentation: subjective assessment and objective measure. IEEE Trans Image Process 24(12):5033–5045. https://doi.org/10.1109/TIP.2015.2473099
Shi W, Meng F, Wu Q (2017) Segmentation quality evaluation based on multi-scale convolutional neural networks. In: 2017 IEEE visual communications and image processing (VCIP), pp 1–4
Shi R, Ngan KN, Li S (2014) Jaccard index compensation for object segmentation evaluation. In: 2014 IEEE international conference on image processing (ICIP), pp 4457–4461
Shi R, Ngan KN, Li S (2017) Objectness based unsupervised object segmentation quality evaluation. In: 2017 seventh international conference on information science and technology (ICIST2017), pp 256–258
Skalski A, Jakubowski J, Drewniak T (2018) LEFMIS: locally-oriented evaluation framework for medical image segmentation algorithms. Phys Med Biol 63(16):2018. https://doi.org/10.1088/1361-6560/aad316
Srubar S (2012) Quality measurement of image segmentation evaluation methods. In: 8th international conference on signal image technology & internet based systems (SITIS 2012), pp 254–258

Strasters K, Gerbrands J (1991) Three-dimensional image segmentation using a split, merge and group approach. Pattern Recognit Lett 12(5):307–325. https://doi.org/10.1016/0167-8655(91)90414-H
Su T (2018) An improved unsupervised image segmentation evaluation approach based on under- and over- segmentation aware. Ann Photogramm Remote Sens Spatial Inf Sci 4:197–204
Su T, Zhang S (2017) Local and global evaluation for remote sensing image segmentation. ISPRS J Photogramm Remote Sens 130:256–276. https://doi.org/10.1016/j.isprsjprs.2017.06.003
Sundara SM, Aarthi R (2019) Segmentation and evaluation of white blood cells using segmentation algorithms. In: 2019 3rd international conference on trends in electronics and informatics (ICOEI), pp 1143–1146
Taha AA, Hanbury A, del Toro OAJ (2014) A formal method for selecting evaluation metrics for image segmentation. In: 2014 IEEE international conference on image processing (ICIP), pp 932–936
Taha AA, Hanbury A (2015) Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool. BMC Med Imaging. https://doi.org/10.1186/s12880-015-0068-x
Tang Y, Zhao L, Ren L (2019) Different versions of entropy rate superpixel segmentation for hyperspectral image. In: 2019 IEEE 4th international conference on signal and image processing (ICSIP), pp 1050–1054
Unnikrishnan R, Pantofaru C, Hebert M (2007) Toward objective evaluation of image segmentation algorithms. IEEE Trans Pattern Anal Mach Intell 29(6):929–944. https://doi.org/10.1109/TPAMI.2007.1046
Vedaldi A, Lenc K (2015) MatConvNet convolutional neural networks for MATLAB. In: MM'15: proceedings of the 2015 acm multimedia conference, pp 689–692. https://doi.org/10.1145/2733373.2807412
Wang Y, Qi Q, Liu Y (2018) Unsupervised segmentation evaluation using area-weighted variance and jeffries–Matusita distance for remote sensing images. Remote Sens 10(8):2018. https://doi.org/10.3390/rs10081193
Wang Y, Qi Q, Jiang L, Liu Y (2020) Hybrid remote sensing image segmentation considering intrasegment homogeneity and intersegment heterogeneity. IEEE Geosci Remote Sens Lett 17(1):22–26
Wiesmann V, Bergler M, Palmisano R, Prinzen M, Franz D, Wittenberg T (2017) Using simulated fluorescence cell micrographs for the evaluation of cell image segmentation algorithms. BMC Bioinform. https://doi.org/10.1186/s12859-017-1591-2
Wu J, Li B, Ni W, Yan W, Zhang H (2019) Optimal segmentation scale selection for object-based change detection in remote sensing images using Kullback–Leibler divergence. IEEE Geosci Remote Sens Lett. https://doi.org/10.1109/LGRS.2019.2943406
Xia Y, Zhang B, Coenen F (2016) Face occlusion detection using deep convolutional neural networks. Int J Pattern Recognit Artif Intell. https://doi.org/10.1142/S0218001416600107
Yan Z, Yang X, Cheng K-T (2018) Joint segment-level and pixel-wise losses for deep learning based retinal vessel segmentation. IEEE Trans Biomed Eng 65(9):1912–1923. https://doi.org/10.1109/TBME.2018.2828137
Yang J, Li P, He Y (2014) A multi-band approach to unsupervised scale parameter selection for multi-scale image segmentation. ISPRS J Photogramm Remote Sens 94:13–24. https://doi.org/10.1016/j.isprsjprs.2014.04.008
Yang J, He Y, Caspersen J, Jones T (2015) A discrepancy measure for segmentation evaluation from the perspective of object recognition. ISPRS J Photogramm Remote Sens 101:186–192. https://doi.org/10.1016/j.isprsjprs.2014.12.015
Ye P, Kumar J, Kang L, Doermann D (2012) Unsupervised feature learning framework for no-reference image quality assessment. In: 2012 IEEE conference on computer vision and pattern recognition (CVPR), pp 1098–1105
Zagoruyko S, Komodakis N (2015) Learning to compare image patches via convolutional neural networks. In: 2015 IEEE conference on computer vision and pattern recognition (CVPR), pp 4353–4361
Zeng Y, Niu X, Dou Y (2019) Aircraft segmentation from remote sensing image by transferring natual image trained forground extraction CNN model. In: 2019 IEEE 4th international conference on signal and image processing (ICSIP), pp 817–822
Zhang Hui, Cholleti S, Goldman SA, Fritts JE (2006) Meta-evaluation of image segmentation using machine learning. In: 2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06), vol 1, pp 1138–1145
Zhang Y (1996) A survey on evaluation methods for image segmentation. Pattern Recognit 29(8):1335–1346. https://doi.org/10.1016/0031-3203(95)00169-7
Zhang L, Yang K (2014) Region-of-interest extraction based on frequency domain analysis and salient region detection for remote sensing image. IEEE Geosci Remote Sens Lett 11:916–920

Zhang H, Fritts J, Goldman S (2004) An entropy-based objective evaluation method for image segmentation. Storage Retr Methods Appl Multimed 5307(2004):38–49

Zhang H, Fritts JE, Goldman SA (2008) Image segmentation evaluation: a survey of unsupervised methods. Comput Vis Image Underst 110(2):260–280. https://doi.org/10.1016/j.cviu.2007.08.003

Zhang X, Xiao P, Feng X (2012) An unsupervised evaluation method for remotely sensed imagery segmentation. IEEE Geosci Remote Sens Lett 9(2):156–160. https://doi.org/10.1109/LGRS.2011.2163056

Zhang X, Feng X, Xiao P, He G, Zhu L (2015) Segmentation quality evaluation using region-based precision and recall measures for remote sensing images. ISPRS J Photogramm Remote Sens 102:73–84. https://doi.org/10.1016/j.isprsjprs.2015.01.009

Zhang L, Li A, Zhang Z, Yang K (2016) Global and local saliency analysis for the extraction of residential areas in high-spatial-resolution remote sensing image. IEEE Trans Geosci Remote Sens 54(7):3750–3763. https://doi.org/10.1109/TGRS.2016.2527044

Zhang L, Ma J, Lv X, Chen D (2020) Hierarchical weakly supervised learning for residential area semantic segmentation in remote sensing images. IEEE Geosci Remote Sens Lett 17(1):117–121

Zhao Y, Hao K, He H, Tang X, Wei B (2020) A visual long-short-term memory based integrated CNN model for fabric defect image classification. Neurocomputing 380:259–270. https://doi.org/10.1016/j.neucom.2019.10.067

Zhao Q, Liu F, Zhang L, Zhang D (2010) A comparative study on quality assessment of high resolution fingerprint images. In: 2010 IEEE international conference on image processing, pp 3089–3092. https://doi.org/10.1109/ICIP.2010.5648800

Zhou S, Nie D, Adeli E, Yin J, Lian J, Shen D (2020) High-resolution encoder-decoder networks for low-contrast medical image segmentation. IEEE Trans Image Process 29:461–475

Ziolko B, Emms D, Ziolko M (2018) Fuzzy evaluations of image segmentations. IEEE Trans Fuzzy Syst 26(4):1789–1799. https://doi.org/10.1109/TFUZZ.2017.2752130