Visible colour difference-based quantitative evaluation of colour segmentation

H.-C. Chen and S.-J. Wang

Abstract: The authors present the use of visible colour difference in a new quantitative evaluation scheme for colour segmentation. To avoid directly evaluating the subjectively perceived quality of colour segmentation, two objective visual quantities, the quantity of missing boundaries and the quantity of fake boundaries, are considered. To explore how missing boundaries and fake boundaries affect the perceived quality of colour segmentation, a few visual rating experiments are made. On the other hand, to fit for humans' visual perception on colour difference, the visible colour difference is defined. Based on the experiments and the visible colour difference, two measures, named intra-region visual error and inter-region visual error, are designed to estimate the degrees of missing boundaries and fake boundaries, respectively. With these two measures, a complete scheme for the evaluation of colour segmentation is proposed. The simulation results demonstrate that this new scheme may evaluate segmentation results without any ground truth, and could help the automatic selection of parameter settings for a given segmentation algorithm.

1 Introduction

Colour segmentation is a crucial step in image analysis and pattern recognition. The performance of colour segmentation may significantly affect the quality of an image understanding system. So far, hundreds of colour segmentation algorithms have already been developed to deal with various kinds of image-related applications [1, 2]. Among these colour segmentation algorithms, the automatic setting of controlling parameters is usually a difficult task. These control parameters are often adjusted by the users in an interactive and tiresome manner. Moreover, the selection of control parameters is also image dependent. For most colour segmentation algorithms, there exists no parameter setting that is universally applicable.

On the other hand, it is well known that performance evaluation of segmentation algorithms is critical and essential in the development of image understanding systems. However, as compared with the tremendous efforts spent in the development of segmentation algorithms, relatively fewer efforts have been spent on the subject of image segmentation evaluation [3–7]. According to the classification scheme proposed by Zhang [6, 7], existing evaluation methods for image segmentation could be roughly classified into three categories: (1) analytical methods, (2) discrepancy methods and (3) goodness methods. As shown in Fig. 1, analytical methods directly evaluate segmentation algorithms by analysing their principles, requirements, utilities, complexity and so on [7]. On the contrary, both discrepancy methods and goodness methods evaluate the performance of segmentation by judging the quality of segmentation results. Especially, discrepancy methods measure the difference between the segmentation result and the reference result, which is usually an expected result or a ground truth [8, 9]. On the other hand, goodness methods evaluate the segmentation results with certain quality measures directly, without the use of any reference result [10–13].

Due to the lack of a general theory for image segmentation, analytical methods work well only for some particular models or for some desirable properties of the algorithms. Moreover, these analytical methods themselves are seldom to be used alone [7]. For discrepancy methods, the reference result is essential for the evaluation of segmentation. However, the acquirement of reference results is usually non-trivial, and the acquired reference results are usually user-dependent [8]. Hence, in normal circumstances, the third type of methods, the goodness methods, tends to be more practical. For this type of method, a given algorithm can be evaluated by simply computing some goodness measures over the segmentation results. So far, several goodness measures have already been proposed [10–13]. For example, based on the total number of segmented regions and a colour difference defined in the RGB colour space, evaluation functions are proposed by Liu and Yang [10] and Borsotti *et al.* [11] to measure the difference between the original image and the segmented image. In order to avoid segmenting an image into too many small regions, the factor of region area is often considered in these evaluation functions.

Although several goodness methods have already been proposed, not too many of them are directly based on human visual perception. Instead, most goodness methods combine several existing measures together to formulate their evaluation functions. The selection and the combination of different measures are usually subjective. The adjustment of weighting coefficients is often troublesome. Moreover, some commonly used measures, such as the number of homogeneous regions, could be very different for different images. When these image-dependent measures are combined into a single evaluation function,

IEE Proceedings online no. 20045221

doi:10.1049/ip-vis:20045221

Paper first received 23rd October 2004 and in revised form 24th October 2005

The authors are with the Department of Electronics Engineering, Institute of Electronics, National Chiao Tung University, Hsin-Chu 30050, Taiwan, Republic of China

E-mail: hsinchia.ee88g@nctu.edu.tw

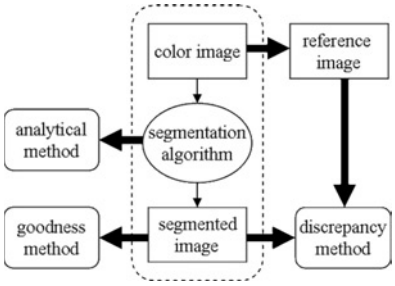


Fig. 1 *Approaches for evaluating image segmentation [6, 7]*

it would be fairly difficult to perform segmentation evaluation, without prior knowledge of image contents.

In this article, we propose a new evaluation scheme that is basically a goodness approach. To mimic the way a human perceives the performance of segmentation, two objective visual quantities, the quantity of missing boundaries and the quantity of fake boundaries, are considered and a set of visual rating experiments was made. Moreover, to fit for humans' visual perception on colour difference, a so-called 'visible colour difference' is defined. Based on these visual experiments and the defined visible colour difference, two measures, named 'intra-region visual error' and 'inter-region visual error', are designed to estimate the degrees of 'visible' false negative and 'visible' false positive, respectively. Based on these two error measurements, a complete scheme is then proposed to evaluate the results of colour segmentation. This evaluation scheme may not only evaluate the segmentation results without any ground truth, but could also be used to assist the selection of parameter settings for a given segmentation algorithm.

2 Existing goodness methods and visual rating experiments

2.1 Existing goodness methods

As mentioned earlier, most existing goodness methods combine several different measures together to compose an evaluation function that fits for visual judgement [10–13]. However, not all of them are directly based on human visual perception. Moreover, the adaptation of some measures in the evaluation function may require some prior knowledge of image contents. For example, Liu and Yang [10] define an evaluation function as

$$F(\mathrm{I}) = \frac{1}{1000(N \times M)} \sqrt{R} \times \sum_{i=1}^{R} \frac{e_i^2}{\sqrt{A_i}} \tag{1}$$

where I is the image to be segmented, R is the number of regions in the segmented image, e_i is the colour error of the ith region, A_i is the area of the ith region and N and M represent the length and the width of the image. Here, e_i is defined as the sum of the Euclidean distance of the colour vectors between the original image and the segmented image in the ith region. Note that, in (1), the square root of region number is adopted in the evaluation function to avoid segmenting the image into too many regions.

Based on (1), two further improved evaluation functions are proposed by Borsotti *et al.* [11] that are defined as

$$F'(\mathrm{I}) = \frac{1}{10\,000(N \times M)} \sqrt{\sum_{A=1}^{\mathrm{Max}} [R(A)]^{1+1/A}} \sum_{i=1}^{R} \frac{e_i^2}{\sqrt{A_i}} \tag{2}$$

and

$$Q(\mathrm{I}) = \frac{1}{10\,000(N \times M)} \sqrt{R} \sum_{i=1}^{R} \left[\frac{e_i^2}{1 + \log A_i} + \left(\frac{R(A_i)}{A_i} \right)^2 \right] \tag{3}$$

where $R(A_i)$ represents the number of regions with area size A_i. In both equations, the areas of regions are considered in the evaluation functions to punish these segmentation results with too many small regions. Similarly, the number of segmented regions is also included in these two equations to achieve segmented results with an appropriate number of homogeneous regions.

For these evaluation functions, two primary requirements are adopted for preferred segmentation: smaller colour difference and a smaller number of segmented regions. However, colour difference and the number of segmented regions are very different in physical meanings. The trade-off between them would be very difficult. Moreover, the preferred numbers of segmented regions could be very different from image to image. When this image-dependent measure is involved in a single evaluation function, it would be fairly difficult to perform segmentation evaluation without prior knowledge of image contents.

In summary, although several goodness functions have already been proposed, not many of them are directly based on human visual perception. Instead, most goodness methods combine several existing measures together to formulate their evaluation functions. The selection and the combination of different measures are usually subjective. The adjustment of weighting coefficients is often troublesome. Moreover, for most evaluation methods, the quality of segmentation is usually represented in one single function, which mixes together several weakly related measures. Without knowing the erroneous information about the segmented result under evaluation, these approaches could be very unreliable.

2.2 Visual rating experiments

In the proposed scheme for segmentation evaluation, we aim to develop a goodness approach that could mimic the way humans evaluate segmentation results. To explore the way a human perceives the performance of segmentation, a set of visual rating experiments were first made. In these experiments, to avoid evaluating directly the subjectively perceived quality of colour segmentation, two more objective quantities, the degree of 'visible' missing boundaries and the degree of 'visible' fake boundaries, are considered. Based on these experiment results, the correlations between the visual quality of colour segmentation and the degrees of missing-boundary and/or fake-boundary are then investigated.

In this section, three colour segmentation algorithms used in the visual rating experiments are to be introduced first. These algorithms will also be used in the following sections to demonstrate the performance of the proposed evaluation scheme. After a brief introduction of these three segmentation algorithms, the details of the visual rating experiments will be described.

2.2.1 Adopted colour segmentation algorithms: In general, current colour segmentation algorithms could be roughly classified into three major categories: (1) image domain-based approaches, (2) feature space-based approaches and (3) physics-based approaches [2]. For image domain-based approaches [1, 2, 14–17], most methods could be further classified into two groups: (1)

edge-based methods and (2) region-based methods. For edge-based methods, the discontinuity of local information is usually used as the gauge for segmentation, while for region-based methods, the similarity of neighbouring pixels is usually used. That is, edge-based methods mark boundaries with large enough intensity/colour variations, while region-based methods merge together pixels with small intensity/colour variations. For feature space-based approaches [1, 2, 18, 19], the data distribution of the entire image plays a crucial role. Clustering or grouping techniques are usually applied over the data distribution to allocate image data into groups. On the other hand, for the third type of approaches, the physics-based approaches, the adopted mathematical tools are basically the same as the former two types of approaches, while an underlying physical model is used to account for the reflection properties of the captured objects [2].

To select appropriate colour segmentation algorithms for the visual rating experiments, the segmentation algorithms are picked based on three criteria:

1. *Diversity:* these algorithms represent different types of image segmentation algorithms;
2. *Visibility:* all these algorithms had been presented to the vision community through a refereed publication and
3. *Availability:* the codes of these algorithms are readily available.

Based on these three criteria, we pick three different kinds of segmentation algorithms for our visual rating experiments. For edge-based approaches, we picked Ma and Manjunath's edge flow algorithm [14]; for region-based approaches, we picked Deng and Manjunath's JSEG algorithm [16] and for feature space-based approaches, we picked Comaniciu and Meer's mean shift algorithm [18]. As physics-based approaches are much less popular than the others, this type of approach is not considered in our experiments.

2.2.2 Visual rating experiments: To mimic the way humans evaluate segmentation results, we consider the degree of 'visible' missing boundaries and the degree of 'visible' fake boundaries. To explore how missing boundaries and fake boundaries affect the perceived quality of image segmentation, few visual rating experiments are made. In our experiments, 20 observers, 19 graduate students and 1 professor, with normal sight were involved. The ages of these observers were from 25 to 45 years. There was no special training for these observers before the experiments. To acquire more accurate experiment results with less sensitivity to context, the stimulus-comparison method was used [20]. In the stimulus-comparison method, any two of the subjects should be compared. Hence, this type of approach is usually time consuming. To avoid heavy time consumption but without sacrificing the diversity of colour images, six different colour images were used. These six images are shown in Figs. 2*a–f* and are named 'Fruit', 'Lena', 'House', 'Tower', 'Room' and 'Table tennis', respectively. On the other hand, as the attributes of segmentation results produced by different algorithms are quite different, it would be fairly difficult to compare segmentation results among different algorithms. For example, one algorithm may generate segmentation results with inaccurately located boundaries, while another algorithm may generate segmentation results with accurate boundaries but with some extra fake boundaries. Hence, in this article, we only focus on the comparison of segmentation results produced by the same algorithm but with different settings of control parameters.

Fig. 2 *Six test images*
a Fruit
b Lena
c House
d Tower
e Room
f Table tennis

In the experiments, the segmented results of 'Fruit' and 'Lena' are produced by the edge flow algorithm [14]. The 'Fruit' image is less textured, while the 'Lena' image includes slight texture over the hat fringe region. On the other hand, the segmented results of 'House' and 'Tower' are produced by the JSEG algorithm [16], while the segmented results of 'Room' and 'Table tennis' are produced by the mean-shift algorithm [18]. Table 1 shows the summary of these six colour images and the corresponding segmentation algorithms. Actually, the pairings of colour images and the used segmentation algorithms are just arbitrary. There is no special reason why the edge-flow algorithm is chosen for the 'Fruit' image, but not for the 'House' image.

For each image, nine segmentation results are produced by one of the three algorithms with nine different settings. Then, every two of these nine segmentation results are displayed in a random order on an LCD (liquid crystal display) monitor for comparisons, as shown in Fig. 3. That is, for each colour image, there are $C_2^9 = 36$ segmentation pairs to be compared. Totally, for all six colour images, there are $6 \times 36 = 216$ segmentation pairs to be compared. For each pair, the 20 observers are asked to subjectively compare the right segmentation result, named Seg I, with the left segmentation result, named Seg II, in terms of the perceived segmentation quality, the perceived degree of missing boundaries and the perceived degree of fake boundaries. Here, we use the seven-grade scales with a set of categories defined in semantic terms (e.g. much better, better, slightly better).

The experiment results for Figs. 2*a–f* are shown in Figs. 4*a* and *b*. In Fig. 4*a*, the nine triangles represent the averaged segmentation quality against the averaged degree of missing boundaries for the nine segmentation results of the 'Fruit' image. The term 'averaged' means that value is computed based on the grades from all 20 observers. Similarly, the asterisks, pentagrams, squares, circles and plus-signs represent the averaged segmentation quality against the averaged degree of missing boundaries for the segmentation results of 'Lena', 'House', 'Tower',

Table 1: Colour images against the applied segmentation algorithms

Segmentation algorithm	Colour image					
	Fruit	Lena	House	Tower	Room	Table tennis
Edge-flow algorithm [14]	◎	◎				
JSEG algorithm [16]			◎	◎		
Mean-shift algorithm [18]					◎	◎

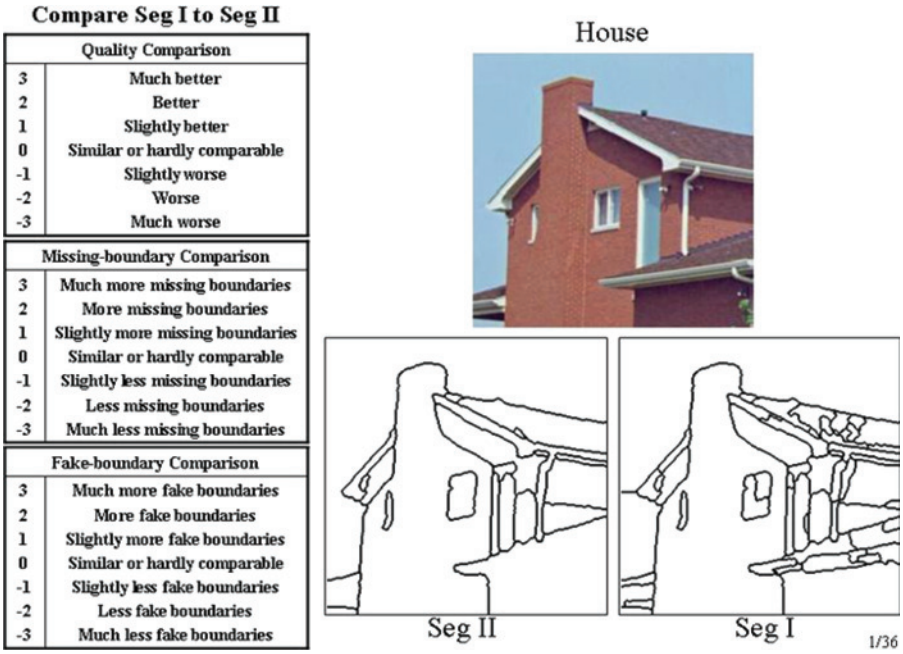


Fig. 3 *One of the 36 comparison patterns for the 'House' image*

'Room' and 'Table tennis', respectively. In Fig. 4*a*, it seems there is no apparent correlation between the perceived segmentation quality and the degree of missing boundaries. On the other hand, Fig. 4*b* shows the plot of the averaged segmentation quality against the averaged degree of fake boundaries for the segmentation results of the six colour images. Similarly, the correlation between the segmentation quality and the degree of fake boundaries is not very clear.

To measure the correlation between the averaged segmentation quality and the averaged degree of missing-boundary/fake-boundary, we calculate the correlation coefficient defined as

$$r = \frac{\sum(X - \bar{X})(Y - \bar{Y})}{\sqrt{\sum(X - \bar{X})^2}\sqrt{\sum(Y - \bar{Y})^2}} \tag{4}$$

where $\bar{X}$ is the mean of the scores on the X variable, while $\bar{Y}$ is the mean of the scores on the Y variable. In Table 2, we list the correlation coefficient representing the correlation between the segmentation quality and the degree of missing-boundary/fake-boundary. As listed in Table 2, we can see that the correlation between the segmentation

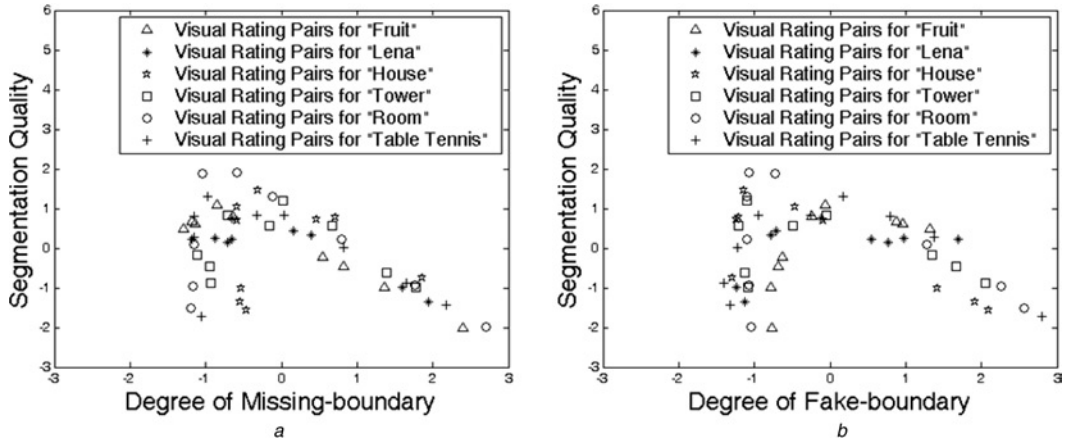


Fig. 4 *Results of the visual rating experiments*

a Segmentation quality against the degree of missing boundary for Figs. 2*a–f*
b Segmentation quality against the degree of fake boundary for Figs. 2*a–f*

Table 2: Averaged segmentation quality against missing boundary and/or fake boundary

Correlation	Colour image					
	Fruit	Lena	House	Tower	Room	Table tennis
Quality against missing-boundary	−0.954[b]	−0.840[b]	0.001	−0.206	−0.432	−0.496
Quality against fake-boundary	0.642	0.509	−0.788[a]	−0.396	−0.434	−0.114
Quality against missing boundary + fake boundary	−0.944[b]	−0.797[a]	−0.994[b]	−0.964[b]	−0.994[b]	−0.920[b]

[a]Correlation is significant at the 0.05 level (two-tailed) [21]
[b]Correlation is significant at the 0.01 level (two-tailed) [21]

quality and the degree of missing boundaries is not always significant at the 0.05 level for these six images. Neither is the correlation between the segmentation quality and the degree of fake boundaries. Here, the value of the significance level is defined as a value that is larger than or equal to a rejection probability under a two-class hypothesis. For example, with a 0.05 significant level, the probability is <0.05 that we would be wrong in rejecting the hypothesis that the correlation is zero. With such a low probability of error, we might confidently reject this hypothesis, and accept that there is a positive/negative correlation [22].

As the correlation between the averaged segmentation quality and the averaged degree of missing-boundary/fake-boundary is not always strong, we try to explore the correlation between the averaged segmentation quality and the combination of missing-boundary and fake-boundary. In Figs. 5*a–f*, the horizontal axis represents the degree of missing boundaries, increasing from left to right; while the vertical axis represents the degree of fake boundaries, increasing from bottom to top. Figs. 5*a–f* could be referred to the plots of visual false negatives against visual false positives for the segmentation results. These figures are closely related to the commonly used ROC (receiver operating characteristics) curves, which are plots of the true positive rates against false positive rates. Each of the nine segmentation results for Fig. 2*a* is represented by a square in Fig. 5*a*. The colour of the square denotes the normalised averaged grade of segmentation quality, increasing from dark red to white. It is not surprising that the best quality scores usually occur at the lower-left corner of the figure. That is, the preferred segmentation results are these results with both a lower degree of missing boundaries and a lower degree of fake boundaries. Similarly, the simulation results for Figs. 2*b–f* are shown in Figs. 5*b–f*, respectively. All these figures reveal the same phenomenon. To confirm this phenomenon, we plot the averaged segmentation quality against the averaged degree of missing boundaries, plus the averaged degree of fake boundaries for all six colour images, as shown in Fig. 5*g*. In Fig. 5*g*, we can easily see that the correlation between the visual quality of colour segmentation and the combination of these two visual errors is strong. To verify the strong correlation between the segmentation quality and the degree of missing boundaries plus fake boundaries, we also calculate the corresponding correlation coefficients. As listed in Table 2, it can be seen that the correlation coefficients between the segmentation quality and the degree of missing boundaries plus fake boundaries is significant at the 0.05 level, or even at the 0.01 level, for all six images. Moreover, as the sign of the correlation coefficient is negative, it implies that the preferred segmentation result is a segmentation result with a lower degree of missing boundaries plus a lower degree of fake boundaries. Hence, once if we can find some reasonable measures to estimate the degree of missing boundaries

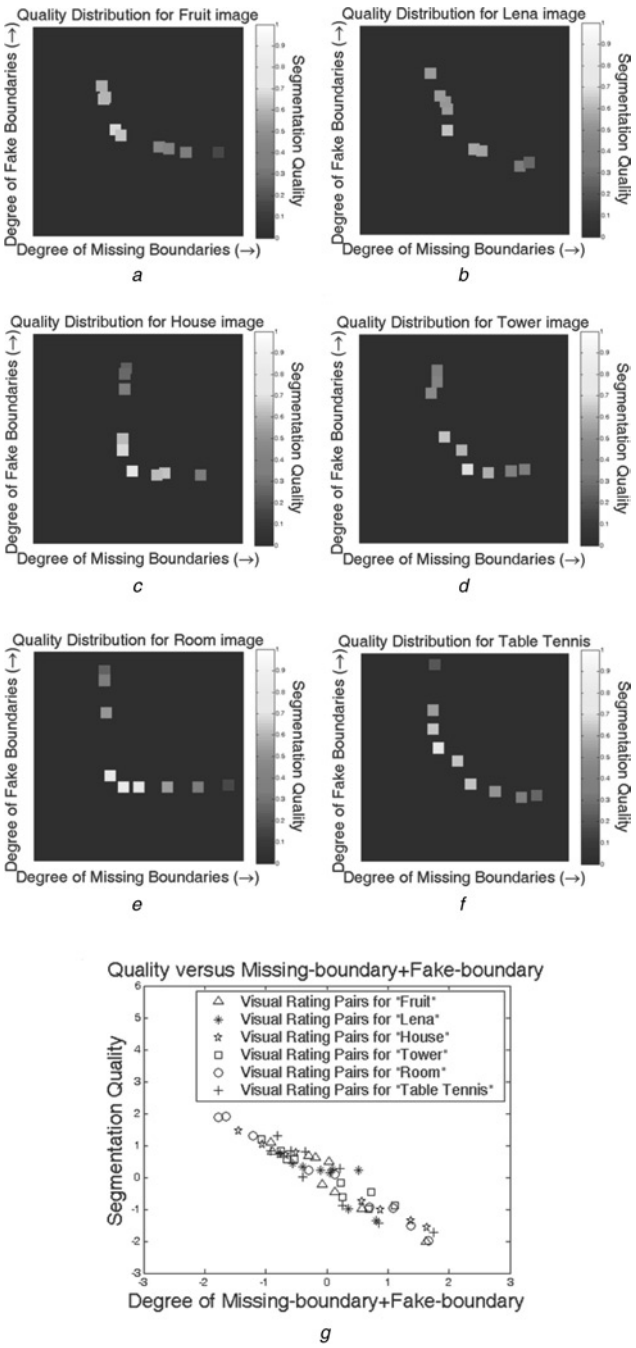


Fig. 5 *Segmentation quality with respect to the degree of missing boundary and the degree of fake boundary*

a Fruit
b Lena
c House
d Tower
e Room
f Table tennis
g Quality of segmentation against the degree of missing-boundary plus the degree of fake-boundary

and the degree of fake boundaries, we may use these measures to evaluate the segmentation quality in a reasonable and practical way.

3 Visible colour difference and error measures

3.1 Visible colour difference

In this article, we propose a new evaluation scheme that is based on the preference of having less visual errors in the segmented results. Based on the experiment results deduced in Section 2.2, it appears that the combination of the degree of missing boundaries and the degree of fake boundaries is closely related to humans' subjective evaluation over segmentation performance. Therefore if we can formulate some appropriate measures to estimate the degrees of missing boundaries and fake boundaries, we may find some quantitative and effective ways to evaluate segmentation algorithms.

To evaluate the quality of colour segmentation, we first propose the use of 'visible colour difference'. Among various definitions regarding colour difference [23], we choose the CIE ΔE^*_{Lab} definition as the basis of colour difference. As mentioned in the literature [23, 24], the value of ΔE^*_{Lab} is perceptually analogous to humans' visual perception of colour difference. This colour difference definition is defined over the CIE L*a*b* colour space, which is a roughly perceptually uniform colour space. The formula for converting an RGB image into the (L*, a*, b*) coordinates can be found in several colour-related articles [23, 24]. In this CIE L*a*b* colour space, the colour difference between two colours, (L_1^*, a_1^*, b_1^*) and (L_2^*, a_2^*, b_2^*), is defined as

$$\begin{aligned}\Delta E^*_{Lab} &\equiv \left\|(L_1^*, a_1^*, b_1^*) - (L_2^*, a_2^*, b_2^*)\right\|_{L^*a^*b^*} \\ &= \sqrt{(L_1^* - L_2^*)^2 + (a_1^* - a_2^*)^2 + (b_1^* - b_2^*)^2} \end{aligned} \quad (5)$$

Moreover, the values of ΔE^*_{Lab} could be roughly classified into three different levels to reflect three different degrees of colour difference perceived by humans [25]. As indicated in Table 3, the colour difference is hardly perceptible when ΔE^*_{Lab} is smaller than 3, is perceptible but still tolerable when ΔE^*_{Lab} is between 3 and 6, and is usually not acceptable when ΔE^*_{Lab} is larger than 6 [25]. Hence, in this article, we define a colour difference is 'visible' if its ΔE^*_{Lab} value is larger than 6.

3.2 Measures of visual errors

To estimate the degree of missing boundaries and fake boundaries, some appropriate quantitative goodness measures are formulated in this section. In general, for goodness methods, three basic types of measures are considered: (1) intra-region measure, (2) inter-region measure, and (3) region-shape measure [4]. Usually, intra-region measures are designed to measure the homogeneity within segmented regions, while inter-region measures are designed to measure the heterogeneity between adjacent regions. On the other hand, region-shape measures are usually designed to measure the regularity of region shape. Intuitively, the two former types of measures may be closely linked to the way humans evaluate the quality of segmentation at the discrimination level, while the third type of measures is more likely to be linked to the evaluation at the recognition level. Moreover, the intra-region measure that evaluates the homogeneity within segmented regions could be adopted to estimate the degree of missing boundaries, while the inter-region measure that evaluates the heterogeneity between adjacent regions could be used to estimate the degree of fake boundaries. Hence, in this article, we focus on the discussion of intra-region measure and inter-region measure.

Table 3: The effect of colour difference in the CIE L*a*b* colour space on human visual perception [25]

ΔE^*_{Lab}	Effect
<3	hardly perceptible
3 < 6	perceptible, but acceptable
>6	not acceptable

3.2.1 Intra-region visual error: To evaluate the degree of missing boundaries, a measure, named 'intra-region visual error', is designed. In each segmented region, these pixels with visible colour difference away from the average colour of that region are regarded as pixels with visible colour errors. Intuitively, a properly segmented region should contain as few visible colour errors as possible. Any missing boundary will cause the increase of intra-region visual errors. Given an $N \times M$ colour image $f(x, y)$, we first denote $\hat{f}(x, y)$ as the segmented colour image, with the colour of each segmented region being filled with the average colour of that region. We then define the intra-region error as

$$E_{intra}(I) = \frac{\sum_{x=1}^{N}\sum_{y=1}^{M} u\left(\left\|f(x,y) - \hat{f}(x,y)\right\|_{L^*a^*b^*} - th\right)}{N \times M} \quad (6)$$

where $\|\cdot\|_{L^*a^*b^*}$ denotes the colour difference in the CIE L*a*b* space, th denotes the threshold for visible colour difference and $u(\cdot)$ denotes the step function that is defined as

$$u(t) = \begin{cases} 1, & t > 0 \\ 0, & \text{otherwise} \end{cases} \quad (7)$$

In this article, we choose the threshold th to be 6, as explained in Section 3.1.

In (6), given a segmented result, we tend to calculate the total amount of the pixels with visible colour errors to estimate the degree of missing boundaries. Actually, when a segmented region contains more missing boundaries, the average colour of that region will have a larger colour difference from the original colours of those pixels. Once the colour difference is too large to be visible, the number of the pixels will be counted in (6). Hence, as the degree of missing boundaries increases, there will be more amounts of pixels with visible colour errors counted in (6).

In Table 4, for each of the six colour images, we confirm that the correlation between the intra-region visual error and the degree of missing boundaries is significant at the 0.01 level, and the sign of the correlation is positive. This implies that, given a segmentation result, the intra-region visual error could be an effective way to estimate the perceived degree of missing boundaries.

3.2.2 Inter-region visual error: On the other hand, the second measure, named 'inter-region visual error', is designed to evaluate the degree of fake boundaries. Given a colour segmentation result, we take into account these boundary pixels with invisible colour difference across the boundary. Intuitively, these pixels are not supposed to be

Table 4: Missing-boundary against intra-region visual error

Correlation	Colour image					
	Fruit	Lena	House	Tower	Room	Table tennis
Missing-boundary against intra-error	0.978[a]	0.980[a]	0.901[a]	0.980[a]	0.866[a]	0.918[a]

[a]Correlation is significant at the 0.01 level (two-tailed) [21]

Table 5: Fake-boundary against inter-region visual error

Correlation	Colour Image					
	Fruit	Lena	House	Tower	Room	Table tennis
Fake-boundary against inter-error	0.961[a]	0.964[a]	0.988[a]	0.991[a]	0.982[a]	0.853[a]

[a]Correlation is significant at the 0.01 level (two-tailed) [21]

detected as boundaries. Hence, as more fake boundaries appear in the segmented image, more inter-region visual errors are expected. In this article, we define the inter-region visual error of a segmented image as

$$\mathrm{E}_{\mathrm{inter}}(I) = \frac{\sum_{i=1}^{R}\sum_{\substack{j=1\\ j\neq i}}^{R} w_{ij} \times u\left(th - \left\|\hat{f}_i - \hat{f}_j\right\|_{\mathrm{L^*a^*b^*}}\right)}{N \times M} \quad (8)$$

where R denotes the number of segmented regions, and w_{ij} denotes the joined length between Region i and Region j. Here, w_{ij} is equal to zero if Region i and Region j are not connected.

Similarly, in (8), given a segmented result, we tend to calculate the total amounts of the boundary pixels with invisible colour errors to measure the degree of fake boundaries. Actually, when two adjacent regions contain an invisible boundary, the colour difference between the average colours of the adjacent regions will be small. Therefore once the average colour difference between any two adjacent regions is too small to be visible, the joined boundary pixels will be counted. Hence, if the degree of fake boundaries increases, more boundary pixels with invisible colour errors will be counted in (8).

In Table 5, for each of the six colour images, we confirm the correlation between the inter-region visual error and the degree of fake boundaries, obtained in the visual rating experiments, is also significant at the 0.01 level and the sign of the correlation is positive. This implies that the inter-region visual error could be an effective measure for the perceived degree of fake boundaries.

3.2.3 Intra-region visual error against inter-region visual error plot: With these two newly designed measures, we can estimate the degrees of missing boundaries and fake boundaries. Even though each of these two measures is still image dependent, the alliance of both measures may provide an effective way for segmentation evaluation.

In Fig. 6, we show the plot of intra-region visual error against inter-region visual error. As the given image is under-segmented, more boundaries are missing and the intra-region visual error increases. On the contrary, as the image is over-segmented, more fake boundaries appear and the inter-region error increases. These two measures are complementary to each other. With these two measures, the segmentation could be evaluated in quite a reasonable way. These two measures are also closely related in physical meaning. This makes the trade-off between these two measures much easier.

3.2.4 Ratio of intra-region visual error to inter-region visual error: As shown in Figs. 5*a–f*, the preferred segmented results are usually located at the lower-left corner in the plot of the quality segmentation against the sum of the degree of missing boundaries and the degree of fake boundaries. As the defined intra-region visual error is proportional to the degree of missing boundaries, and the inter-region visual error is proportional to the degree of missing boundaries, a preferred segmentation result is expected to locate at the lower-left corner of the inter-region error/intra-region error plot, as shown in Fig. 6. Analogous to the phenomenon that perceived segmentation quality is closely correlated with the sum of the degree of missing boundaries and the degree of fake boundaries, we assume the visual quality of a segmentation result can be evaluated based on a linear combination of intra-region visual error and inter-region visual error. That is, for I_j^i, the jth segmentation result of the ith image, we define its total visual error E_j^i as

$$E_j^i = \alpha_j^i \mathrm{E}_{\mathrm{intra}}(I_j^i) + \beta_j^i \mathrm{E}_{\mathrm{inter}}(I_j^i) \quad (9)$$

The total visual error E_j^i may also be normalised with respect to α_j^i and we have

$$\hat{E}_j^i = \mathrm{E}_{\mathrm{intra}}(I_j^i) + \lambda_j^i \mathrm{E}_{\mathrm{inter}}(I_j^i) \quad (10)$$

In (10), the coefficient λ_j^i is to balance the contributions of visual error from $\mathrm{E}_{\mathrm{intra}}(I_j^i)$ and $\mathrm{E}_{\mathrm{inter}}(I_j^i)$. For different image contents and different segmentation algorithms, λ_j^is are expected to be different. Based on the results of the visual experiments made in Section 2.2, we may perform linear fitting over perceived segmentation quality, measured intra-region errors and measured inter-region error to estimate the value of λ. Table 6 shows the estimated values of λ for these six colour images. In general, as the value of λ increases, the preferred segmented results are the

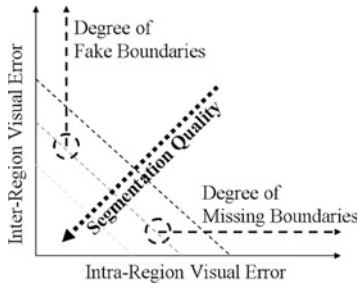


Fig. 6 *The intra-region visual error against inter-region visual error plot*

Table 6: Values of λ

	Colour image					
	Fruit	Lena	House	Tower	Room	Table tennis
λ_i	0.586	1.36	9.05	7.70	3.18	5.34
Segmentation algorithm	edge flow	edge flow	JSEG	JSEG	mean shift	mean shift

results with more intra-region visual errors; while as the value of λ decreases, the preferred segmented results are the results with more inter-region visual errors. We can easily see that for different images and different algorithms the values of λ are quite different. However, for each segmentation algorithm, the values of λ are roughly of the same order of magnitude.

To further investigate the impact of coefficient λ over the performance of the proposed evaluation scheme, we discuss the correlation of the averaged segmentation quality and the total visual error with respect to different λs. As shown in Fig. 7, the nine triangles denote the minus values of the correlation coefficients for the segmented results of 'Fruit' with $\lambda = 1, 2, \ldots, 9$, respectively. Similarly, the asterisks, pentagrams, squares, circles and plus-signs denote the minus correlation coefficients for the segmented results of 'Lena', 'House', 'Tower', 'Room' and 'Table tennis'. It can be seen that the λs in Table 6 correspond to the λ that causes the maximum correlation. For example, for the case of the 'Room' image, the minus value of the correlation coefficient reaches its local maximum around $\lambda = 3$. On the other hand, for most images, the value of the correlation coefficient remains large even if λ is changed. The averaged values of the minus correlation coefficients are represented in the black solid line. It can be seen that, with λ ranging from 2 to 7, the correlation of the averaged segmentation quality and the total visual error remains significant at the 0.05 level. Hence, in this article, we use the averaged value of λs in Table 6, as calculated in (11), to be a typical choice of λ

$$\lambda \equiv \frac{1}{6}\sum_{i=1}^{6}\lambda_i = 4.54 \tag{11}$$

Of course, this typical choice of λ is only a rough estimate and may not work for all types of images. How to automatically choose an appropriate value of λ for a given image deserves further investigation in the future.

4 Evaluation of segmentation with the inter-region error/intra-region error plot

In this section, the use of the inter-region error/intra-region error plot in the evaluation of colour segmentation is introduced. Also, the automatic selection of parameter settings for a given segmentation algorithm is described.

Fig. 8*a* shows the 'Fruit' image. Figs. 8*b–f* show several segmentation results of Fig. 8*a* produced by the edge-flow algorithm [14], with different parameter settings. Subjectively, Fig. 8*c* is preferable. In comparison with Fig. 8*c*, Fig. 8*b* has a higher degree of fake boundaries, while Figs. 8*d–f* have higher degrees of missing boundaries. As shown from left to right in Fig. 8*s*, the five blue circles represent the 'intra-region visual error' against 'inter-region visual error' pairs of Figs. 8*b–f*, respectively.

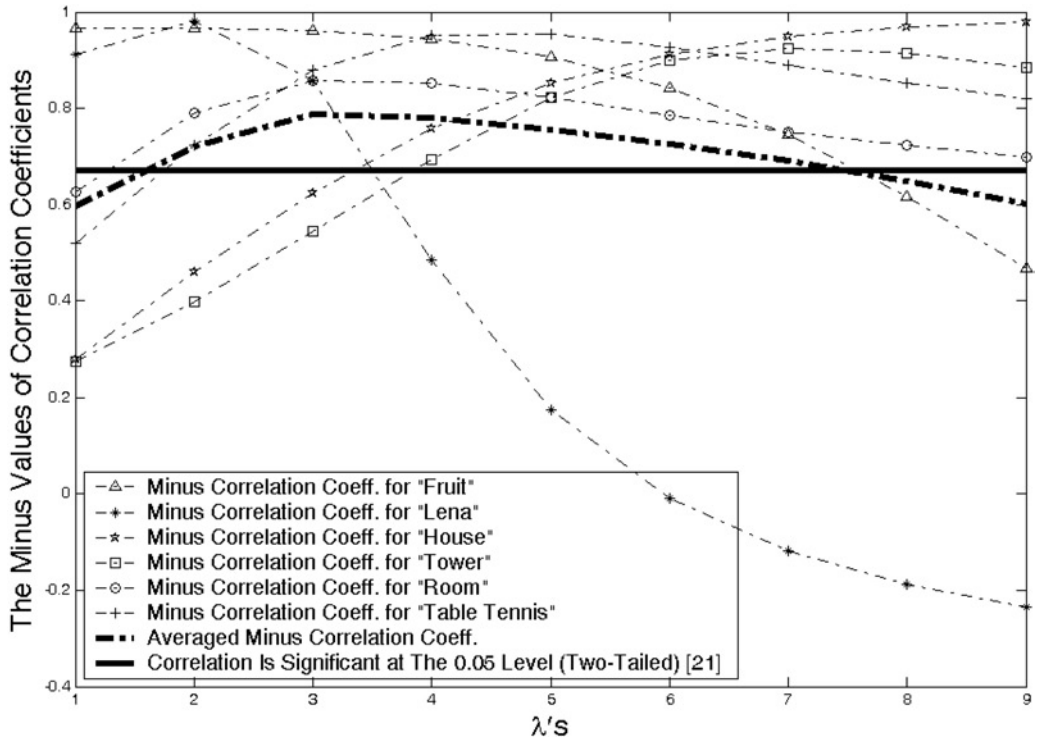


Fig. 7 *Correlation plot of averaged segmentation quality and total visual error with respect to different λs*

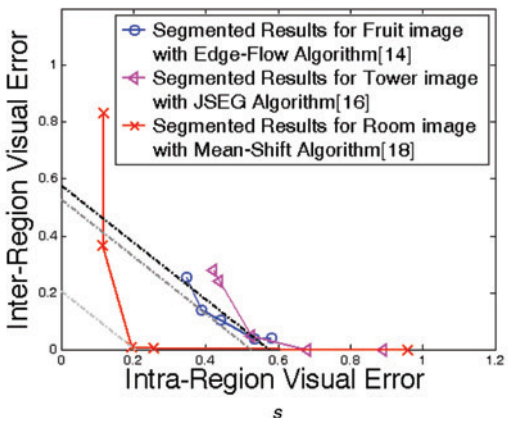
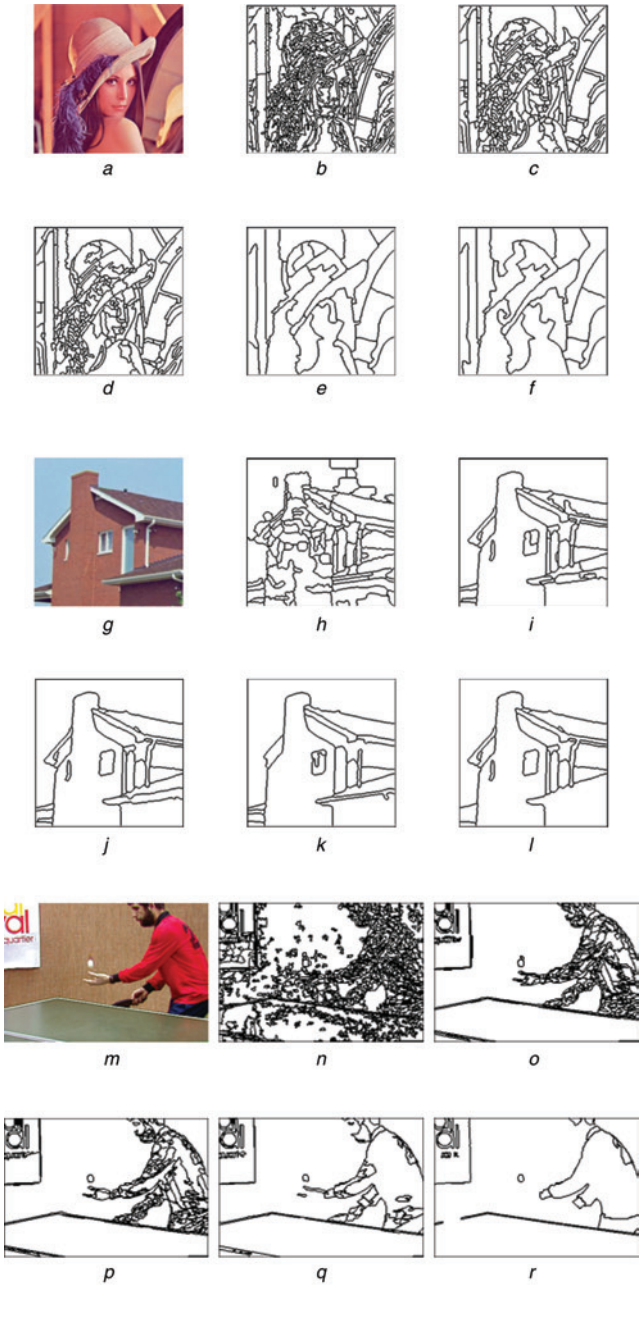


Fig. 8 *Evaluation of segmentation results*

a Fruit image
b–f Segmented results of *a* by using the edge-flow algorithm [14]
g Tower image
h–l Segmented results of *g* by using the JSEG algorithm [16]
m Room image
n–r Segmented results of *m* by using the mean-shift algorithm [18]
s Inter-region error against intra-region error plot of *b–f*, *h–l* and *n–r*



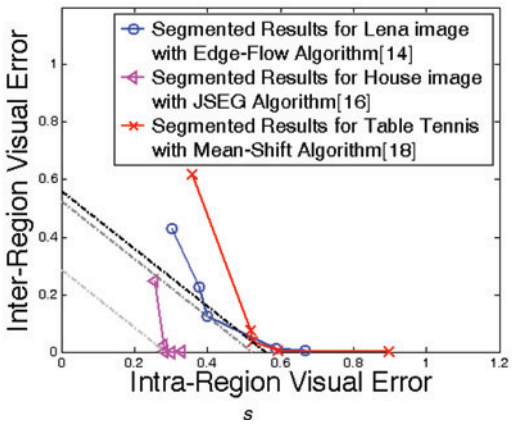


Fig. 9 *Evaluation of segmented results*

a Lena image
b–f Segmented results of *a* by using the edge-flow algorithm [14]
g House image
h–l Segmented results of *g* by using the JSEG algorithm [16]
m Table tennis image
n–r Segmented results of *m* by using the mean-shift algorithm [18]
s Intra-region error against inter-region error plot of *b–f*, *h–l* and *n–r*

Table 7: Evaluation comparison for the 'Fruit' image

Evaluation	Segmented result				
	Fig. 8*b*	Fig. 8*c*	Fig. 8*d*	Fig. 8*e*	Fig. 8*f*
Averaged visual quality	0.488 (3)	1.100 (1)	0.806 (2)	−0.225 (4)	−0.469 (5)
$E_{intra}(I) + E_{inter}(I)\|_{\lambda=4.54}$	0.600 (4)	0.528 (1)	0.542 (2)	0.572 (3)	0.625 (5)
$F(I)$	0.753 (5)	0.730 (4)	0.440 (3)	0.379 (2)	0.285 (1)
$F'(I)$	0.075 (5)	0.073 (4)	0.044 (3)	0.038 (2)	0.029 (1)
$Q(I)$	0.201 (5)	0.183 (4)	0.154 (3)	0.143 (2)	0.128 (1)

It can be seen that, with similar intra-region errors, Fig. 8*b* has larger inter-region error values than those of Fig. 8*c*. On the other hand, with similar inter-region errors, Figs. 8*d–f* have larger intra-region error values than those of Fig. 8*c*. Hence, in the selection of parameter setting, the sum of $E_{intra}(I)$ and $E_{inter}(I)|_{\lambda=4.54}$ may serve as a suitable criterion for the evaluation of segmentation performance. As the sum reaches a smaller value, the parameter setting is expected to achieve better segmentation. In Fig. 8*s*, we use the quality straight line, $E_{intra}(I) + E_{inter}(I)|_{\lambda=4.54}$ to illustrate this idea. Here we use grey straight quality lines to denote the lines $E_{intra}(I) + E_{inter}(I)|_{\lambda=4.54} =$ constant. It can be easily seen that Fig. 8*c* does have the smallest sum if compared with the other four.

In Figs. 8*g* and *m*, we show another two examples of colour images. Figs. 8*h–l* show the segmentation results of Fig. 8*g* produced by the JSEG algorithm [16], while Figs. 8*n–r* show the segmentation results of Fig. 8*m* produced by the mean-shift algorithm [18], all with different parameter settings. Similarly, in Fig. 8*s*, from left to right, the 'intra-region error' against 'inter-region error' pairs of Figs. 8*h–l* are represented by triangles; while the error pairs of Figs. 8*n–r* are represented by crosses. It can be easily seen that Fig. 8*j* has the smallest error sum if compared with Figs. 8*h–l*; while Fig. 8*p* has the smallest error sum if compared with Figs. 8*n–r*. In perception, the segmented results in Figs. 8*j* and *p* do appear to be the most preferable results among these candidates.

Similarly, in Figs. 9*a*, *g* and *m*, we show the other three of colour images. Figs. 9*b–f* show the segmentation results of Fig. 9*a* produced by the edge-flow algorithm [14], while Figs. 9*h–l* show the segmentation results of Fig. 9*g* produced by the JSEG algorithm [16] and Figs. 9*n–r* show the segmentation results of Fig. 9*m* produced by the mean-shift algorithm [18], all with different parameter settings. Similarly, in Fig. 9*s*, from left to right, the 'intra-region error' against 'inter-region error' pairs of Figs. 9*b–f* are represented by circles, while the error pairs of Figs. 9*h–l* are represented by triangles and the error pairs of Figs. 9*n–r* are represented by crosses. It can be easily seen that Fig. 9*d* has the smallest error sum if compared with Figs. 9*b–f*, Fig. 9*j* has the smallest error sum if compared with Figs. 9*h–l* and Fig. 9*p* has the smallest error sum if compared with Figs. 9*n–r*. In perception, the segmented results in Figs. 9*d*, *j* and *p* do appear to be the most preferable results among these candidates.

In summary, we use these three simulation results to demonstrate how the inter-region error/intra-region error plot can be used to automatically select the parameter setting based on the performance of segmentation results. In fact, $E_{intra}(I)$ and $E_{inter}(I)$ can be combined in various forms based on user's requirements. So far, we found that the simple form $E_{intra}(I) + E_{inter}(I)$ performs pretty well when applied to various types of colour images.

In Table 7, we compare the proposed evaluation measure, $E_{intra}(I) + E_{inter}(I)|_{\lambda=4.54}$, with three existing evaluation measures in the literature. These three measures are the $F(I)$ measure [10], the $F'(I)$ measure [11] and the $Q(I)$ measure [11]. In Table 7, the five segmentation results shown in Figs. 8*b–f* are used as the test inputs for the comparison. Here, the 'averaged visual quality' denotes the subjective evaluation results based on the visual experiment mentioned in Section 2.2, with a larger number indicating a better rating of perceived quality. The other four rows indicate the evaluation scores based on the proposed measure, $E_{intra}(I) + E_{inter}(I)|_{\lambda=4.54}$, the $F(I)$ measure the $F'(I)$ measure, and the $Q(I)$ measure, respectively. For these four evaluation measures, a smaller number indicates a better rating of the measurement. Moreover, the numbers in parentheses indicates the ranking of these five test inputs based on the applied evaluation measure. According to the subjective visual experiment results, Fig. 8*c* is ranked as the best segmentation results among Figs. 8*b–f*. As shown in Table 7, the proposed evaluation measure does pick Fig. 8*c* as the best segmentation with the smallest visual errors, while all the other three evaluation measures pick Fig. 8*f* as the best segmentation result. Similarly, in Tables 8–12, we show the comparisons over the other five colour images. All these tables illustrate that the proposed evaluation measure $E_{intra}(I) + E_{inter}(I)$ does provide a reasonable and reliable way for the evaluation of colour segmentation.

Table 8: Evaluation comparison for the 'Tower' image

Evaluation	Segmented result				
	Fig. 8*h*	Fig. 8*i*	Fig. 8*j*	Fig. 8*k*	Fig. 8*l*
Averaged visual quality	−0.469 (4)	−0.181 (3)	0.569 (1)	0.556 (2)	−0.625 (5)
$E_{intra}(I) + E_{inter}(I)\|_{\lambda=4.54}$	0.702 (4)	0.677 (2)	0.576 (1)	0.681 (3)	0.893 (5)
$F(I)$	1.090 (5)	0.987 (4)	0.194 (3)	0.091 (1)	0.184 (2)
$F'(I)$	0.109 (5)	0.099 (4)	0.019 (3)	0.009 (1)	0.018 (2)
$Q(I)$	0.354 (5)	0.330 (4)	0.153 (2)	0.092 (1)	0.180 (3)

Table 9: Evaluation comparison for the 'Room' image

Evaluation	Segmented result				
	Fig. 8*n*	Fig. 8*o*	Fig. 8*p*	Fig. 8*q*	Fig. 8*r*
Averaged visual quality	−0.963 (4)	0.100 (3)	1.869 (1)	1.306 (2)	−1.988 (5)
$E_{intra}(I) + E_{inter}(I)\|_{\lambda=4.54}$	0.950 (4)	0.479 (3)	0.205 (1)	0.262 (2)	0.959 (5)
$F(I)$	1.021 (5)	0.540 (4)	0.389 (2)	0.392 (3)	0.076 (1)
$F'(I)$	0.102 (5)	0.054 (4)	0.039 (2)	0.039 (2)	0.008 (1)
$Q(I)$	0.155 (5)	0.094 (2)	0.085 (1)	0.107 (3)	0.135 (4)

Table 10: Evaluation comparison for the 'Lena' image

Evaluation	Segmented result				
	Fig. 9*b*	Fig. 9*c*	Fig. 9*d*	Fig. 9*e*	Fig. 9*f*
Averaged visual quality	0.219 (3)	0.231 (2)	0.765 (1)	−0.999 (4)	−1.363 (5)
$E_{intra}(I) + E_{inter}(I)\|_{\lambda=4.54}$	0.733 (5)	0.606 (3)	0.525 (1)	0.602 (2)	0.673 (4)
$F(I)$	0.902 (5)	0.769 (4)	0.614 (3)	0.224 (2)	0.173 (1)
$F'(I)$	0.090 (5)	0.077 (4)	0.061 (3)	0.022 (2)	0.017 (1)
$Q(I)$	0.175 (5)	0.173 (4)	0.149 (3)	0.127 (2)	0.113 (1)

Table 11: Evaluation comparison for the 'House' image

Evaluation	Segmented result				
	Fig. 9*h*	Fig. 9*i*	Fig. 9*j*	Fig. 9*k*	Fig. 9*l*
Averaged visual quality	−1.363 (5)	1.050 (2)	1.469 (1)	0.725 (4)	0.769 (3)
$E_{intra}(I) + E_{inter}(I)\|_{\lambda=4.54}$	0.502 (5)	0.303 (2)	0.286 (1)	0.303 (2)	0.333 (4)
$F(I)$	0.473 (5)	0.194 (4)	0.125 (3)	0.100 (1)	0.116 (2)
$F'(I)$	0.047 (5)	0.019 (4)	0.013 (3)	0.001 (1)	0.012 (2)
$Q(I)$	0.148 (5)	0.072 (4)	0.059 (2)	0.050 (1)	0.064 (3)

Table 12: Evaluation comparison for the 'Table tennis' image

Evaluation	Segmented result				
	Fig. 9*n*	Fig. 9*o*	Fig. 9*p*	Fig. 9*q*	Fig. 9*r*
Averaged visual quality	−1.713 (5)	0.794 (3)	1.306 (1)	0.831 (2)	−0.894 (4)
$E_{intra}(I) + E_{inter}(I)\|_{\lambda=4.54}$	0.977 (5)	0.595 (2)	0.560 (1)	0.598 (3)	0.897 (4)
$F(I)$	6.320 (5)	2.195 (4)	1.738 (3)	0.704 (2)	0.218 (1)
$F'(I)$	0.632 (5)	0.220 (4)	0.174 (3)	0.070 (2)	0.022 (1)
$Q(I)$	1.124 (5)	0.576 (4)	0.462 (3)	0.268 (2)	0.184 (1)

5 Conclusions

In this article, we describe a new evaluation scheme based on the visible colour difference for colour segmentation. To avoid directly evaluating the subjective quality of colour segmentation, we estimate the degrees of missing boundaries and fake boundaries first. With the combination of these two quantities, we could approach the subjective evaluation of colour segmentation. Also, based the definition of visible colour difference, we design two measures, the intra-region visual error and inter-region visual error, to estimate the degrees of missing boundaries and fake boundaries, respectively. We found these measures, based on these two types of visible colour differences, have significant correlation with the degree of missing boundaries and the degree of fake boundaries. With these two measures, an evaluation scheme is proposed to evaluate the segmentation results and help the automatic selection of the parameters for a given segmentation algorithm. The simulation results have demonstrated the potential of this approach in providing reliable and efficient evaluations over colour segmentation. Moreover, given that the measures of segmentation quality presented here are designed to fit for subjective evaluations of segmentation quality, these measures are particularly applicable to tasks such as content-based image retrieval.

6 References

1 Cheng, H.D., Jiang, X.H., Sun, Y., and Wang, J.: 'Color image segmentation: advances and prospects', *Pattern Recognit.*, 2001, **34**, (6), pp. 2259–2281

2 Lucchese, L., and Mitra, S.K.: 'Color image segmentation: a state-of-the-art survey'. Proc. Indian National Science

Academy (INSA-A), New Delhi, India, 2001, vol. 67, A, (2), pp. 207–221
3 Heath, M.D., Sarkar, S., Sanocki, T., and Bowyer, K.W.: 'A robust visual method for assessing the relative performance of edge-detection algorithms', *IEEE Trans. Pattern Anal. Mach. Intell.*, 1997, **19**, (12), pp. 1338–1359
4 Hoover, A., Jean-Baptiste, G., Jiang, X., Flynn, P.J., Bunke, H., Goldgof, D.B., Bowyer, K., Eggert, D.W., Fitzgibbon, A., and Fisher, R.B.: 'An experimental comparison of range image segmentation algorithms', *IEEE Trans. Pattern Anal. Mach. Intell.*, 1996, **18**, (7), pp. 673–689
5 Correia, P.L., and Pereira, F.: 'Objective evaluation of video segmentation quality', *IEEE Trans. Image Process.*, 2003, **12**, (2), pp. 186–200
6 Zhang, Y.J.: 'A survey on evaluation methods for image segmentation', *Pattern Recognit.*, 1996, **29**, (8), pp. 1335–1346
7 Zhang, Y.J.: 'A review of recent evaluation methods for image segmentation'. Proc. 6th Int. Symp. on Signal Processing and its Applications, Kuala Lumpur, Malaysia, 2001, pp. 148–151
8 Martin, D.D., Fowlkes, C.C., Tal, D., and Malik, J.: 'A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics'. Proc. IEEE Int. Conf. on Computer Vision, Vancouver, Canada, 2001, vol. 2, pp. 416–423
9 Martin, D.R., Fowlkes, C.C., and Malik, J.: 'Learning to detect natural image boundaries using local brightness, colour, and texture cues', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2004, **26**, (5), pp. 530–549
10 Liu, J., and Yang, Y.H.: 'Multiresolution colour image segmentation', *IEEE Trans. Pattern Anal. Mach. Intell.*, 1994, **16**, (7), pp. 689–700
11 Borsotti, M., Campadelli, P., and Schettini, R.: 'Quantitative evaluation of colour image segmentation results', *Pattern Recognit. Lett.*, 1998, **19**, (8), pp. 741–747
12 Rosenberger, C., and Chehdi, K.: 'Genetic fusion: application to multi-components image segmentation'. Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, Istanbul, Turkey, 2000, vol. 4, pp. 2223–2226
13 Huang, Q., and Dom, B.: 'Quantitative methods of evaluating image segmentation'. Proc. IEEE Int. Conf. on Image Processing, Washington, DC, USA, 1995, vol. 3, pp. 53–56
14 Ma, W.Y., and Manjunath, B.S.: 'Edge flow: a technique for boundary detection and image segmentation', *IEEE Trans. Image Process.*, 2000, **9**, (8), pp. 1375–1388
15 Haris, K., Efstratiadis, S.N., Maglaveras, N., and Katsaggelos, A.K.: 'Hybrid image segmentation using watersheds and fast region merging', *IEEE Trans. Image Process.*, 1998, **7**, (12), pp. 1684–1699
16 Deng, Y., and Manjunath, B.S.: 'Unsupervised segmentation of colour-texture regions in images and video', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2001, **23**, (8), pp. 800–810
17 Shi, J., and Malik, J.: 'Normalized cuts and image segmentation'. Proc. IEEE Conf. on Computer Vision Pattern Recognition, 1997, pp. 731–737
18 Comaniciu, D., and Meer, P.: 'Mean shift: a robust approach toward feature space analysis', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2002, **24**, (5), pp. 603–619
19 Cheng, H.D., and Sun, Y.: 'A hierarchical approach to colour image segmentation using homogeneity', *IEEE Trans. Image Process.*, 2000, **9**, (12), pp. 2071–2082
20 ITU-R Recommendation BT. 500-11: 'Methodology for the subjective assessment of the quality of television pictures', Geneva, 2002. Available at http://www.itu.org
21 Ronald, R.A., and Yates, F.: 'Statistical methods for research workers' (Oliver and Boyd Ltd, Edinburgh, 1970)
22 Siegel, S.: 'Nonparametric statistics for the behavioral sciences' (McGraw-Hill Kogakusha Ltd, Tokyo, 1956)
23 Sharma, G., and Trussell, H.J.: 'Digital colour image', *IEEE Trans. Image Process.*, 1997, **6**, (7), pp. 901–932
24 Wyszecki, G., and Stiles, W.: 'Color science: concepts and methods, quantitative data and formulae' (John Wiley & Sons, New York, 1982, 2nd edn.)
25 Hardeberg, J.Y.: 'Acquisition and reproduction of colour images: colourimetric and multispectral approaches'. PhD thesis, Ecole Nationale Supérieure des Télécommunications', Paris, France, 1999