Mutual Information Criterion-Based Optimal Scale Selection for Image Denoising and Segmentation

Tiexiang Wen[1,2,3]*, Qingsong Zhu* [1,3,*] *and Jia Gu*[1,3,*]

[1] Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, 518055, P. R. China
[2] University of Chinese Academy of Sciences, Beijing, 100049, P. R. China
[3] The Shenzhen Key Laboratory for Low-cost Healthcare, Shenzhen, 518055, P. R. China

Abstract: Scale-spaces play an important role in many computer vision tasks. Automatic scale selection is at the foundation of multi-scale image analysis, but its performance is still very subjective and empirical. To automatically select the appropriate scale for a particular issue, a scale selection model based on information theory is proposed in this paper. The proposed model utilizes mutual information as a measuring criterion of similarity for the optimal scale selection in multi-scale analysis, with applications to image denoising and segmentation. Firstly, we focus on the morphological operator based scale selection to image denoising. This technique does not require the prior knowledge of the noise variance and can effectively eliminate the variation of illumination. Secondly, we develop a clustering based unsupervised image segmentation algorithm by recursively pruning the Huffman coding tree. The proposed clustering algorithm can preserve the maximum amount of information at a specific clustering number from the information-theoretical point of view. Finally, for the feasibility of the proposed algorithms, its theoretical properties are analyzed mathematically and its performance is tested by a series of experiments, which demonstrate that it yields the optimal scale for our developed image denoising and segmentation algorithms.

Keywords: scale selection, mutual information, denoising, segmentation

1 Introduction

Scale is of an important concept in the mathematical modeling process for a variety of modeling fields. But the choice of scale is quite complex. In a wide range of chemical, physical, and biological systems, macroscopic, coherent behavior emerges from interactions between microscopic entities (molecules, cells, individuals in a population) among themselves and with their environment. In many cases, a macroscopic model (such as the Navier-Stokes equations for fluid flow or a reaction-diffusion equation) has been formally derived that quantitatively describes behavior at this level to perform a variety of tasks (e.g. simulation, optimization, bifurcation analysis) using analytical and numerical techniques [1–3]. For many complex systems, however, although evolution is observed at a macroscopic scale of interest, accurate models are only given at a more detailed (fine-scale, microscopic) level of description (e.g., lattice Boltzmann, kinetic theory of active particles, molecular

dynamics, cellular automata) [4–6]. Recently, effects have been made to build the connection between macroscopic and microscopic level in physicochemical modeling [7–9].

It is vast for the application of scale-space theory and its relationship among scales is complex. This study focuses only on the optimal scale selection for the image processing task. Scale analysis is one of the most useful frameworks for many image processing tasks [10–12]. The fundamental ideal behind hierarchical multi-scale representation is to analyze with respect to scales [13, 14] and it is commonly defined as the collection of the filtered image.

In some classical literatures, multi-scale representation could be distinguished as three types. The first one is the wavelet transform [15], which is widespread used in the field of image processing, such as image denosing [16], image and video compression [17, 18], and multi-modality registration [19]. By employing the windowing technique

---

* Corresponding author e-mail: qs.zhu@siat.ac.cn, jia.gu@siat.ac.cn

with variable-size region, wavelet transform is capable of decomposing a signal into multi-scale representation. The second type is derived from the diffusion process and based on partial differential equations [20]. It originates from the conventional isotropic heat flow equation and has been widespread used in image filtering [21–24], segmentation [25–27] and inpainting [28–30]. And the third type is based on the mathematical morphology [31–33]. Most of the morphological filters generate nonlinear scales, which have exhibited excellence performance in most image processing tasks comparing with the linear scale analysis. For example, the features (e.g. edges, corners) of an image could be well preserved by the nonlinear operator.

Scale is the key parameter in the application of multi-scale analysis and its performance critically determines the success of a multi-scale algorithm. Several existing techniques for selecting the proper scale in image denoising have been proposed recently by using optimal control theory [34, 35], asymptotic perturbation methods [36], minimization of the correlation of signal and noise [37], statistical model [38, 39], and Markov random field model [40]. However, all these methods require some prior of noise variance and its evaluation is time-consuming and sometimes impractical. To improve the robustness of scale selection, we extend the information-theoretic model [41] and propose a new scale selection algorithm, which is based on the maximization of mutual information for morphological multi-scale representation.

Further, our proposed scale selection model is used for unsupervised image segmentation, where the determination of number of cluster (NC) is treated as the selection of optimal scale. And the measuring criterion is from the computation of mutual information. Such criterion has been adopted towards image segmentation along with binary space partition method [42], simulated annealing method [43] and fuzzy C-means method (FCM) [44, 45]. Differ from those methods, our proposed algorithm could maintain the maximum amount of information between the original image and segmented images, and the convexity of solution could be guaranteed by employing the Huffman coding strategy. And all these well-pose properties are proved mathematically.

The remainder of this paper is organized as follows. Section 2 describes the proposed scale selection model and its application to image denoising and segmentation based on the objective functional of mutual information maximization. Experimental results are presented in section 3 and conclusions are drawn in section 4.

2 The proposed model

2.1 The model description

In the information theory, the mutual information is defined as a measuring metric of statistical correlation for two random variables X and Y. We model an image $u(x)$ as a random variable. Thus, the probability density of $u(x)$ is denoted by $p_u(i)$ and is usually estimated from image histogram bins. Given two images $u(x)$ and $v(x)$, and then the mutual information $I(u,v)$ is written as follows:

$$I(u,v) = H(u) + H(v) - H(u,v), \qquad (1)$$

$$H(u) = -\sum_{i=1}^{n} p_u(i) log(p_u(i)), \qquad (2)$$

$$H(u,v) = -\sum_{i=1}^{n} \sum_{j=1}^{m} p_{u,v}(i,j) log(p_{u,v}(i,j)), \qquad (3)$$

where $H(u)$ denotes the Shannon entropy. It expresses the average information or uncertainty of a random variable and obtains its maximum entropy only in the case of equally occurrence of the probability bins. $H(u,v)$ is the joint entropy of $u(x)$ and $v(x)$ and expresses the misalignment between them. $p_{u,v}(i,j)$ is the joint probability and measures the similarity between $u(x)$ and $v(x)$. The joint probability is defined by:

$$p_{u,v}(i,j) = \left| \frac{x_k | u(x_k) = i \bigcap v(x_k) = j}{\Omega_{uv}} \right|. \qquad (4)$$

The mutual information maximizes the amount of information if random variables $u(x)$ and $v(x)$ are correctly aligned. To express the relationship between the original image $u(x)$ and the processed image with multi-scale algorithm, we define an information-theoretic functional J_t as follows:

$$J_t = I(u(x), P_t(u(x))), \qquad (5)$$

where P_t is an operational operator parameterized by scale t, for instance, the morphological open operator for image denoising and clustering operator for image segmentation.

We seek to estimate the scale t that parameterizes the multi-scale operator by maximizing its mutual information over the scale t by:

$$t = arg \max_t J_t. \qquad (6)$$

We denote the mutual information between the original image and the processed image as follows:

$$I_1 = I(u, P_t u). \qquad (7)$$

And the computation for the mutual information between the original image and the residual of the processed image is formulated as:

$$I_2 = I(u, u - P_t u). \qquad (8)$$

With the increase of the scale variable t, it is observed that the growth for the amount of mutual information between (7) and (8) is opposite. One is increasing while

the other is decreasing. For the convenience of comparison between (7) and (8), we normalize and rewrite them as follows:

$$I_1 = \frac{I(u, P_t u)}{I(u, u)}, \qquad (9)$$

and

$$I_2 = \frac{I(u, u - P_t u)}{I(u, u)}. \qquad (10)$$

In the following sections, we apply the above normalized equations as a determining criterion for the problem of scale selection in image denoising and segmentation, respectively.

2.2 Maximization of mutual information for denoising

2.2.1 Description of the denoising algorithm

In this section, we focus on the denoising algorithm in the multi-scale representation of morphological open operator. Accordingly, we replace the function operator in (9) and (10) with the specific open operator O_t. Open operation is a combination of dilation and erosion operators, which are fundamental operators in mathematical morphology. And the selection of the shape and size of the structuring element (SE) is very fundamental and crucial. Unfortunately, the selection of SE is subjective and empirical to a specific image processing task. To automate the selection of SE, we parameterize the size of SE with scale t and re-formulate the fundamental morphological operator from the multi-scale point of view. Then the problem of determining the size of the SE becomes the problem of selecting the proper or optimal scale for a specific image processing task.

The dilation operator D that computes the maximized gray value over a given size of SE B with scale t is described as:

$$D_t u(x) = \sup_{y \in x + tB} u(y). \qquad (11)$$

When an image is dilated by the dilation operator, finer details (or noise) whose size is smaller than the SE B will be eliminated from the image. By increasing the scale t from 0 to larger scale, a stack of filtered images is obtained.

Conversely, the erosion operator E computes the minimized gray value near the neighborhood and is defined as follows:

$$E_t u(x) = \inf_{y \in x + tB} u(y). \qquad (12)$$

Then the open operator O is defined as the dilation operator applied after the erosion:

$$O_t u(x) = \sup_{B' \in tB} \inf_{y \in x + B'} u(y). \qquad (13)$$

Using Taylor expansion to (13), we have:

$$O_t u(x) = \sup_{y \in tB} \inf u(y) = u(x) - t \cdot \bigtriangledown u(x) + o(t), \quad (14)$$

where $o(t)$ is an infinitesimal of higher order to scale t, and $\bigtriangledown u(x)$ is the derivative with respect to original image $u(x)$. Thus, we obtain the limit of the open operator as scale t vanishes:

$$\lim_{t \to 0} O_t u(x) = u(x). \qquad (15)$$

Meanwhile, we obtain easily the limitation of the open operator as scale t tends to infinite:

$$\lim_{t \to \infty} O_t u(x) = \lim_{t \to \infty} \sup_{B' \in tB} \inf_{y \in x + B'} u(y) = C, \qquad (16)$$

where C is a constant. Equation (16) indicates that if the selected scale is sufficiently large, the filtered image by open operator is a flat (or constant) image.

The residual of the opening compared to the original image represents the top-hat transformation:

$$T_t u(x) = u - O_t u. \qquad (17)$$

The open operator suppresses fine details in the images successively. Thus, when the opened image is subtracted from the original, the desired detail is obtained [46]. One of the potential applications for the top-hat transformation is to estimate uneven background illumination. However, top-hat transformation would underestimate the uneven background for small scale. Otherwise, it would overestimate the uneven background for large scale. In other words, if the selected scale is too small, the result obtained from the top-hat transformation is sensitive to noise. On the other hand, if the selected scale is too large, the top-hat transformation has no effect on the original image, since the opening is a flat image. Utilizing (16) and (17), we formulated previous remark as followings:

$$\lim_{t \to 0} \frac{T_t u}{t} = \lim_{t \to 0} \frac{u - O_t u}{t} = \bigtriangledown u(x), \qquad (18)$$

and

$$\lim_{t \to \infty} T_t u(x) = \lim_{t \to \infty} (u - O_t u) = u - C. \qquad (19)$$

In order to choose the proper scale in the top-hat transformation based image denoising, we utilize the mutual information of (9) and (10) as the quantitative criterion and present the following scale selection algorithm (*i-De*).

Algorithm: *i-De*
Parameter:
    Denote T as the number of iteration,
    ε as a small threshold parameter.
    Calculate $I_1^t(u, O_t u)$ at the initial scale $t = 1$.
Do for $t = 2, ..., T$

1.Compute top-hat transformation as in (17);

2.Calculate mutual information I_1 and I_2 as follows:

$$I_1 = \frac{I(u, O_t u)}{I(u, u)}, \qquad (20)$$

$$I_2 = \frac{I(u, u - O_t u)}{I(u, u)}. \qquad (21)$$

3.If $I_1 < I_2$ and $abs\left(I_1^{t-1} - I_1^t\right) < \varepsilon$, then output selected scale t.

In the following experiments, we set the small threshold parameter ε to 0.01.

2.2.2 Analysis of the *i-De* algorithm

For the feasibility of the *i-De* algorithm, we give some proofs for the monotonicity of the proposed formula. We firstly introduce the following theorem [33] upon the open (or sup-inf) operator.

Theorem 1. Let $\mathbf{F}\,(u \in \mathbf{F})$ be a set of functions and it is stable under contrast change, and O be a function operator having a sup-inf form, namely $O_t u(x) = \sup\limits_{B' \in tB}\ \inf\limits_{y \in x+B'} u(y)$. Then the sup-inf operator is monotone and contrast invariant.

According to monotonicity of the open operator given in theorem 1, we introduce the following two propositions that describe the monotonicity of (20) and (21).

Proposition 1. Let O be an open operator, T be a top-hat transformation, and $I(u, v)$ be the mutual information for two random variables $u(x)$ and $v(x)$. Then the mutual information defined by (20) is monotonically decreasing, and the mutual information defined by (21) is monotonically increasing.

Proof. Equations (15) and (16) show that the opened image is the original image at scale 0 and it tends to a flat image at infinite scale. Combining with the monotonicity stated in theorem 1 we tell that the opened image $O_t u(x)$ is decreasing monotonously from original image to the flat image as the scale increasing from 0 to infinite. Meanwhile, maximum mutual information is obtained only when we calculate the mutual information of the image to itself, i.e., $I(u, u) \geq I(u, v)$. Thus we conclude that the mutual information of (20) is decreasing monotonously. Similarly, the mutual information of (21) is increasing monotonously. $\square$.

Remark. In our experimental results, we observed that the mutual information of (20) is indeed decreasing monotonously, which conforms to our proposition 1. However, the mutual information of (21) is increasing gradually with very small oscillation, but not monotonicity. Such a phenomenon comes from the fact that the computation of (20) is insensitive to noise since

the opened image could be viewed as a smoothed version to the original image. Conversely, the top-hat transformed image is the residual of the opened image to the original image. In other words, top-hat transformed image is comparable to noise at small scale, just as demonstrated in (18). Such an oscillation is caused by the existence of noise, but the trend of the mutual information of (21) is increasing.

Proposition 2. The formula of mutual information defined in (20) and (21) will intersect at a certain scale.

Proof. By substituting (15) and (16) into (20), we could easily conclude that the mutual information of (20) is decreasing from maximum mutual information (i.e.1) to minimum mutual information (i.e. 0). Similarly, by substituting (18) and (19) into (21), we conclude that the (21) is increasing from 0 to 1. Thus there is an intersection for (20) and (21) at a certain scale. $\square$.

Remark. According to proposition 2, there is an intersection point for the curves of (20) and (21) as the scale t increases. And this point is indeed a turning point as both of the mutual information become convergent. However, the scale at this point is not optimal as it is demonstrated in our experimental results. And the *i-De* algorithm gives a more optimized answer.

Not only does the scale model resolve the morphological denoising but also unsupervised clustering. Next we present another algorithm of using the scale model to determine the NC in unsupervised segmentation.

2.3 Maximization of mutual information for segmentation

2.3.1 Description of the segmentation algorithm

Most of the unsupervised segmentation algorithms require some prior knowledge on the determination of NC. For example, we have the prior that the brain image of magnetic resonance imaging (MRI) can be classified into cerebrospinal fluid, white matter, grey matter and background. However, it is sometimes difficult or impossible for us to know the true NC without the specific domain knowledge.

In this section, we present an algorithm for solving the problem of selecting the proper or optimal NC in unsupervised image segmentation, where level set functions are commonly used to describe the contrast invariant properties [33]. The level set of an image $u(x)$ with level t is written as $\Phi_t u(x) = \{x : u(x) > t\}$, for $t \in \mathbf{R}$. Here we treat a specific level t as a threshold for image segmentation. By adopting the Huffman entropy coding methodology, our algorithm subdivides an image into similar regions recursively according to a predefined criterion. The algorithm and its analysis are based on the information-theoretical functional described in section 2.1 and the Huffman coding methodology. The Huffman

sorting for the histogram bins in each stage guarantees that our solution to the problem of mutual information maximization is not a local one, but a global one. We denote our proposed segmentation method as *i-Se* algorithm and its pseudo-code is described as follows.

Algorithm: *i-Se*

Parameter:
Denote T as the number of maximum intensity bin for an image $u(x)$, S as the clustering operation.
  Calculate $I_1^t(u, S_t u)$ at scale $t = T$.

Do for $t = T - 1, ..., 2, 1$

1. Calculate probability density $p_u(i), i \in [1, t]$;
2. Sort $p_u(i)$ in descending order;
3. Combine the lowest two ordered probabilities, and produce a weighted intensity as follows:

$$k = \frac{i * p_u(i) + j * p_u(j)}{p_u(i) + p_u(j)}. \qquad (22)$$

Now the probability of intensity k becomes:

$$p_u^{t-1}(k) = p_u^t(k) + p_u^{'}(k), \qquad (23)$$

where

$$p_u^{'}(k) = p_u(i) + p_u(j). \qquad (24)$$

And the following diagram illustrates our clustering process based on Huffman coding strategy:

t	t−1	t−2
$p_u(r)$	$p_u(r)$	
$\vdots$	$\vdots$	
	$\to p_u(k)$	$\to \vdots$
$p_u(j)$	$p_u(s)$	
$p_u(i)$		
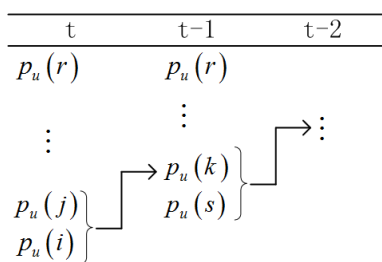
4. Calculate following mutual information:

$$I_1^t = (u, S_t u), \qquad (25)$$

$$I_2^t = (u, u - S_t u). \qquad (26)$$

5. If $I_1 < I_2$ and $abs\left(I_1^{t+1} - I_1^t\right) < \varepsilon$, then output selected NC t.

2.3.2 Analysis of the segmentation algorithm

The main idea behind the analysis for *i-Se* algorithms is to derive monotonic and smooth properties on the mutual information based formula of (25) and further to point out that the solution to the *i-Se* algorithm is convex. We begin with the following two inequality equations (proofs for the following two lemmas are given in the appendix).

Lemma 1. Let p, q be the probability density function and $r = p + q$, then we have $-r log(r) < -(p log(p) + q log(q))$.

Lemma 2. Let $S_t u(x)$ be the segmented image at the level of t for an image $u(x)$. Then the joint entropy $H(u, S_t u)$ is invariable at any clustering level of t.

  With the two lemmas, we have the following proposition.

Proposition 3. Let $S_t u(x)$ be the segmented image at the clustering level of t for a random variable of $u(x)$. With the attenuation of level variable t, the mutual information for $I(u, S_t u)$ is monotonically decreasing.

Proof. To prove the above proposition, we only need to prove that the mutual information of $I(u, S_t u)$ is greater than $I(u, S_{t-1} u)$. From (1), we have $I(u, S_t u) = H(u) + H(S_t u) - H(u, S_t u)$ and $I(u, S_{t-1} u) = H(u) + H(S_{t-1} u) - H(u, S_{t-1} u)$. According to the definition of Shannon entropy and lemma 1, we have $H(S_t u) > H(S_{t-1} u)$. With the proved lemma 2, we immediately have $I(u, S_t u) > I(u, S_{t-1} u)$. $\square$

  Construction of Huffman coding tree in the *i-Se* algorithm plays an important role in obtaining global minima. Otherwise, the convexity for the mutual information of (25) is not guaranteed. Such non-convex solution is commonly existed in classical FCM algorithm and could be obviously observed in our following experimental result for the classification of the computerized tomography (CT) phantom image. In other words, the solution for FCM is local minima. Thus, to find globally minima solution, the *i-Se* algorithm adopts the method for the construction of minimum-redundancy codes in Huffman entropy coding by merging the lowest two ordered probability density bins successively. Meanwhile, from proposition 3 we have:

$$dI^t = I_1^t(u, S_t u) - I_1^{t-1}(u, S_{t-1} u) = \triangle H^t(S_t u). \qquad (27)$$

  Equation (27) indicates that it is redundancy to compute some of the terms for the mutual information in (25). Thus, we simplify the *i-Se* algorithm to increase computation speed by deleting redundancy term.

Algorithm: *i-Se (simplified version)*

Parameter:
Denote T as the number of maximum intensity bin for an image $u(x)$, S as the clustering operation.
  Calculate entropy $H^t(S_t u)$ at scale $t = T$.

Do for $t = T - 1, ..., 2, 1$

1. Calculate probability density $p_u(i), i \in [1, t]$;
2. Sort $p_u(i)$ in descending order;
3. Combine the lowest two ordered probabilities, and produce a weighted intensity as in (22);
4. Calculate Shannon entropy as follows:

$$H^t = H^t(S_t u). \tag{28}$$

5. If $\Delta H^t = H^{t+1} - H^t < \varepsilon$, then output selected NC t.

3 Experimental results

3.1 Denoising results

In the experiments of multi-scale *i-De* denoising algorithm, we focus on the non-uniform background, which could be effectively subtracted from the original image through top-hat transformation [46]. And the transformed images are further segmented to validate the *i-De* scale selection algorithm.

Figure 1(a) is an original rice grains image with uneven illumination. After the threshold segmentation applied to the original image, we obtain the segmented result in Fig. 1(d), where we could observe that the rice grains at the bottom of the image could not be extracted ideally from the uneven background. It is pointed out in [46] that as long as the SE is large enough, a reasonable estimation of the background illumination could be obtained by opening the original image. But it is subjective to select the appropriate scale. Figure 2 demonstrates the diagram for the mutual information of (20) and (21) with the increasing scale t and it conforms to our prediction to the I_1 and I_2 in proposition 1: I_1 decreases monotonously and I_2 increases monotonously. However, I_1 and I_2 are not convex functions. Its non-convex property is demonstrated in Fig. 3, where it illustrates the oscillation for the difference of I_1 and I_2. Figure 1(b) illustrates the opened image at scale 6, where the mutual information of (20) and (21) is intersected. And it is envisaged that the scale at this intersection point is not optimal. It is incorrect for the estimation of a rice grain marked out in a circle in Fig. 1(b). Conforming to the empirical selection of the scale in [46], the *i-De* scale selection algorithm has chosen the scale 10 and the promising results are illustrated in Figs. 1(c) and (f).

Figure 4(a) is an X-ray vessel medical image with uneven background. Figure 4(e) demonstrates the segmentation result by using the level set method [47] to the original image. The level set stop evolving to the left bottom part of the vessel and fail to segment such a vessel image with uneven background. In this study, top-hat transformation is used to eliminate the uneven background. The mutual information and its difference of (20) and (21) for the vessel image are demonstrated in Figs. 5 and 6, respectively. And the *i-De* scale selection
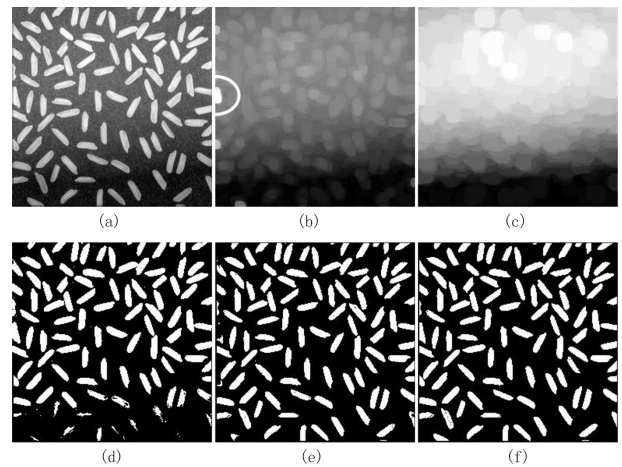

Fig. 1: Denoising for rice grains image. (a) Original image. (b) Opened image at scale 6. (c) Opened image at scale 10. (d) Threshold image at scale 0. (e) Threshold image at scale 6. (f) Threshold image at scale 10.
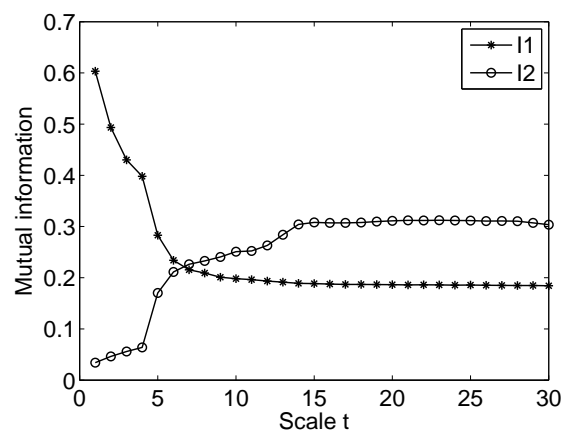

Fig. 2: Mutual information of (20) and (21) for grains image at various scales.

algorithm predicts that the optimal scale is 15. For comparison with other scales, we have chosen the representative scales at 5 and 30. The images in the first row of Fig. 4 are the top-hat transformed images at scale 0, 5, 15, and 30, respectively. The images in second row are the corresponding segmentation results using level set method. The third row shows the estimation of uneven background. From these images, we could conclude that if the scale is not sufficiently large, it would produce an underestimated image, which could lead to some leakages in the vessels boundary marked as circles in Fig. 4(f). On the other hand, larger scale selection would come to an overestimation to the background. Only the proper scale could result in the promising segmentation result demonstrated in Fig. 4(g).
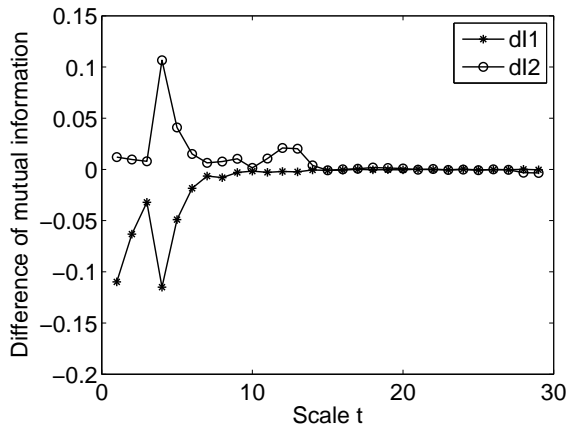
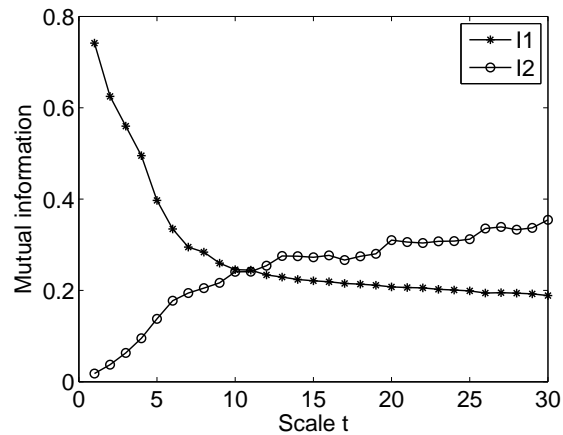**Fig. 3:** Difference of mutual information for Fig. 2.

Fig. 5: Mutual information of (20) and (21) for vessel image at various scales.
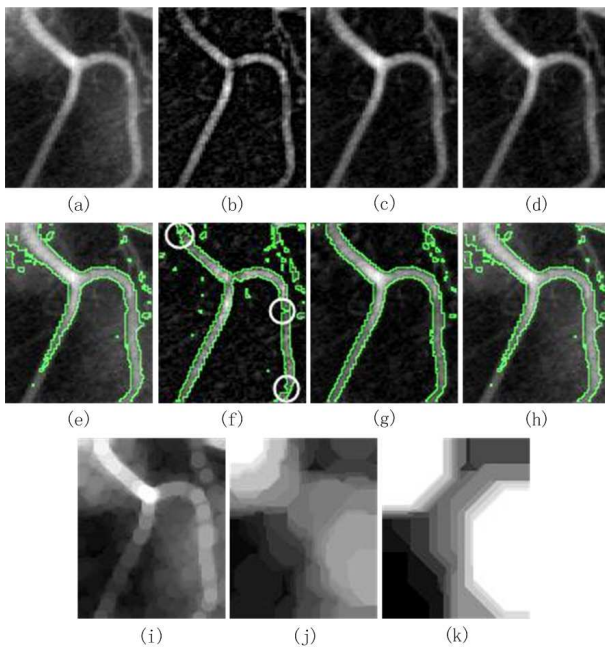

Fig. 4: Denoising for vessel image. The images in first row are $T_t u$ at scale 0, 5, 15, and 30, respectively. The images in second row are the segmentation results corresponding to the first row using level set method. The third row shows the estimation of uneven background: $O_t u$, corresponding to the scale t=5, 15, 30.

Fig. 6: Difference of mutual information for Fig. 5.

3.2 Segmentation results

The first experiment for testing the *i-Se* clustering algorithm is conducted on a Shepp-Logan head phantom image, which is used widely by researchers for validating their numerical accuracy of any two-dimensional CT reconstruction algorithms. We choose such a test image for its six known clusters, which has the intensity value of 0, 26, 51, 77, 102, and 255, respectively.
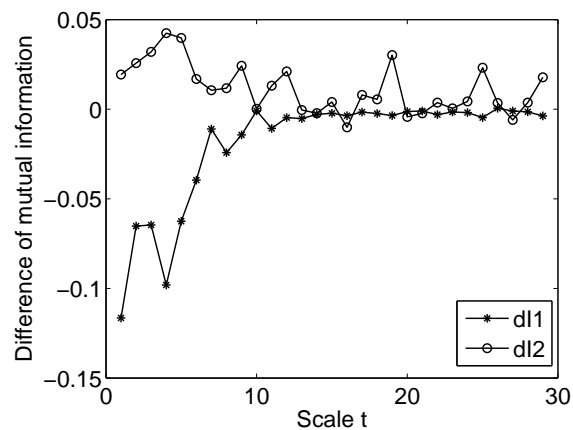
Figures 7 and 8 illustrate the segmentation results by using the FCM and the *i-Se* clustering methods. The relationship between mutual information and NC are illustrated in Figs. 9 and 10, where the mutual information of formulas (25) and (26) are plotted. It verified to our proposition 3 that the mutual information of I_1 is increasing monotonously and convergent to 1.0 as the NC increases. Meanwhile, I_2 is decreasing monotonously and reaches to 0.0 as the NC $\geq$ 6.

Figure 11 demonstrates that the amount of mutual information for the *i-Se* algorithm is more than for the FCM clustering algorithm when the NC is 2 or 3. It means that the segmented result from the *i-Se* algorithm contains more detail information in the image, which is envisaged by comparing Figs. 7 (b) and (c) with Figs. 8 (b) and (c). Meanwhile, the FCM method is a local algorithm and leads to different segmentation results due to the different initialization for clustering center. Such
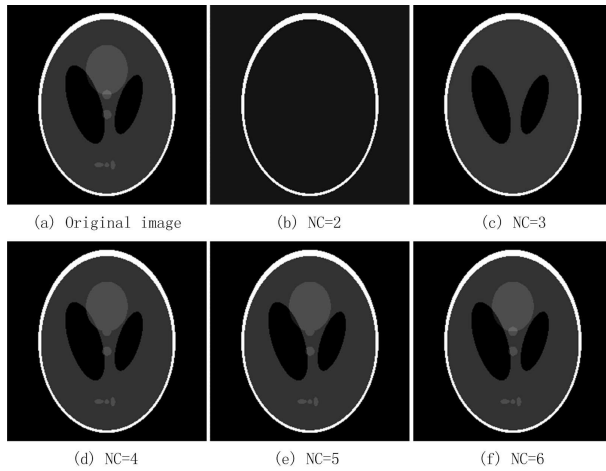
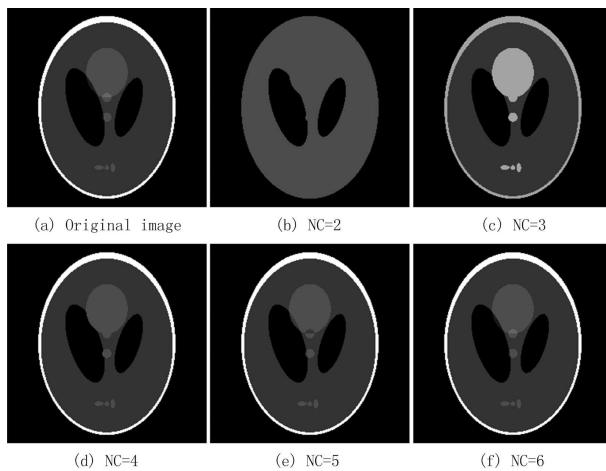**Fig. 7:** Segmentation of head phantom image using FCM algorithm.

Fig. 9: Mutual information of FCM algorithm for head phantom image.

Fig. 8: Segmentation of phantom image using *i-Se* algorithm.
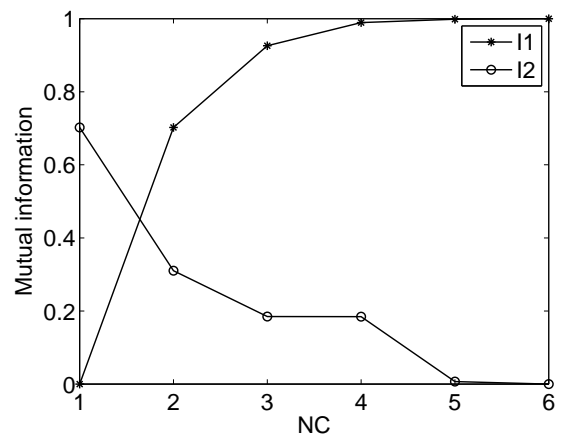
Fig. 10: Mutual information of *i-Se* algorithm for head phantom image.

local phenomenon is illustrated in Fig. 12, where such local method could not guarantee the monotonicity for the difference of mutual information. On the contrary, Fig. 12 demonstrates that the curve of the difference of mutual information for the *i-Se* algorithm is smoother. More importantly, it is monotonically decreasing.

The second experiment for testing the *i-Se* algorithm is conducted on a real cerebral CT image. It is not a synthetic image and we have no prior knowledge to its exact NC. Figure 13 shows the testing CT image and the segmentation results using FCM and *i-Se* algorithms. The relationships between mutual information and NC, and between the difference of mutual information and NC are demonstrated in Figs. 14 and 15, respectively. As shown in Fig. 14, with the growth of NC, the mutual information of equation (25) is increasing both for the *i-Se* and FCM algorithms. But it could also be observed that the mutual
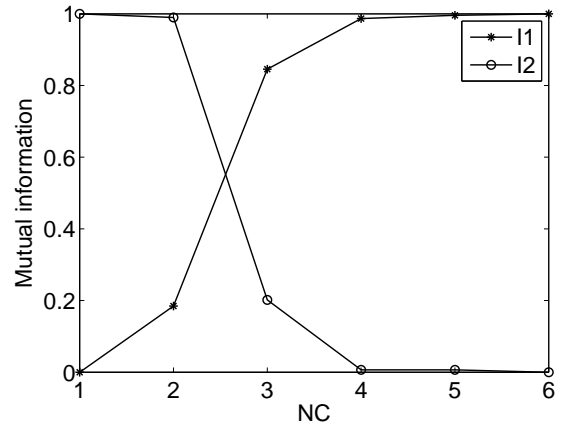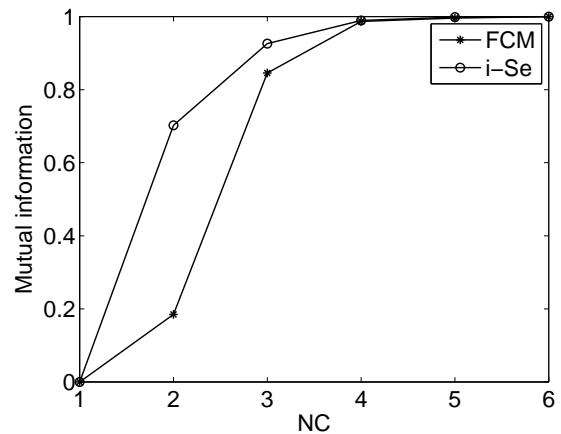

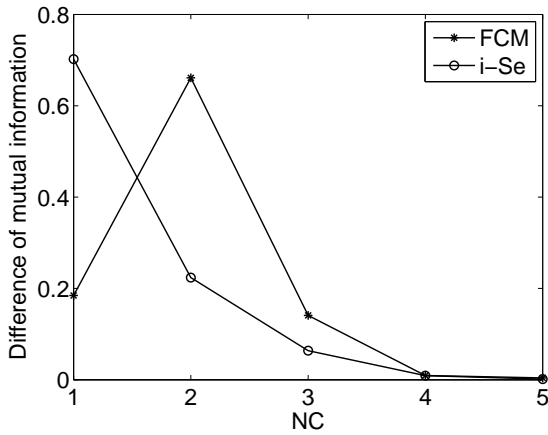**Fig. 11:** Comparison of mutual information for FCM and *i-Se* algorithms.

Fig. 12: Comparison of the difference of mutual information for FCM and *i-Se* algorithms.
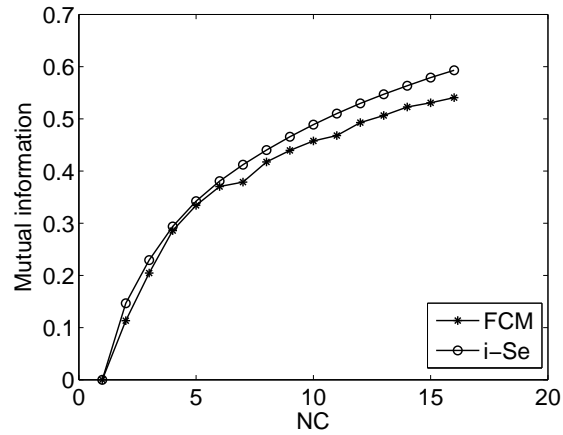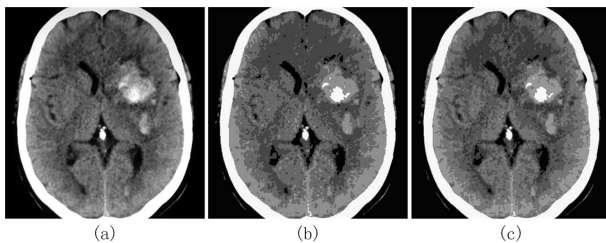

Fig. 13: Segmentation for CT image. (a) Original image. (b) FCM segmentation result. (c) *i-Se* segmentation result.

Fig. 14: Comparison of mutual information for CT image using FCM and *i-Se* algorithms.

Fig. 15: Comparison of the difference of mutual information for CT image using FCM and *i-Se* algorithms.

Fig. 16: Segmentation for Lena image. (a) Original image. (b) FCM segmentation result. (c) *i-Se* segmentation result.
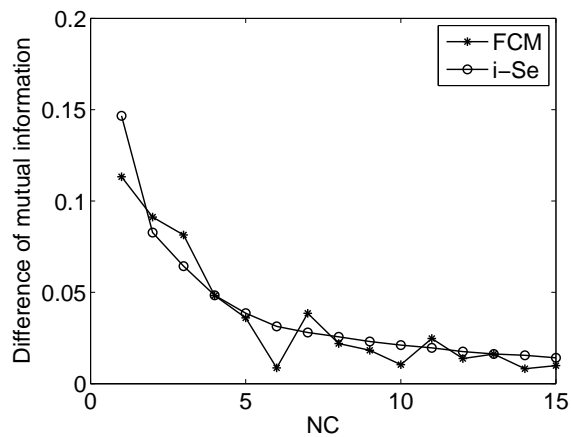
information is greater for the *i-Se* algorithm than for the FCM algorithm in the case of the same NC. In comparison, more detail information is preserved in the *i-Se* segmented image compared with the FCM segmentation result. Such advantage is even more obvious with the increase of NC. As is shown in Fig. 15, with the growth of NC, the difference of mutual information for the FCM algorithm is oscillatory decreasing, while the decreasing for the *i-Se* algorithm is monotonically decreasing.

The third experiment for testing the *i-Se* algorithm is conducted on the Lena image. The experimental results are demonstrated in Figs. 16-18. To compare the results between cerebral CT image and Lena image, our method can effectively preserve the detailed information in the segmented images. More intensive experiments are conducted on other testing images to validate the effectiveness of the *i-Se* algorithm, which reveals the internal relationship among the original image and the segmented images.
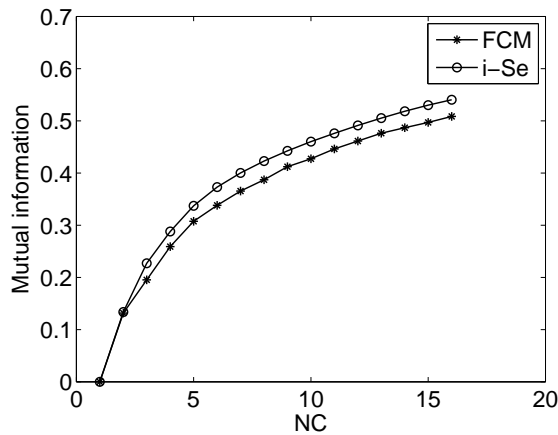
Fig. 17: Comparison of mutual information for Lena image using FCM and *i-Se* algorithms.
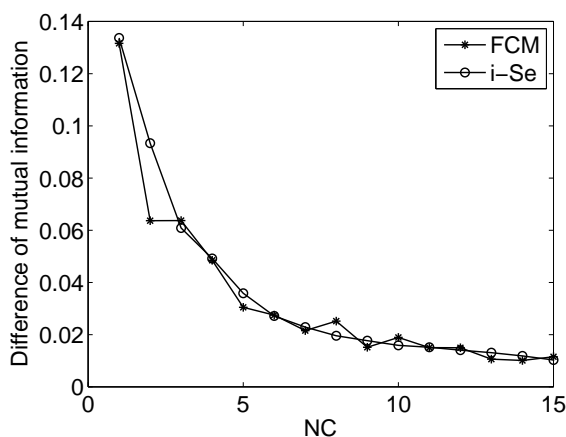

Fig. 18: Comparison of the difference of mutual information for Lena image using FCM and *i-Se* algorithms.

4 Conclusion and discussion

In this paper, we make use of mutual information as a measuring criterion of similarity for the optimal scale selection in multi-scale analysis, with applications to image denoising and segmentation. Firstly we express the image processing operator in multi-scale space and propose a mutual information based functional. To morphological open operator, we present a *i-De* scale selection algorithm, whose advantage is that it can automate the scale selection for the multi-scale representation of images. To clustering operator, we present a *i-Se* clustering algorithm by treating the specific level *t* as a threshold for image segmentation. For the feasibility of the *i-De* and *i-Se* algorithms, we make the detailed mathematical analysis, mainly to the monotonicity and smoothness of the objective formulas.

Using the *i-De* algorithm, we are capable of selecting the proper scale to eliminate uneven background. Such non-uniform illumination is a common phenomenon in most camera or medical images. Finally the *i-Se* algorithm outputs the proper level for clustering segmentation. At such scale we can obtain the maximum amount of information for segmented image.

A weakness of our *i-Se* algorithm is that only the grey feature is considered in the model, while other features (such as spatial correlation among pixels, boundary, and texture) are not taken into account. A more detailed model for their local correlation would be needed in order to produce more reliable results.

Many other modeling tasks, such as chemical, physical, and biological modeling, could have been explored using the scale-space theory. However, in the present study our main aim was to develop a scale selection criterion for image denoising and segmentation tasks. In the future, we will do more investigation for the application of our algorithm to other modeling tasks. And most of these limitations will be dealt with in future research.

Appendix

Following we give the detailed proof for the lemmas in section 2.3.

Lemma 1. Let p, q be the probability density function and $r = p+q$, then we have $-rlog(r) < -(plog(p)+qlog(q))$.

Proof. As p, q are probability functions, we have $0 \leq p \leq 1$ and $0 \leq q \leq 1$. By an increasing argument for the function $f(x) = x^a$, $a \in [0,1]$, $x \in (0,\infty)$, it can be shown that $(p+q)^p > p^p$ and $(p+q)^q > q^q$. So we have $r^r = (p+q)^{p+q} > p^p * q^q$. Applying log operator to both side yields $rlog(r) > (plog(p) + qlog(q))$. The lemma follows immediately. □

Lemma 2. Let $S_t u(x)$ be the segmented image at the level of t for an image $u(x)$. Then the joint entropy $H(u, S_t u)$ is invariable at any clustering level of t.

Proof. For the self-joint entropy $H(u,u)$, we have $p_{uu}(i,j)|_{i=j} \neq 0$ and $p_{uu}(i,j)|_{i \neq j} = 0$. In other words, the self-joint entropy is a diagonal matrix. Providing that we combine any of two intensity probabilities estimated from $u(x)$, e.g. i and j, it produces a weighted intensity as (22). Now the probability of intensity k for $S_t u(x)$ becomes (23). So we have the following shifting for the joint histogram in the diagonal matrix of $H(u,u)$: $p_{uu}(k,k) \rightarrow p_{uS_t u}(k,k)$, $p_{uu}(i,i) \rightarrow p_{uS_t u}(i,k)$ and $p_{uu}(j,j) \rightarrow p_{uS_t u}(j,k)$. As clustering only shifts the element in the joint distribution, it does not change the total value for the joint entropy. That means $H(u,u) = H(u, S_t u)$. Similarly, we have $H(u,u) = H(u, S_{t-1}u)$. So we have $H(u, S_t u) = H(u, S_{t-1}u)$ and the lemma follows immediately. □

Acknowledgement

References

[1] H. Spohn, Large Scale Dynamics of Interacting Particles, Springer-Verlag, New York, (1991).

[2] M. A. Katsoulakis, A. J. Majda, D. G. Vlachos, Journal of Computational Physics, **186**, 250-278 (2003).

[3] L. Xiao, L. X. Chen, J. Z. Xiao, Appl. Math. Inf. Sci., **6**, 397-400 (2012).

[4] F. Pappalardo, A. Palladini, M. Pennisi, F. Castiglione, S. Motta, Mathematical Modelling of Natural Phenomena, **7**, 186-203 (2012).

[5] R. Keunings, Rheology Reviews, 67-98 (2004).

[6] R. Erban and H. G. Othmer, SIAM Journal on Applied Mathematics, **65**, 361-391 (2004).

[7] D. Alemani, F. Pappalardo, M. Pennisi, S. Motta, V. Brusic, Journal of Immunological Methods, **376**, 55-68 (2012).

[8] C. Bradley, A. Bowery, R. Britten, et al., Progress in Biophysics and Molecular Biology, **101**, 32-47 (2011).

[9] A. Bellouquid and C. Bianca, Mathematical and Computer Modelling, **52**, 802-813 (2010).

[10] D. Marr, Vision, Freeman Publishers, Dallas, (1982).

[11] J. J. Koenderink, Biological Cybernetics, **50**, 363-370 (1984).

[12] T. Lindeberg, Scale-Space Theory in Computer Vision, Kluwer Academic Publishers, Norwell, (1994).

[13] A. P. Witkin, Proceedings of the 8th International Joint Conference on Artificial Intelligence, **2**, 1019-1022 (1983).

[14] J. Babaud, A. P. Witkin, M. Baudin, R. O. Duda, IEEE Transactions on Pattern Analysis and Machine Intelligence, **8**, 26-33 (1986).

[15] S. Mallat, A Wavelet Tour of Signal Processing, Academic Press, Utah, (1988).

[16] J. Portilla, V. Strela, M. J. Wainwright, E. P. Simoncelli, IEEE Transactions on Image Processing **12**, 1338-1351 (2003).

[17] A. S. Lewis and G. Knowles, IEEE Transactions on Image Processing, **1**, 244-250 (1992).

[18] N. Adami, A. Signoroni, R. Leonardi, IEEE Transactions on Circuits and Systems for Video Technology, **17**, 1238-1255 (2007).

[19] R. L. Allen, F. A. Kamangar, E. M. Stokely, IEEE Transactions on Signal Processing, **41**, 3536-3541 (1993).

[20] G. Sapiro, Geometric Partial Differential Equations and Image Analysis, Cambridge University Press, Cambridge, (2001).

[21] P. Perona and J. Malik, IEEE Transactions on Pattern Analysis and Machine Intelligence, **12**, 629-639 (1990).

[22] L. I. Rudin, S. Osher, E. Fatemi, Physica D, **60**, 259-268 (1992).

[23] M. J. Black, G. Sapiro, D. H. Marimont, D. Heeger, IEEE Transactions on Image Processing, **7**, 421-432 (1998).

[24] D. Tschumperle, International Journal of Computer Vision, **68**, 65-82 (2006).

[25] K. L. Vincken, A. S. E. Koster, M. A. Viergever, IEEE Transactions on Pattern Analysis and Machine Intelligence, **19**, 109-120 (1997).

[26] Y. Leung, J. S. Zhang, Z. B. Xu, IEEE Transactions on Pattern Analysis and Machine Intelligence, **22**, 1396-1410 (2000).

[27] A. Petrovic, O. D. Escoda, P. Vandergheynst, IEEE Transactions on Image Processing, **13**, 1104-1114 (2004).

[28] M. Bertalmio, G. Sapiro, V. Caselles, C. Ballester, Proceedings of SIGGRAPH 2000, 417-424 (2000).

[29] M. Bertalmio, et al., IEEE Computer Society Conference on Computer Vision and Pattern Recognition, **1**, 355-362 (2001).

[30] T. F. Chan and J. H. Shen, Communications on Pure and Applied Mathematics, **58**, 579-619 (2005).

[31] J. Serra, Image Analysis and Mathematical Morphology, Academic Press, Utah, (1982).

[32] P. T. Jackway and M. Deriche, IEEE Transactions on Pattern Analysis and Machine Intelligence, **18**, 38-51 (1996).

[33] F. Guichard, L. Moisan, J. M. Morel, Journal de Physique IV France, **12**, 137-154 (2002).

[34] T. Lindeberg, International Journal of Computer Vision, **30**, 117-154 (1998).

[35] I. C. Dolcetta and R. Ferretti, Applied Mathematics and Optimization, **43**, 245-258 (2001).

[36] V. Solo, IEEE International Conference on Acoustics, Speech, and Signal Processing, **2**, 1661-1664 (2002).

[37] P. Mrazek and M. Navara, International Journal of Computer Vision, **52**, 189-203 (2003).

[38] G. Papandreou and P. Maragos, International Conference on Image Processing, **1**, 1033-1036 (2005).

[39] G. Papandreou and P. Maragos, IEEE Computer Society Conference on Computer Vision and Pattern Recognition, **1**, 625-630 (2005).

[40] J. Sun, Z. B. Xu, Pattern Recognition, **43**, 2630-2645 (2010).

[41] T. X. Wen, J. Gu, Z. Q. Zhang, L. Wang, 3rd International Congress on Image and Signal Processing, **3**, 1092-1096 (2010).

[42] J. Rigau, M. Feixas, M. Sbert, A. Bardera, I. Boada, Medical Image Computing and Computer-Assisted Intervention, **3216**, 135-142 (2004).

[43] Q. W. Lu and W. F. Chen, Science in China Series F-Information Sciences, **49**, 484-493 (2006).

[44] Z. T. Lu, Q. J. Feng, P. C. Shi, W. F. Chen, IEEE/ICME International Conference on Complex Medical Engineering, **1**, 513-516 (2007).

[45] J. C. Bezdek, Pattern Recognition with Fuzzy Objective Function Algorithms, Plenum Press, New York, (1981).

[46] R. C. Gonzalez, R.E. Woods, S.L. Eddins, Digital Image Processing Using MATLAB, Publishing House of Electronics Industry, Beijing, (2004).

[47] T. Chan and L. Vese, IEEE Transactions on Image Processing, **10**, 266-277 (2001).

1742

T. Wen et al.: Mutual Information Criterion-Based Optimal Scale...

Tiexiang Wen received his BS and MS degrees in biomedical engineering from Southern Medical University, Guangzhou, China. He is currently an assistant professor in Research Center for Medical Robotics and Minimally Invasive Surgical Devices, Institute of Biomedical and Health Engineering, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences. His research interests are in the areas of computer vision and image processing.

Qingsong Zhu received the BS and MS degrees in computer science from University of Science and Technology of China (USTC), Hefei, China. He is currently an assistant professor in Research Center for Medical Robotics and Minimally Invasive Surgical Devices, Institute of Biomedical and Health Engineering, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences. His current research interests focus on computer vision, statistical pattern recognition, machine learning and robotics. He is a member of the IEEE.

Jia Gu received his master and Ph.D. degree in 2001 and 2005 from Southeast University (China) and University of Rennes (France), respectively. Dr. Gu has working experiences in Samsung Advanced Institute of Technology (Beijing) and STI Medical Systems (Hawaii), on a wide variety of research projects on computer vision and medical imaging. Since 2008, Dr. Gu was recruited by Shenzhen Institutes of Advanced Technology of Chinese Academy of Sciences, as a full professor.