Computers & Graphics 37 (2013) 553–564

Contents lists available at SciVerse ScienceDirect

Computers & Graphics

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

SMI 2013

New evaluation metrics for mesh segmentation

Zhenbao Liu [a], Sicong Tang [a], Shuhui Bu [a,*], Hao Zhang [b]

[a] Northwestern Polytechnical University, China
[b] Simon Fraser University, Canada

ARTICLE INFO

Article history:
Received 18 March 2013
Received in revised form
30 May 2013
Accepted 30 May 2013
Available online 10 June 2013

Keywords:
Mesh segmentation
Evaluation metric
Similarity Hamming Distance
Adaptive Entropy Increment

ABSTRACT

3D model segmentation avails to skeleton extraction, shape partial matching, shape correspondence, texture mapping, shape deformation, and shape annotation. Many excellent solutions have been proposed in the last decade. How to efficiently evaluate these methods and impartially compare their performances are important issues. Since the Princeton segmentation benchmark has been proposed, their four representative metrics have been extensively adopted to evaluate segmentation algorithms. However, comparison to only a fixed ground-truth is problematic because objects have many semantic segmentations, hence we propose two novel metrics to support comparison with multiple ground-truth segmentations, which are named Similarity Hamming Distance (SHD) and Adaptive Entropy Increment (AEI). SHD is based on partial similarity correspondences between automatic segmentation and ground-truth segmentations, and AEI measures entropy change when an automatic segmentation is added to a set of different ground-truth segmentations. A group of experiments demonstrates that the metrics are able to provide relatively higher discriminative power and stability when evaluating different hierarchical segmentations, and also provide an effective evaluation more consistent with human perception.

1. Introduction

Decomposing 3D models into meaningful parts has been an increasing topic in the shape analysis community. The tasks such as skeleton extraction, shape partial matching, shape correspondence, texture mapping, shape deformation, and shape annotation heavily rely on 3D model segmentation. Many methods have attempted to provide better segmentation solutions, however, determining which method is superior to other methods is not an easy task. Similar as many shape retrieval benchmarks proposed previously, Benhabiles et al. [1] first provide a pioneering framework to quantitatively evaluate segmentation algorithms. Chen et al. [2] also propose a benchmark, which comprises a dataset with 4300 manually generated segmentations for 380 surface meshes of 19 different object categories. In addition, it offers four quantitative metrics for comparison of segmentations. The four metrics are obtained by extending metrics from image segmentation, and researchers adopt part or all of metrics to test their methods. Although these metrics are widely accepted by researchers, one-to-one comparison between automatic and ground-truth segmentation, and the way of averaging on all the comparisons limits their performance. Moreover, they are unable to be directly applied to multiple standard comparison.

In order to provide this supplement to the Princeton segmentation benchmark, in this paper, we focus on introducing two metrics, Similarity Hamming Distance (SHD) and Adaptive Entropy Increment (AEI). They jointly adopt all the ground-truth segmentations of each model to generate a score for automatic segmentation, which is different from averaging one-to-one comparisons between automatic segmentation and standard segmentation. SHD is based on partial similarity correspondences between automatic segmentation and ground-truth segmentations. For any segment of the input mesh, the metric searches its optimal corresponding part from all the ground-truth segmentations of the same model instead of only one ground-truth segmentation. These corresponding parts possibly from different segmentations are used to calculate the final error. Semantic information contained in the corresponding relationship between each segment of the input mesh and its corresponding part makes evaluation more rational and intuitive. The other metric, AEI, is based on the entropy concept from information theory, which measures the uncertainty associated with a random variable. We consider diversity and disorder of different segmentations on the same shape, and model this type of diversity and disorder using a group of random variables. Their entropy can be introduced to measure diversity and disorder of segmentations, and the problem of estimating segmentation quality can also be converted to entropy comparison. The entropy of all the different ground-truth segmentations

* Corresponding author.
E-mail addresses: liuzhenbao@nwpu.edu.cn (Z. Liu), bushuhui@nwpu.edu.cn (S. Bu), haoz@cs.sfu.ca (H. Zhang).

0097-8493/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.cag.2013.05.021

forms a baseline. When a novel automatic segmentation generated via an algorithm is added, the entropy increases from the baseline. Amplitude of entropy increment is adopted to evaluate the quality of automatic segmentation.

A group of experiments shows that the proposed metrics are able to provide higher discriminative power and effective evaluation consistent with human perception, and also robust to different hierarchical segmentations. We will integrate the two metrics into the Princeton segmentation benchmark for making them used conveniently in the future.

The rest of this paper is organized as follows. Recent works in shape segmentation and evaluation metrics will be discussed in Section 2. The first novel metric SHD will be introduced in Section 3, and the other metric AEI will be given in Section 4 and 5. We will demonstrate a group of experimental results and compare the two proposed metrics to the existent four metrics in Section 6. The work will be concluded in Section 7.

2. Related works

3D model segmentation has become a fundamental issue in computer graphics, which absorbed many researchers in the recent decade. The tasks have mainly focused on segmentation of a single shape and co-segmentation of a set of shapes. In this section, we briefly survey two closely related topics: segmentation methods and segmentation evaluation. According to the differences between segmentation techniques employed, we divide recent methods into three categories, including low-level geometric segmentation, learning based segmentation, and interactive segmentation. We will also discuss recent representative metrics for segmentation evaluation.

2.1. Segmentation methods

Low-level geometric segmentation: Many efforts have been made to find meaningful segmentations of 3D shapes in the recent decade. The early works usually focus on finding geometrical features used to provide segmentation criteria, and a detailed survey [16] classified previous segmentation solutions into meaningful part-type segmentation and surface-type segmentation partitioning the surface mesh into patches under some geometric criteria. Recent progress in discovering geometric properties includes diffusion distance [17], heat kernel [3], intrinsic primitive decomposition [13], heat walk [10], concavity-sensitive scalar fields [12], and minimum slice perimeter [14]. These geometric features are clustered in a descriptor space using clustering techniques such as recent Gaussian mixture models [18], greedy algorithm [10], and the Mumford–Shah model [15].

Learning based segmentation: To overcome the limitations inherent in segmentation of single shapes, a supervised learning based approach [4] has been considered to utilize a priori manual segmentations to obtain higher segmentation accuracy. It realized a data-driven approach by optimizing a conditional random field whose objective function is learned from labeled train data. Another work [7] learned an objective boundary edge function from a set of segmented training meshes in an off-line step, and the learned function is used to segment any input model in an on-line step. This type of methods needs a large set of manual segmentation data to train classifiers. In order to avoid the problem and simultaneously enhance the robustness to large shape variation, researchers have turned to unsupervised methods based on co-analysis of a set of shapes from the same class. Huang et al. [6] presented an unsupervised approach which optimizes over possible segmentations of individual shapes as well as over possible correspondences between segments from multiple shapes. Sidi et al. [19] and Hu et al. [20] described unsupervised co-segmentation methods based on descriptor-space spectral clustering and subspace clustering respectively. Iterative multi-label optimization [21] could also be applied to improve the co-segmentation, which is implemented via clustering over-segmented patches. Semi-supervised segmentation [22] is a trade-off solution to supervised and unsupervised methods.

Interactive segmentation: Because it is difficult for fully automatic segmentation to adapt to complex models and different applications, several works have introduced user assistance such as interactive sketching to obtain a relatively satisfactory segmentation. Under user's guidance such as defining the initial area coarsely or labeling foreground and background, they commonly provide intuitive interactive segmentation tools to find optimal cuts. Recent interactive segmentation tools include constrained random walk [23], bottom-up aggregation [24], graph-cut segmentation [25,26], harmonic field based method [27,11], geodesic curvature flow [28], dot scissor based on concavity-aware field [5], and semi-supervised learning based on cannot-link and must link constraints [29]. These algorithms generate natural segmentation seams by finding least cost paths. Meng et al. [9] investigated several popular foreground/background sketch-based interactive mesh segmentation algorithms, and performed an extensive comparative evaluation of these methods.

2.2. Segmentation evaluation

There are two types of evaluation methods: visual comparison and quantitative metrics. Visual comparison [30] is first adopted to visualize several segmented models generated by five early methods in different colors. Because visual results are limited to selected models, it is difficult to fully and fairly compare performances of algorithms. Four quantitative metrics are introduced into 3D segmentation by Chen et al. [2], which include Cut Discrepancy (CD), Hamming Distance (HD), Rand Index (RI), and Consistency Error (CE). These metrics are used to evaluate partition curves and regions generated by algorithms. The performances of seven representative methods including K-means (KM) [31], fitting primitives (FP) [32], core extraction (CE) [33], random walks (RW) [34], shape diameter (SD) [18], normalized cuts (NC) and randomized cuts (RC) [35], are investigated. They found that segmentation based on low-level geometric criteria did not perform well on all the test groups because these features are commonly sensitive to local surface perturbation, non-rigid deformation, and topology change. We summarized recent segmentation methods evaluated on the representative four metrics from the Princeton segmentation benchmark in Table 1. Although many methods are evaluated on the four metrics simultaneously, we consider that they cannot help to generate consistent evaluation

Table 1
Summary of recent papers adopting four metrics.

Papers	Metrics
Skraba et al. (2010) [3]	CD, HD, RI, CE
Kalogerakis et al. (2010) [4]	CE, RI
Zheng et al. (2011) [5]	CD, HD, RI, CE
Huang et al. (2011) [6]	CD, HD, RI, CE
Benhabiles et al. (2011) [7]	RI
Bergamasco et al. (2011) [8]	CD, HD, RI, CE
Meng et al. (2011) [9]	HD, RI, CE
Benjamin et al. (2011) [10]	CD, HD, RI, CE
Meng et al. (2011) [11]	RI, CD
Solomon et al. (2011) [13]	RI
Au et al. (2012) [12]	CD, HD, RI, CE
Ho et al. (2012) [14]	CD, HD, RI, CE
Zhang et al. (2012) [15]	RI, CE

results. There are two reasons: (1) CD measures segmentation boundaries while HD, RI, and CE are based on region differences of segmented surfaces. (2) CD, HD, and RI are sensitive to hierarchical segmentations, while CE is robust to them.

Benhabiles et al. [36] systematically analyzed the four metrics from several different viewpoints and suggested representative desirable properties of a metric. This is a pioneering work in the study of segmentation metrics. They also improved Rand Index by introducing a probabilistic interpretation. Their probabilistic Rand Index adopted a fast and efficient mean estimator over a generative model of correct segmentations, which can be understood as averaging the RI over multiple ground-truths, as mentioned in Section 3 of their paper. They then normalized it in order to increase its dynamic range, and obtained a higher performance of segmentation evaluation.

Kalogerakis et al. [4], Sidi et al. [19], and Lv et al. [22] adopted an accuracy measure on segmented regions to evaluate their methods. An indicator function of face is defined via comparison between the ground-truth face label and recognized label, and the segmentation score relies on area-weighted summation of indicator functions, divided by the total area. This evaluation method is based on one-to-one comparison.

Differently from previous works, we do not average metric values over multiple ground-truths, but fully exploit manual segmentation datasets and integrate possible similarity information of all ground-truth segmentations to give a comprehensive evaluation on an automatic segmentation. In order to provide the effective multi-standard comparison, we propose two evaluation metrics based on multiple ground-truth segmentations, Similarity Hamming Distance (SHD) and Adaptive Entropy Increment (AEI), to enhance discriminability against unreasonable segmentations, adaptability to complex and simple models, and tolerance to hierarchical segmentations.

3. Similarity Hamming Distance

The Hamming distance proposed by Huang et al. [37] is extended to 3D mesh segmentation evaluation [2], which compares region differences between two segmentations A and G. Suppose A is an automatic segmentation generated by a given algorithm and G is a ground-truth segmentation generated manually. In most cases G is not unique, and this is the reason why Chen et al. [2] collected segmentations of each model from multiple people. We consider two questions: (1) how to evaluate an automatic segmentation on multi-standard segmentations acquired from multiple people? (2) how to handle many cases in which there is no ground-truth segmentation corresponding to the automatic segmentation?

Consider the following situations illustrated in Fig. 1, where we assume that A is the automatic segmentation, and G_1 and G_2 are both ground-truth segmentations. From the point of view of human perception, it seems that A is a reasonable segmentation, but A is obviously different from G_1 and G_2. Comparing A with either of G_1 and G_2 would lead to an undesired error. And this problem exists in all the metrics based on one-to-one comparison. In fact, for the vast majority of 3D models, there are often many standard segmentations in accordance with human perception, especially for those models which contain many semantic parts (for example, humans and four-leg animals with many joints).

Here we introduce a new metric, Similarity Hamming Distance (SHD), which integrates possible similarity information of all ground-truth segmentations to give a comprehensive evaluation on an automatic segmentation. This metric is based on partial similarity correspondences between the automatic segmentation and ground-truth segmentations. For any segment of the input mesh, we search its corresponding part from all the ground-truth segmentations of the same model instead of only one ground-truth segmentation. These corresponding parts possibly from different segmentations are used to calculate the final evaluation score. It explores a potential semantic relation between each segment of the input mesh and its corresponding part, and makes evaluation more reasonable and intuitive.

As illustrated in Fig. 2, the steps of computing SHD are as follows:

Step 1: Let $\{G_1, \ldots, G_n\}$ be all the different ground-truth segmentations of the same model, and G_i^l is the l-th segment in the segmentation G_i. A denotes the automatic segmentation to be evaluated. We first choose one segment a_k from C segments in the segmentation A, and then search its overlapping segments in each ground-truth segmentation G_i. The overlapping segments of a_k in G_i compose a set $O(a_k)$, and each element $O_j(a_k)$ satisfies

$$O_j(a_k) \cap a_k = \{f : f \in O_j(a_k) \wedge f \in a_k\} \neq \varnothing, \tag{1}$$

where f denotes a face, and $\varnothing$ represents the empty set.

Step 2: The segment a_k is compared with each overlapping segment in the set $O(a_k)$ of G_i by defining a geometric similarity

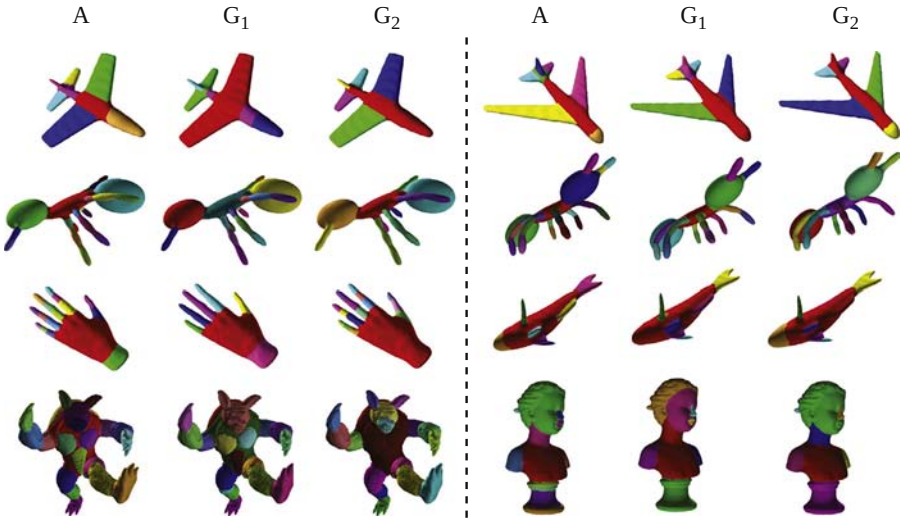

Fig. 1. One-to-one comparison between a perfect segmentation A generated via an algorithm and any ground-truth segmentation G_i leads to an unexpected error, because A is different from all the ground-truth segmentations, for example, G_1 and G_2.

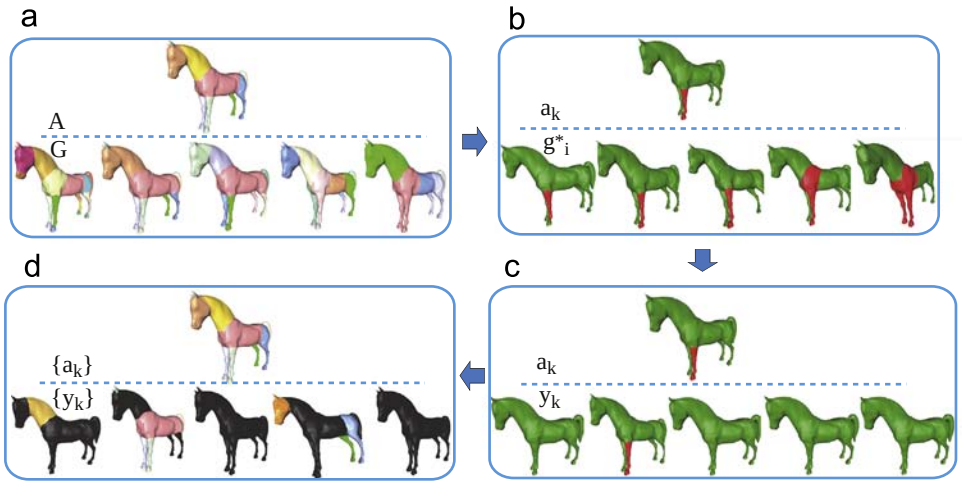

Fig. 2. Overview of the steps in SHD computation. (a) An automatic segmentation A and multiple ground-truth segmentations G_i. (b) Search corresponding parts (in red) of each segment a_k in ground-truth segmentations. (c) Obtain most similar part (in red) for the segment. (d) Find all the corresponding parts (in color not black) for segments of A. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this article.)

distance as follows:

$$SD = (1-\beta)*EMD(a_k, O_j(a_k)) + \beta*\tilde{d}(C(a_k), C(O_j(a_k))), \quad (2)$$

where EMD is the Earth Mover's distance of the D2 distribution [38] between a_k and each overlapping part $O_j(a_k)$ in G_i. The second term $\tilde{d}(C(a_k), C(O_j(a_k)))$ is the scaled Euclidean distance $d(C(a_k), C(O_j(a_k)))$ between the center $C(a_k)$ of a_k and the center $C(O_j(a_k))$ of $O_j(a_k)$. β is a weight. Because EMD lies in the range of [0,1], their Euclidean distance should be scaled to the range of [0,1] by introducing the following equation:

$$\tilde{d}(C(a_k), C(O_j(a_k))) = \frac{d(C(a_k), C(O_j(a_k)))}{\sum_{j=1}^{M} d(C(a_k), C(O_j(a_k)))}, \quad (3)$$

where M is the size of set $O(a_k)$. We use the above distance measure to find a desired part in the segmentation G_i, whose shape looks like the segment a_k and also its geometric location is very close to the segment a_k. The best corresponding part g_i^* with the smallest similarity distance is selected. β is the weight balancing the two terms. The method for determining β will be shown later.

Step 3: After choosing the parts $\{g_1^*, \ldots, g_n^*\}$ from different ground-truth segmentations $\{G_1, \ldots, G_n\}$, we re-compute and compare the similarity distances SD from each part g_i^* to the segment a_k using Eq. (2). The first term has been computed in the above step, and only the second term should be re-scaled by the summation of Euclidean distances between $C(a_k)$ and $C(g_i^*)$. We search an optimum part from $\{g_i^*\}$, which satisfies two conditions: (1) it has smaller similarity distance SD, and (2) its area should be also greater than half the area of a_k. If there are no parts satisfying $Area(g_i^*) > \frac{1}{2}Area(a_k)$, we will select the part with smallest similarity distance. We name the optimum part y_k corresponding to a_k.

Step 4: Repeat the above steps until all the corresponding parts $\{y_k\}$ to $\{a_k\}$ are found. $\{y_1, \ldots, y_C\}$ compose a new set Y, called similarity ground-truth segmentation. Now the number of elements in Y and A are the same, and each part a_k of A has a one-to-one geometric matching part y_k in Y. This solves the problem of different partition number of algorithms, and makes the corresponding relation more meaningful.

*Step*5: Calculate the Hamming distance between the automatic segmentation A and the similarity ground-truth segmentation Y as follows:

$$D_H(A, Y) = \tfrac{1}{2}(R_m(A, Y) + R_f(A, Y)). \quad (4)$$

Taking into account the particularity of Y, for $R_m(A, Y)$ and $R_f(A, Y)$, we make the following changes:

$$R_m(A, Y) = \frac{\sum_{k=1}^{C} |a_k \backslash y_k|}{\sum_{k=1}^{C} |a_k|}, \quad (5)$$

$$R_f(A, Y) = \frac{\sum_{k=1}^{C} |y_k \backslash a_k|}{\sum_{k=1}^{C} |y_k|}, \quad (6)$$

where "\" is the set difference operator, and "|.|" denotes the cardinality of a set. Here it is the total area of faces in the set. The normalization constant in the traditional Hamming distance [2] is $|S|$, which means that $|A| = |Y| = |S|$. Our algorithm introduces the Similarity Hamming Distance, which makes $|A| \neq |Y|$. Hence we changed the normalization coefficient. The normalized Hamming distance will take a value with a lower bound of 0 and an upper bound of 1, where 0 represents perfect segmentation, and 1 means no similarity between the automatic segmentation and the ground-truth segmentations.

Discussion: We discuss two possible questions about the SHD metric. The first question is why not to adopt the size of the overlapping area to find the corresponding segment in G_i for each segment a_k. The traditional Hamming distance requires that the corresponding relation is determined via the rule of "maximizing the overlapping area". However, this rule cannot be applied to "one to many" matching. Fig. 3(a) shows the reason. The left is an automatic segmentation A, and the right segmentations are ground-truth segmentations $\{G_i\}$. While sorting segmentations by the overlapping area, the whole body of G_1 is selected as the best corresponding part to the cup's body (in red) in A. In fact, the cup's body (in red) in segmentation G_2 has more similar geometry as the cup's body in A. Obviously, a comparison process without similarity information will cause an unreasonable correspondence. Accordingly, in order to avoid the unexpected correspondence, we introduced the concept of a similarity distance.

The second question is how to choose the weight β in Step 2, where we defined a similarity distance in Eq. (2) to consider the shape feature along with location information for segment correspondences. If only one of them is used, it will lead to an unexpected result. For instance, sometimes the D2 distance [38]

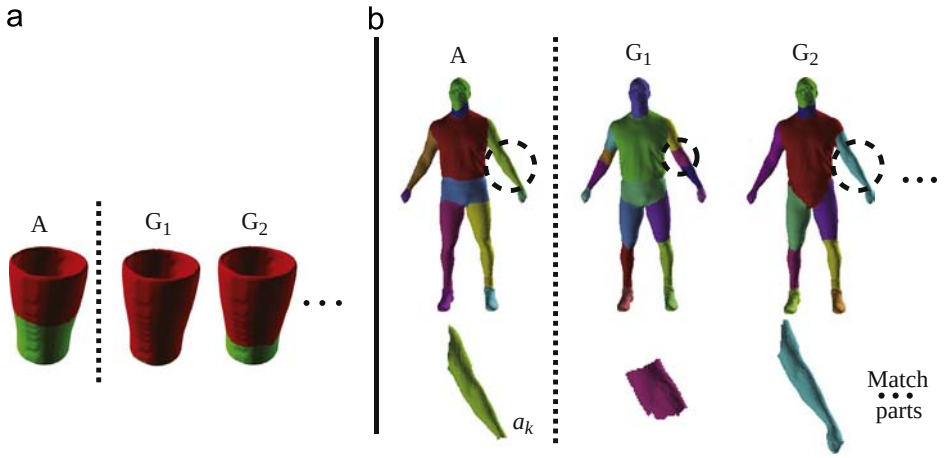

Fig. 3. (a) Illustration of the reason of why we do not adopt the size of overlap area to find the corresponding segment in G_i for each segment a_k. G_2 is a more consistent segmentation with A but its overlap area with A on the cup's body is less than that between A and G_1. (b) We show a case of wrong correspondence caused by only using location information while computing similarity distance. After sorting segmentations by the Euclidean distance between the center of the elbow part a_k and its corresponding parts in G_i, it is found that although a part of G_1 has the nearest distance to a_k, the arm of G_2 has a more similar geometry as a_k. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this article.)

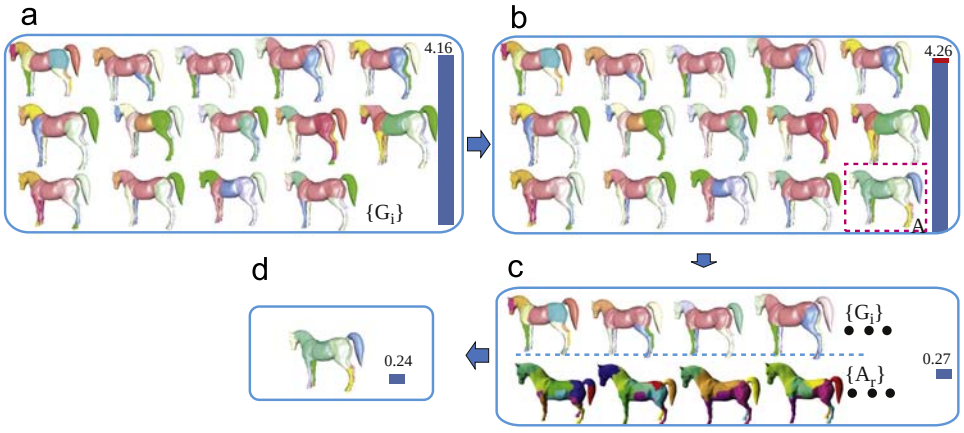

Fig. 4. Overview of the steps in AEI computation. (a) The entropy of multiple ground-truth segmentations $\{G_i\}$ forms a baseline (a blue bar with its value). (b) The entropy increases when an automatic segmentation A is added. The red bar is the incremental from the baseline, which is 0.1 in this example and then normalized by the upper bound of the entropy increment ($H(A)=1.64$). (c) The adaptive expectation of entropy increment $E(\Delta H)$ is computed by introducing many random segmentations $\{A_r\}$. (d) The final AEI is obtained after normalizing it via the adaptive expectation. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this article.)

cannot separate two very similar parts, such as the upper leg and lower leg. This can result in finding a wrong corresponding part if only D2 distance is used. Similarly, if we use the location information alone, Fig. 3(b) shows a case of wrong correspondence caused by only using location information. In order to integrate the two items to get the right correspondence result, the coefficient β plays the important role of balancing the two items. We observe that the Earth Mover's distance of the D2 distribution in the first term is often smaller than the Euclidean distance in the second term, and both of them lie in [0,1]. The algorithm automatically sets β by choosing one discrete value in the range of [0,1], in order to get the smallest Similarity Hamming Distance.

4. Entropy increment

Entropy is a concept commonly used in information theory, which measures the uncertainty associated with a random variable. The entropy of a system increases while the system status becomes more unordered. We consider diversity and disorder of different segmentations on the same model. If this type of diversity and disorder is modeled in the form of random variables, their entropy can be introduced to measure this type of diversity and disorder. The problem of measuring segmentation quality can be converted to entropy comparison. The entropy of all the different ground-truth segmentations forms a baseline. When a novel automatic segmentation generated via an algorithm is added, the entropy increases from the baseline. The amplitude of the entropy increment is adopted to evaluate the quality of an automatic segmentation. An overview of the steps in AEI computation is illustrated in Fig. 4.

Assume that an arbitrary segment in each ground-truth segmentation G_i is $s(G_i)$, and the probability of segment overlaps from different segmentations $\{G_1,\ldots,G_n\}$ is defined as:

$$P(\{s(G_1),\ldots,s(G_n)\}) = \frac{\|\{\forall f, f\in s(G_1)\}\bigcap\ldots\bigcap\{\forall f, f\in s(G_n)\}\|}{S}, \quad (7)$$

where S denotes the total area of the mesh, and $\|\cdot\|$ is the area size of a given subset of faces. In order to simplify the expression of probability distribution, we introduce the following definition to substitute it:

$$P(G_1,\ldots,G_n) = P(\{s(G_1),\ldots,s(G_n)\}). \quad (8)$$

Eqs. (7) and (8) mean that the joint distribution of $\{G_1,\ldots,G_n\}$ can be estimated by computing overlapped area among segments from different segmentations. Then, their entropy is based on all the possible segment combinations among different segmentations of the same model, which is defined as follows:

$$H(G_1,\ldots,G_n) = -\sum P(G_1,\ldots,G_n)\log(P(G_1,\ldots,G_n)). \quad (9)$$

It describes diversity and disorder of different segmentations on the same model. While computing the distribution of manual segmentations for the model, the possible combination statuses are limited, and the upper bound of status number is the number of faces in the model. We only consider these valid statuses, and therefore the computation complexity is linear. According to the concept of entropy, when a novel automatic segmentation A is added, the entropy will increase and the following inequality should be satisfied:

$$H(G_1,\ldots,G_n) \le H(G_1,\ldots,G_n,A). \quad (10)$$

The above equation implies that the inconsistent degree of segmentations will increase when adding a new segmentation A generated via an algorithm. Nevertheless, the entropy increment will be zero in the following three cases:

1. A contains no segments, which means the mesh is not segmented at all in the automatic segmentation.
2. A is the same as one of $\{G_1,\ldots,G_n\}$.
3. A consists of a combination of segments from $\{G_1,\ldots,G_n\}$.

In the third case, we assume that A is an automatic segmentation to be compared to two manual segmentations $\{G_1,G_2\}$ of a 3D human mesh. Fig. 5 shows this situation using three segmentations and their illustrative partitioned boxes. The blue part in the box describes the upper body of the segmentation A, and the red part corresponds with the legs of A. The blue part appears in the ground-truth segmentation G_1 because A and G_1 have the same segmentation in the upper body. The red part appears in the ground-truth segmentation G_2 because A and G_2 have the same segmentation in the lower body. A is fully consistent with two ground-truth segmentations, and accordingly we think A is a complete combination of G_1 and G_2. Different from previous metrics based on average of one-to-one comparisons, the entropy increment does not change when A is added as follows:

$$H(G_1,G_2) = H(G_1,G_2,A). \quad (11)$$

It shows that adding A will not increase the joint entropy, which means that a segmentation is perfect when it can be expressed as combination of subsets of standard segmentations. To better understand the idea, we give a metaphor that if a child's nose is like his father and the other parts look like his mother, we say he still "looks like his parents". We can use this concept to achieve the multi-criteria evaluation. In contrast, the entropy increases remarkably when A is not related to the set of $\{G_1,\ldots,G_n\}$. It means that A looks like none of the ground-truth segmentations $\{G_1,\ldots,G_n\}$. In this case, we can get the following equation:

$$H(G_1,\ldots,G_n,A) - H(G_1,\ldots,G_n) = H(A). \quad (12)$$

The above case is an extreme case, and also a useful cue that $H(A)$ can be adopted as the upper bound of entropy increment. In order to normalize the entropy increment to the range of [0,1], we use the upper bound $H(A)$ to scale the metric by the following equation:

$$\Delta H = \frac{H(G_1,\ldots,G_n,A) - H(G_1,\ldots,G_n) + \epsilon}{H(A)+\epsilon}. \quad (13)$$

where a very small constant ϵ (set to be the minimum of any floating-point number) is added in order to avoid a special case that an input mesh is not partitioned at all. In this case $H(A)$ is zero and the normalized entropy increment is 1.

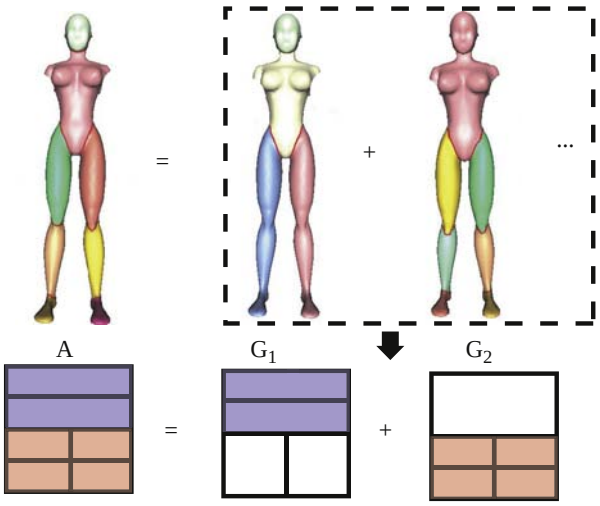

Fig. 5. If A can be completely represented by a combination of segments from G_1 and G_2, then the value of its entropy increment does not change. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this article.)

Additionally, we also tried another different normalization coefficient $\log m$, where m is the number of segments in A. The reason of adopting this normalization is that the maximum value of $H(A)$ is $\log m$. Although choosing $\log m$ as a scale factor can make the scale the same in different segmentations, it possibly leads to incorrect evaluation under the situation of unbalanced segmentation. We use an example to explain it. Fig. 6(a) shows segmentations of a water pot: an automatic segmentation A with very unbalanced segments (such as relatively very small and large segments), and a ground-truth segmentation G_1. We compute the error of the automatic segmentation using Eq. (14), and find that its value is close to 0. This means that the automatic segmentation is very similar to the ground-truth segmentation, while in fact it is not so.

$$\Delta H = \frac{H(G_1,A) - H(G_1) + \epsilon}{\log m + \epsilon} \approx 0. \quad (14)$$

If $H(A)$ is used as the normalization coefficient, the value of entropy increment in Eq. (15) is close to 1. It leads to the correct conclusion that the automatic segmentation is far from the ground-truth segmentation. Moreover, adopting $H(A)$ also makes the upper bound compact and enlarges the variation range of the entropy increment metric

$$\Delta H = \frac{H(G_1,A) - H(G_1) + \epsilon}{H(A) + \epsilon} \approx 1. \quad (15)$$

5. Adaptive entropy increment

We observe that the variation range and discriminative power of entropy increment cannot satisfy the requirement of segmentation evaluation. The values of entropy increment are different for the segmentations of simple and relatively complex models. Fig. 6(b) illustrates two examples, one of which is a simple cup model and another is a relevantly complex horse model. We design an algorithm to randomly partition the two models eleven times, and the number of segments in each random segmentation is consistent with the automatic segmentation to be evaluated. The entropy increment value of each random segmentation against the ground-truth segmentations $\{G_i\}$ of the same model is

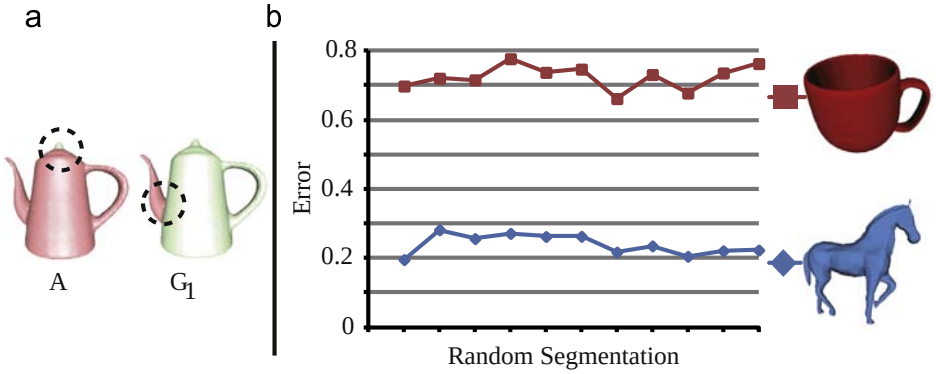

Fig. 6. (a) We use an extreme example of an automatic segmentation A and a ground-truth segmentation G_1 to observe the effects of adopting two scales $\log m$ and $H(A)$. (b) The expectation of entropy increments in segmentations of a simple model "cup" and a complex model "horse". The horizontal axis is the i-th different random segmentations of each model, and the vertical axis is the corresponding entropy increment.

computed. We use manual segmentations in the Princeton dataset as ground-truths. Therefore, we obtain eleven error values for both the cup model and the horse model. The entropy increment of the simple model varies around the center of a large value, while the entropy increment of the complex model varies centered around a small value.

Therefore, we attempt to find an *adaptive expectation* to scale the value of entropy increment for each model and specified segment number, which can enhance the discriminative power of entropy increment. If an automatic segmentation with zero ΔH is viewed as the best segmentation, the random segmentation without any a priori knowledge such as geometric features should be the worst segmentation and its entropy increment should be scaled to 1 by its adaptive expectation. A remaining problem is how to define the expected entropy increment. We estimate it from N random segmentations $\{A_r\}$ in the following equation:

$$E(\Delta H) = \frac{1}{N}\sum_{r=1}^{N} \Delta H(G_1, \ldots, G_n, A_r), \tag{16}$$

where A_r is a random segmentation with same segment number as the automatic segmentation to be evaluated, and $\{\{G_i\}\}$ is the set of ground-truth segmentations for the same 3D mesh. Given a specified segment number, we use random region growing to generate those N random segmentations, which can be seen as the worst cases. We use this way to obtain the expected entropy increment $E(\Delta H)$.

After obtaining $E(\Delta H)$, the *adaptive entropy increment* (AEI) metric is defined as follows:

$$\Delta H_a = \frac{\Delta H}{E(\Delta H)} \tag{17}$$

6. Experiments

In order to investigate the utility of the two proposed metrics for evaluating segmentations, we performed five experiments. The first three experiments are to investigate their discrimination capability in three aspects: (1) standard segmentations and random segmentations, (2) extreme segmentations, (3) segmentations of complex models and simple models. The next study evaluates whether our metrics are robust against hierarchical segmentations. We finally generate the errors of two proposed metrics for manual segmentations in the Princeton segmentation dataset and automatic segmentations generated by representative algorithms.

Protocol: Here we first describe the protocol adopted in the experiment to evaluate a given segmentation. There are two cases according to the origin of the segmentations. While evaluating one manual segmentation, the remaining manual segmentations of the same model are treated as ground-truths. In the case of evaluating one automatic segmentation, the segmentation generated by any state-of-the-art algorithm or a random algorithm designed only for test is compared to all the ground-truth segmentations of the same model. Moreover, the evaluation process is different for each metric when computing the segmentation error. In order to measure the error of one segmentation using one of the CD, HD, RI, and CE metrics, the segmentation is first compared to each ground-truth segmentation to get one error value, and the resulting values are then averaged within all the ground-truths of the same model. In the case of the SHD and AEI metrics, only one value is generated by means of comparing this segmentation to multiple ground-truth segmentations of the same model.

6.1. Discrimination capability on standard segmentations and random segmentations

We first investigate whether the two proposed metrics improve the discrimination capability between standard segmentations and random segmentations. We continue to use the 4300 manual segmentations in the Princeton segmentation dataset, and also generate 4300 new random segmentations. The segment number of each random segmentation is set to be the same as its corresponding manual segmentation. We adopt the above protocol to generate a metric score for each segmentation, and make a histogram of scores of all the segmentations. We compute six metric scores for all the manual segmentations and all the random segmentations, and finally produce six histograms of metric scores.

Fig. 7 illustrates the statistics of the discriminative power on standard segmentations and random segmentations for four previous metrics, and two proposed metrics respectively. The vertical axis represents the number of segmentations with scores located in the horizontal bin. It is expected that manual segmentations have lower errors, and random segmentations have higher errors. Therefore, the most desirable status is that statistical bars counting manual segmentations are grouped to the left, and the bars counting random segmentations are grouped to the right. We see that both SHD and AEI can spatially separate two different types of segmentations in the histograms, and especially AEI generates a satisfactory discrimination result.

We obtain a quantitative difference between two different histograms H_1 and H_2 for each metric as follows:

$$D(H_1, H_2) = \frac{\mu_1 - \mu_2}{\sigma_1 + \sigma_2}, \tag{18}$$

where μ denotes the mean and σ is standard deviation. The difference values of CD, HD, RI, CE, SHD, and AEI are 0.6, 1.4, 0.7, 1.5, 1.9, and

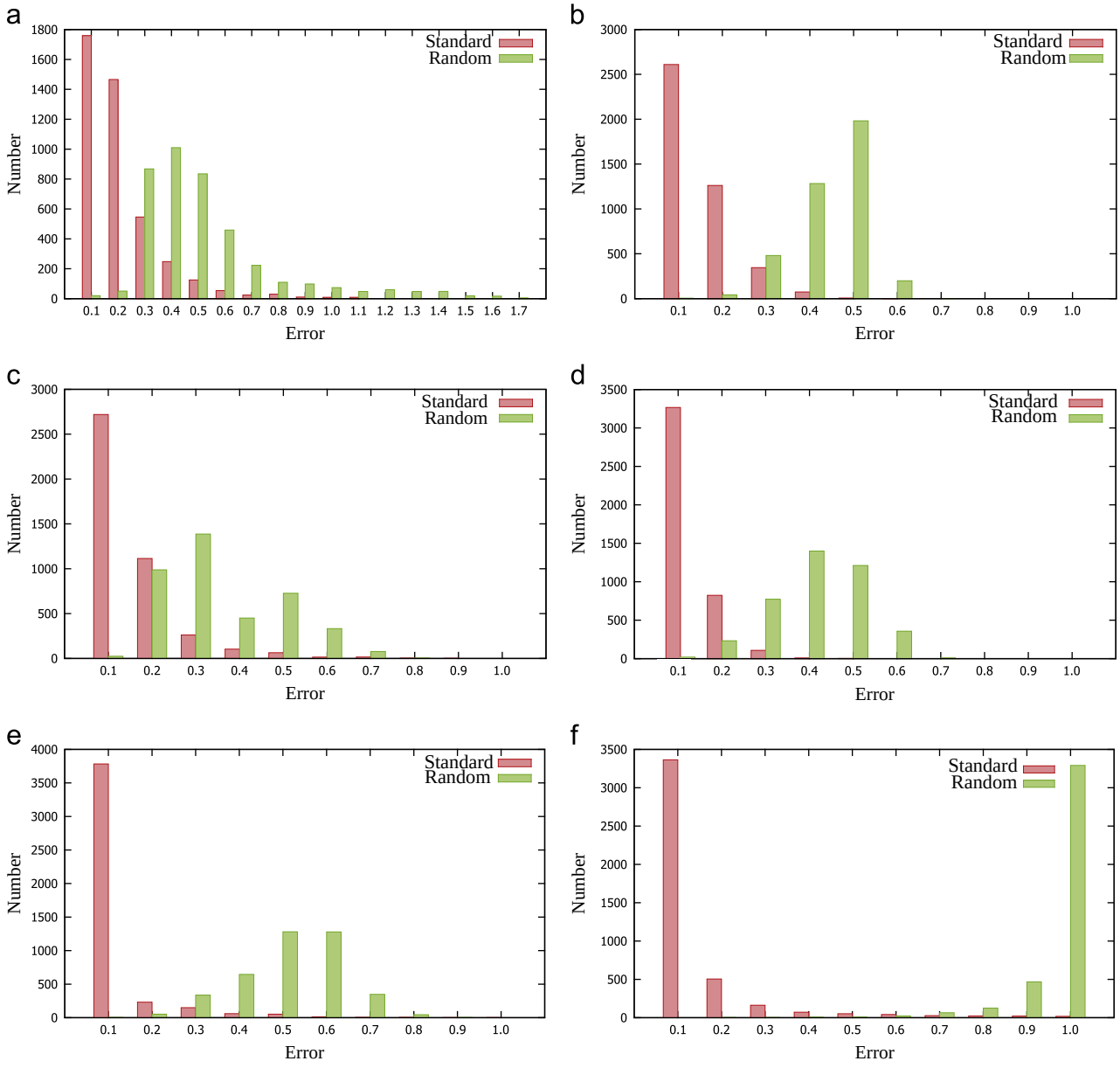

Fig. 7. Discriminative power of six metrics (a)–(f) on standard (manual) segmentations and random segmentations. The vertical axis represents the number of segmentations with scores located in the horizontal bin. (a) CD, (b) HD, (c) RI, (d) CE, (e) SHD and (f) AEI.

5.0 respectively. SHD and AEI have relatively larger difference values between standard and random segmentations, which proves that they have higher discriminative power.

We also compare two histograms of standard and random segmentations for each metric using overlapping bars. Specifically, the function of overlap comparison $O(H_1, H_2)$ is defined as follows:

$$O(H_1, H_2) = \int \min(h_1(e), h_2(e))\, de. \tag{19}$$

where e denotes a metric error and $h(e)$ is the corresponding probability density of error. The function values of CD, HD, RI, CE, SHD, and AEI are 26.0%, 11.1%, 34.3%, 8.5%, 7.9%, and 3.1% respectively. These numerical values show that SHD and AEI have smaller overlap between standard and random segmentations, and achieve relatively higher discrimination.

6.2. Discrimination capability on extreme segmentations

A group of extreme segmentations including unreasonable segmentation, over-segmentation, relatively perfect segmentation, and under-segmentation is used to test quantitative responses of four metrics from the Princeton segmentation benchmark, and the two proposed metrics. Fig. 8 shows the four types of segmentations, and the error values of the six metrics. Unreasonable segmentation and over-segmentation have larger SHD and AEI errors, and under-segmentation has largest AEI error. The values of SHD and AEI on the perfect segmentation are both small. For each metric, if the error value of the perfect segmentation is subtracted from the error value of a bad segmentation (unreasonable segmentation, over-segmentation, and under-segmentation), we see that AEI generates a higher difference than the other metrics. The remarkable difference is desirable while evaluating and comparing segmentation algorithms. Moreover, we see the values of CE on

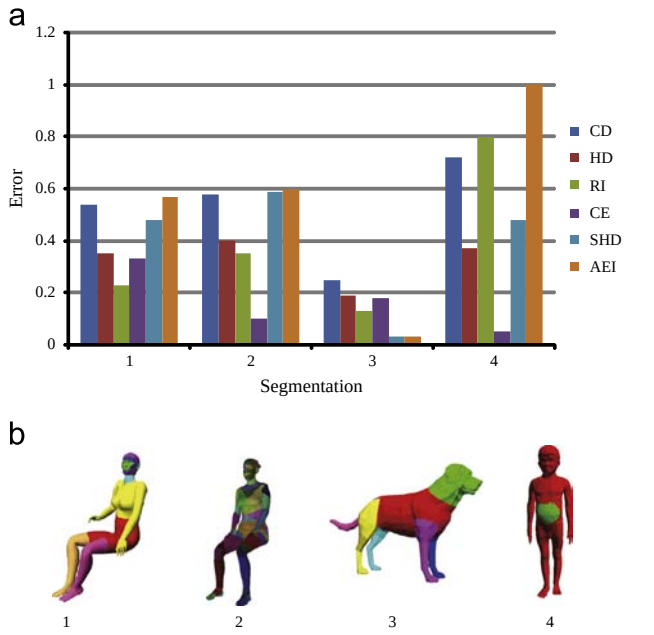

Fig. 8. (a) Errors of six metrics on (b) four extreme segmentation examples, which are unreasonable segmentation, over-segmentation, relatively perfect segmentation, and under-segmentation.

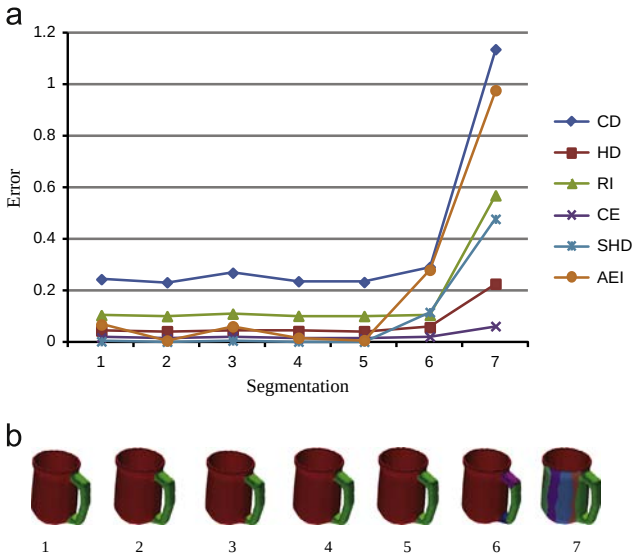

Fig. 9. Comparison of discriminative power on simple models. We use (a) six metrics to evaluate, (b) seven segmentations of a cup, where the last two segmentations are clearly inconsistent with human perception.

the four segmentations. The over-segmentation and under-segmentation obtain lower error than the perfect segmentation, which is an unsatisfactory evaluation of the segmentation.

6.3. *Discrimination capability on complex models and simple models*

We perform two groups of experiments concerning discriminative power on not only simple models but also complex models. Simple models commonly consist of a small number of segments, and complex models consist of relatively many segments. The desirable discriminative power should appear not only on several segmentations of a simple model, but also on segmentations of a complex model. The six metrics including CD, HD, RI, CE and our two metrics SHD and AEI, are adopted to evaluate seven segmentations of a simple cup selected from the Princeton segmentation

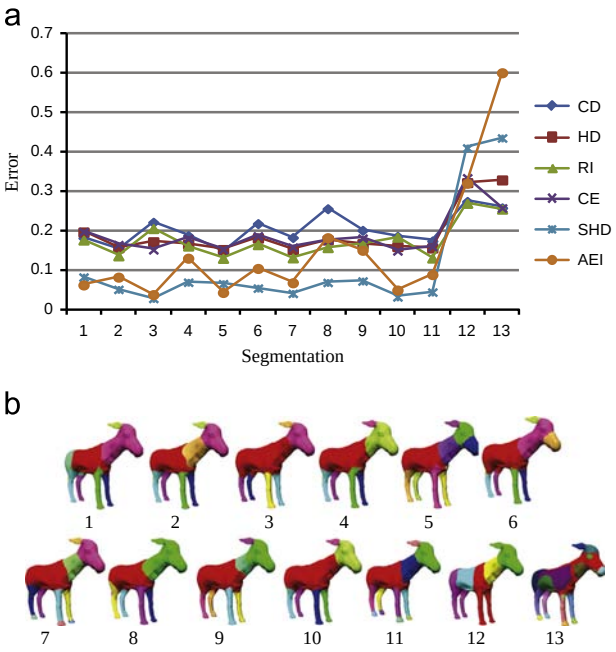

Fig. 10. Comparison of discriminative power on complex models. We use (a) six metrics to evaluate (b) thirteen segmentations of a four-leg animal, where the last two segmentations are clearly inconsistent with human perception.

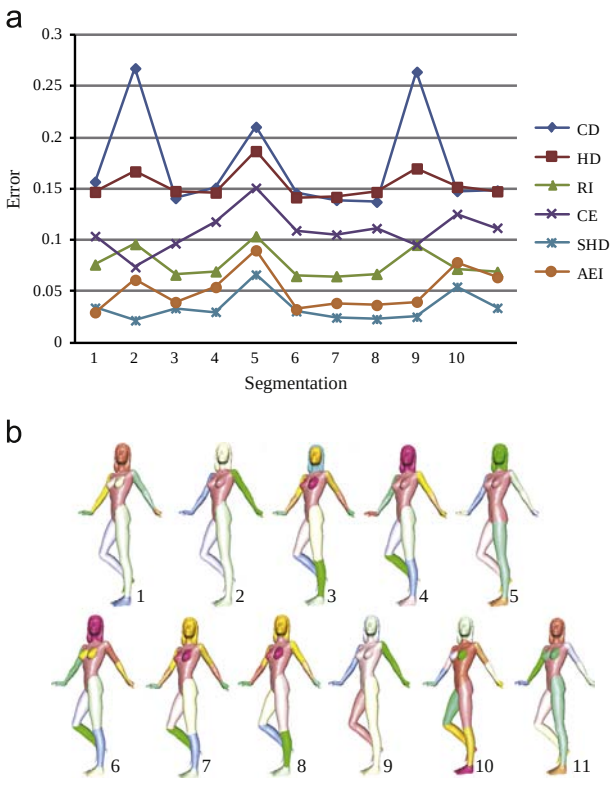

Fig. 11. Tolerance of metrics to hierarchical segmentations. (a) Segmentation errors of various metrics, (b) eleven segmentations of woman model.

dataset, as shown in Fig. 9. The first five segmentations are reasonable and they are slightly different on the boundary of the handle, while the last two segmentations are clearly inconsistent with human perception. SHD and AEI generate lower errors for five good segmentations. CD and AEI generate higher errors on the last two segmentations, especially the last one, while CD and RI have higher errors on five good segmentations. When

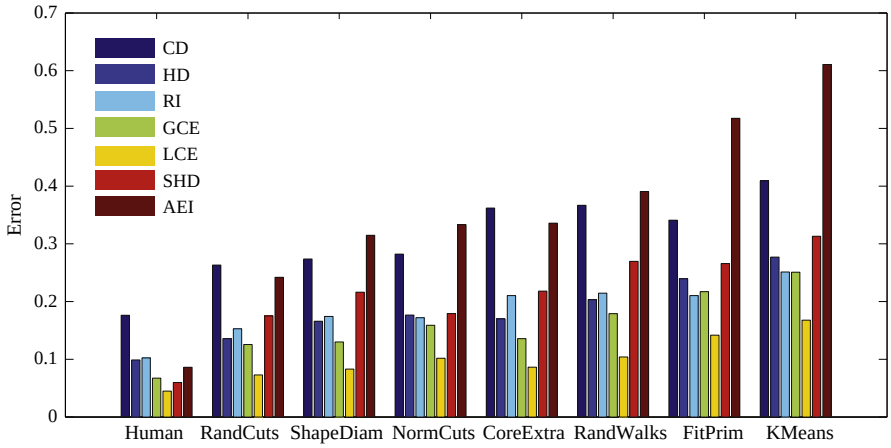

Fig. 12. The errors of SHD, AEI, and other four metrics on segmentations produced by human and algorithms in Princeton dataset.

segmentations become worse from the first five segmentations to the last two segmentations, SHD and AEI give quick response and errors rise significantly. In Fig. 10, we also compute the six metrics on thirteen segmentations of a complex model, where the first eleven segmentations are satisfactory and the last two segmentations are unreasonable or an over-segmentation. As can be seen, AEI and SHD have lower errors on good segmentations, and they rise quickly for the last two segmentations. The two examples show that AEI and SHD have higher discriminative power.

6.4. Robustness against hierarchical segmentations

Next we show an experiment to examine whether our metrics are sensitive to hierarchical segmentations. Fig. 11 shows eleven segmentations of the same model with different refinements. The vertical axis represents the score of each metric, while the horizontal axis indicates the segmentation number. Compared to the first four metrics, the error values of SHD and AEI are relatively stable and change little. Hence, for segmentations with different hierarchical structures, the two proposed metrics can tolerate these refinements and generate consistent results.

6.5. Evaluation on the Princeton segmentation dataset

We finally investigate the effectiveness of adopting the proposed metrics to evaluate the Princeton segmentation dataset and seven representative algorithms, which include randomized cuts (RC), shape diameter (SD), normalized cuts (NC), core extraction (CE), random walks (RW), fitting primitives (FP), and K-means (KM). The 4300 manually generated segmentations of the 380 3D models in the Princeton segmentation benchmark [2] are selected as our ground-truth segmentations.

Human-generated segmentation evaluation: We first compute the proposed metrics, SHD and AEI, on these manual segmentations to measure their errors. As mentioned in the above protocol, the error of each manual segmentation is obtained by viewing the remaining segmentations as ground-truths and comparing against them. Next, we average the errors of all the manual segmentations first within each model, then those results are averaged within each category, and finally a mean error value is provided for the entire dataset for each of the proposed metrics. The same protocol is adopted for each of the previous four metrics, CD, HD, RI, and CE, to report a total mean error for all the segmentations of models generated by humans.

Algorithm-generated segmentation evaluation: Similarly, we study properties of segmentations generated by seven state-of-the-art algorithms and use the two proposed metrics to compute a total mean error for each algorithm. It should be noted that each algorithm produces one segmentation for each model in our experiment, and the above protocol is adopted to compute the error of this segmentation. Five algorithms such as RC, NC, RW, FP, and KM, require the target number of segments as an input parameter. Similarly as the setting in [2], we set it to be the mode (most frequent) of the number of segments appearing in segmentations created by people for that model in the Princeton dataset. SD and CE automatically select the optimal segment number for each model. Similarly, for each of the previous four metrics, a total mean error is obtained by averaging all the errors over the entire segmentation dataset generated by any automatic algorithm.

Table 2
Subjective scores (SS) of seven algorithms, and their subjective rank (SR). Their ranks via CD, HD, RI, CE(GCE,LCE), SHD, and AEI are also given.

Metric	RC	SD	NC	CE	RW	FP	KM
SS	6.0	5.7	5.5	5.2	4.8	4.7	3.9
SR	1	2	3	4	5	6	7
AEI	1	2	3	4	5	6	7
SHD	1	3	2	4	6	5	7
GCE	1	2	4	3	5	6	7
LCE	1	2	4	3	5	6	7
RI	1	3	2	4	6	5	7
HD	1	2	4	3	5	6	7
CD	1	2	3	5	6	4	7

Fig. 12 shows the errors of the two proposed metrics, compared to the other four metrics. For all the six metrics, lower bars represent better segmentation results. Among the six metrics, we see that AEI and SHD achieved relatively larger difference range between manual segmentation and automatic segmentation generated by algorithms, for example, human and K-means. According to the RI metric, the maximum difference of scores between different algorithms is only 0.15. The dynamic ranges are 0.25 and 0.52 for SHD and AEI respectively. They have potential to effectively differentiate the ability of segmentation algorithms. We also rank these algorithms via the two proposed metrics and the four previous metrics in Table 2, and next discuss a user study on these ranks.

User study: To demonstrate the effectiveness of algorithm comparison using the two proposed metrics, we performed a study where we asked 10 participants to subjectively estimate the segmentation quality of 380 models generated by each algorithm, and then sort these seven algorithms via mean subjective scores. Before the test, we asked these participants to observe each standard segmentation in the Princeton dataset, and also each

random segmentation generated in the previous experiment. Next, we provided all the segmentations of 380 models generated by these seven algorithms to each participant. Each participant graded automatic segmentations and was required to compare corresponding standard segmentations at the same time, where the range of given scores was from 0 to 10. According to scores given by each participant, we took the average over models in each category, and then averaged the scores over all the categories. Accordingly, we obtained seven scores for these algorithms from each participant, and took the average for each algorithm over 10 participants. The mean scores and rank of seven algorithms are shown in Table 2. We give ranks provided by the metrics. We can see that the rank provided by AEI is fully consistent with the subjective rank, while SHD and other metrics are slightly different from the subjective rank. SHD and RI are fully consistent on the method rank. We also compute the mean correlation value between subjective scores and the values of each metric. The values of AEI and SHD are 0.55 and 0.48, and CD, HD, RI, and GCE (LCE) are 0.18, 0.47, 0.38, and 0.41(0.37) respectively. It can be shown that AEI and SHD have relatively higher consistency with the subjective evaluation. Finally, it should also be noted that the subjective evaluation contains many uncertainty factors, for example, the scale that different participants assign to good or bad segmentations.

7. Conclusion

This paper describes two metrics supporting evaluation on multiple standard segmentations, which are Similarity Hamming Distance (SHD) and Adaptive Entropy Increment (AEI). SHD is based on partial similarity correspondences between an automatic segmentation and ground-truth segmentations, and AEI measures diversity and disorder of segmentations when an automatic segmentation is added to a set of ground truth segmentations. A group of experiments shows that the new metrics obtain higher discrimination on different types of segmentations and models. We expect that they will be adopted to evaluate the development of algorithms in the future.

Discussion and limitations: The mean computation time of evaluating an automatic segmentation via SHD and AEI is 5.5 s and 0.3 s respectively, executing in an Intel i3 3.3 GHz computer with 8 GB memory. We find that SHD costs much time on the computation of the D2 distributions and parameter selection, although we set a search step 0.01 for the balancing parameter β. It is a limiting factor to practical use. Another limitation is that we only search for corresponding parts of an automatic segmentation, and correlation between segments and contextual relations is not considered in SHD. Moreover, we first compute two terms in the similarity distance, and then find the β by searching its approximate discrete values in order to get the smallest distance. It is a trade-off solution however this can bias the metric and possibly affects the accuracy of SHD. In addition, AEI is based on combinations of small entropies, where summing several small entropies would equal to one big entropy change. In this case, a segmentation that locally is very different would have the same distance as a segmentation which has many small differences. This will lead to a biased evaluation.

Future work: We would like to focus on the evaluation of co-segmentations of shape sets. Co-segmentation has been a hot topic in the segmentation field, and a large dataset was just recently introduced [39], although there is still a lack of metrics designed specifically to evaluate co-segmentation results. Also of interest is to label and evaluate 3D scene segmentation because 3D scene datasets like Sketchup scenes and depth map scenes captured by Kinect are growing rapidly. Besides the automatic segmentation generated by algorithms, segmentations are obtained by interactive techniques in many applications. It is interesting to investigate how the proposed metrics reveal the user interaction during the shape segmentation and evaluate interactive segmentation in real time. First, a dataset composed of standard interactive segmentations should be built, and these standard segmentations should be finished by skilled people. The proposed metrics score online operations via comparing real-time interactive segmentation to standard segmentations. This would assist users to produce satisfactory segmentation results.

Acknowledgments

We would first like to thank the anonymous reviewers for their valuable feedback. We would also like to thank Xiaobai Chen et al. for providing the Princeton segmentation benchmark. Thanks also go to Oliver van Kaick and Honghua Li for fruitful discussions on the paper. This work is supported partly by grants from NSFC (61003137, 61202185), NWPU Basic Research Fund (JC201202, JC201220), Shaanxi Natural Science Fund (2012JQ8037), and Open Fund from State Key Lab of CAD&CG of Zhejiang University.

References

[1] Benhabiles H, Vandeborre J-P, Lavoué G, Daoudi M, A framework for the objective evaluation of segmentation algorithms using a ground-truth of human segmented 3D-models. In: Proceedings of shape modeling international; 2009. p. 36–43.
[2] Chen X, Golovinskiy A, Funkhouser T. A benchmark for 3D mesh segmentation. ACM Trans Graph 2009;28(3) Article no. 73.
[3] Skraba P, Ovsjanikov M, Chazal F, Guibas L. Persistence-based segmentation of deformable shapes. In: IEEE conference on computer vision and pattern recognition workshops; 2010. p. 45–52.
[4] Kalogerakis E, Hertzmann A, Singh K. Learning 3D mesh segmentation and labeling. ACM Trans Graph 2010;29(3):1–11.
[5] Zheng Y, Tai C-L, Au OK-C. Dot scissor: a single-click interface for mesh segmentation. IEEE Trans Vis Comput Graph 2012;18(8):1304–12.
[6] Huang Q, Koltun V, Guibas L. Joint shape segmentation with linear programming. ACM Trans Graph 2011;30(6) Article 125.
[7] Benhabiles H, Lavoué G, Vandeborre J-P, Daoudi M. Learning boundary edges for 3D-mesh segmentation. Comput Graph Forum 2011;30(8):2170–82.
[8] Bergamasco F, Albarelli A, Torsello A. Semi-supervised segmentation of 3D surfaces using a weighted graph representation. In: Proceedings of international conference on graph-based representations in pattern recognition; 2011. p. 225–34.
[9] Meng M, Fan L, Liu L. A comparative evaluation of foreground/background sketch-based mesh segmentation algorithms. Comput Graph 2011;35(3):650–60.
[10] Benjamin W, Polk AW, Vishwanathan S, Ramani K. Heat walk: robust salient segmentation of non-rigid shapes. Comput Graph Forum 2011;30(7):2097–106.
[11] Meng M, Fan L, Liu L. icutter: a direct cut out tool for 3D shapes. Comput Anim Virtual Worlds 2011;22(4):335–42.
[12] Au OK-C, Zheng Y, Chen M, Xu P, Tai C-L. Mesh segmentation with concavity-aware fields. IEEE Trans Vis Comput Graph 2012;18(7):1125–34.
[13] Solomon J, Ben-Chen M, Butscher A, Guibas L. Discovery of intrinsic primitives on triangle meshes. Comput Graph Forum 2011;30(2):365–74.
[14] Ho T-C, Chuang J-H. Volume based mesh segmentation. J Inf Sci Eng 2012;28(4):705–22.
[15] Zhang J, Zheng J, Wu C, Cai J. Variational mesh decomposition. ACM Trans Graph 2012;31(3) Article 21.
[16] Shamir A. A survey on mesh segmentation techniques. Comput Graph Forum 2008;27(6):1539–56.
[17] de Goes F, Goldenstein S, Velho L. A hierarchical segmentation of articulated bodies. Comput Graph Forum 2010;29(5):1349–56.
[18] Shapira L, Shamir A, Cohen-Or D. Consistent mesh partitioning and skeletonisation using the shape diameter function. Vis Comput 2008;24(4):249–59.
[19] Sidi O, van Kaick O, Kleiman Y, Zhang H, Cohen-Or D. Unsupervised co-segmentation of a set of shapes via descriptor-space spectral clustering. ACM Trans Graph 2011;30(6) Article 126.
[20] Hu R, Pan L, Liu L. Co-segmentation of 3D shapes via subspace clustering. Comput Graph Forum (Proc SGP) 2012;31(5):1703–13.
[21] Meng M, Xia J, Luo J, He Y. Unsupervised co-segmentation for 3D shapes using iterative multi-label optimization. Comput Aided Des 2013;45(2):312–20.
[22] Lv J, Chen X, Huang J, Bao H. Semi-supervised mesh segmentation and labeling. Comput Graph Forum 2012;31(7):2241–7.
[23] Zhang J, Zheng J, Cai J. Interactive mesh cutting using constrained random walks. IEEE Trans Vis Comput Graph 2011;17(3):357–67.

[24] Xiao C, Fu H, Tai C-L. Hierarchical aggregation for efficient shape extraction. Vis Comput 2009;25(3):267–78.
[25] Sharf A, Blumenkrants M, Shamir A, Cohen-Or D. Snappaste: an interactive technique for easy mesh composition. Vis Comput 2006;22(9–11):835–44.
[26] Brown S, Morse B, Barrett W. Interactive part selection for mesh and point models using hierarchical graph-cut partitioning. In: Proceedings of graphics interface; 2009. p. 23–30.
[27] Zheng Y, Tai C-L. Mesh decomposition with cross-boundary brushes. Comput Graph Forum 2010;29(2):527–35.
[28] Zhang J, Wu C, Cai J, Zheng J, Tai X cheng. Mesh snapping: robust interactive mesh cutting using fast geodesic curvature flow. Comput Graph Forum 2010;29(2):517–26.
[29] Wang Y, Asafi S, van Kaick O, Zhang H, Cohen-Or D, Chen B. Active co-analysis of a set of shapes. ACM Trans Graph 2012;31(6) Article 165.
[30] Attene M, Katz S, Mortara M, Patanè G, Spagnuolo M, Tal A. Mesh segmentation-a comparative study. In: Proceedings of shape modeling international; 2006. p. 14–25.
[31] Shlafman S, Tal A, Katz S. Metamorphosis of polyhedral surfaces using decomposition. Comput Graph Forum 2002;21(3):219–28.
[32] Attene M, Falcidieno B, Spagnuolo M. Hierarchical mesh segmentation based on fitting primitives. Vis Comput 2006;22(3):181–93.
[33] Katz S, Leifman G, Tal A. Mesh segmentation using feature point and core extraction. Vis Comput 2005;21(8):649–58.
[34] Lai Y, Hu S, Martin RR, Rosin PL. Fast mesh segmentation using random walks. In: Proceedings of ACM symposium on solid and physical modeling; 2008. p. 183–91.
[35] Golovinskiy A, Funkhouser T. Randomized cuts for 3D mesh analysis, ACM Trans Graph 27 (5).
[36] Benhabiles H, Vandeborre J-P, Lavoué G, Daoudi M. A comparative study of existing metrics for 3D-mesh segmentation evaluation. Vis Comput 2010;26 (12):1451–66.
[37] Huang Q, Dom B. Quantitative methods of evaluating image segmentation. In: Proceedings of IEEE international conference on image processing; 1995. p. 53–6.
[38] Osada R, Funkhouser T, Chazelle B, Dobkin D. Shape distributions. ACM Trans Graph 2002;21(4):807–32.
[39] ⟨http://web.siat.ac.cn/∼yunhai/ssl/ssd.htm⟩.