HAL
open science

A comparative study of existing metrics for 3D-mesh segmentation evaluation

Halim Benhabiles, Jean-Philippe Vandeborre, Guillaume Lavoué, Mohamed Daoudi

To cite this version:

Halim Benhabiles, Jean-Philippe Vandeborre, Guillaume Lavoué, Mohamed Daoudi. A comparative study of existing metrics for 3D-mesh segmentation evaluation. The Visual Computer, 2010, 26 (12), pp.1451-1466. 10.1007/s00371-010-0494-2 . hal-00660825

HAL Id: hal-00660825

https://hal.science/hal-00660825v1

Submitted on 17 Jan 2012

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

The Visual Computer manuscript No.
(will be inserted by the editor)

A comparative study of existing metrics for 3D-mesh segmentation evaluation

Halim Benhabiles · Jean-Philippe Vandeborre · Guillaume Lavoué · Mohamed Daoudi

the date of receipt and acceptance should be inserted later

Abstract In this paper, we present an extensive experimental comparison of existing similarity metrics addressing the quality assessment problem of mesh segmentation. We introduce a new metric named the 3D Normalized Probabilistic Rand Index (3D-NPRI) which outperforms the others in terms of properties and discriminative power. This comparative study includes a subjective experiment with human observers and is based on a corpus of manually segmented models. This corpus is an improved version of our previous one [4]. It is composed of a set of 3D-mesh models grouped in different classes associated with several manual ground-truth segmentations. Finally the 3D-NPRI is applied to evaluate six recent segmentation algorithms using our corpus and the Chen's et al. [7] corpus.

Keywords 3D-mesh segmentation · ground-truth · similarity metric · subjective tests · evaluation

1 Introduction

3D-mesh segmentation is a fundamental process in many applications such as shape retrieval [1,29], compression [29], deformation [18], texture mapping [26], etc. It consists in decomposing a polygonal surface into different regions (i.e. connected set of vertices or facets) of uniform properties, either from a *geometric* point of view or from a *semantic* point of view. It is a critical step toward content analysis and mesh understanding. Although some supervised methods exist [12,16], most existing techniques are fully automatic.

According to recent states-of-the-art [3,24], mesh segmentation techniques can be classified into two categories: surface-type (or *geometric*) methods and part-type (or *semantic*) methods. In the first case, the algorithms are based on low level geometric information (e.g. curvature [20]) in order to define segments (i.e. regions) with respect to geometric homogeneity, while in the latter case, the algorithms aim at distinguishing segments that correspond to relevant features of the shape, by following higher level notions such as defined in human perception theory [6]. This kind of approach is particularly suited for object animation / deformation and indexing applications, where the decomposition has to be meaningful.

Although development of mesh segmentation algorithms for both approaches has drawn extensive and consistent attention, relatively little research has been done on segmentation *evaluation*. For the first approach (surface-type), some tools exist depending on the end application as texture mapping [23] or medical imaging [13]. Recently, two main works, Benhabiles et al. [4] (our previous work) and Chen et al. [7], have been proposed to study the quality assessment problem of part-type 3D-mesh segmentation. Both works propose a benchmark for segmentation evaluation which is based on a ground-truth corpus. The corpus is composed of a set of 3D-models grouped in different classes and associated with several manual segmentations produced by human observers. These two benchmarks comprise

H. Benhabiles
LIFL (UMR USTL/CNRS 8022), University of Lille, France
Tel.: +33-20-335517
Fax: +33-20-335599
E-mail: halim.benhabiles@lifl.fr

J-P. Vandeborre · M. Daoudi
Institut TELECOM ; TELECOM Lille 1, France
LIFL (UMR USTL/CNRS 8022), University of Lille, France
E-mail: {jean-philippe.vandeborre, mohamed.daoudi}@lifl.fr

G. Lavoué
Université de Lyon, CNRS, INSA-Lyon, LIRIS, UMR5205, F-69621, France
E-mail: glavoue@liris.cnrs.fr

the ground-truth corpus and a set of similarity metrics, then the evaluation of a segmentation algorithm consists in measuring the similarity between the reference segmentations from the corpus and that obtained by this algorithm (on the same models). In this kind of benchmark the quality of the evaluation depends on the quality of the corpus but also on the quality of the segmentation similarity measure. This leads to conclude that the choice of an accurate measure is quite critical in order to provide a strict evaluation and to reflect the real quality of an automatic segmentation with comparison to a manual one. In this context, less efforts were investigated to propose a reliable measure of mesh segmentation similarity. Indeed, the previous works [4, 7] focused their interests on the design of the ground-truth corpus and presented rather simple metrics suffering from degeneracies and low discriminative power.

In this context the objective of the present work is to evaluate the existing metrics and to propose a new one which is more reliable. This paper introduces three main contributions. Firstly, we propose a thorough study and comparisons of existing metrics addressing the assessment problem of mesh segmentation, using a corpus of manually segmented models. This corpus is an improved version of our previous one [4] and is available on-line[1]. Secondly, we propose a new measure of segmentation similarity that allows to quantify the consistency between multiple segmentations of a model. We show that this new metric outperforms existing ones in terms of properties and discriminative power. To quantitatively compare the discriminative power of the metrics, we have conducted subjective tests using a set of human observers. Thirdly, we apply this measure together with two corpuses (our corpus and Chen's et al. [7] corpus) to evaluate six recent 3D-mesh segmentation algorithms.

This paper is organized as follows. In section 2, we provide a review of the state-of-the-art of segmentation evaluation and an analytic study of the measures that have been proposed in this context. In section 3 and 4, we define a new objective metric to perform a quantitative comparison between a segmentation algorithm and a set of ground-truth segmentations (of the same model). In section 5, we present our corpus which will be used for the experimental comparison of the metrics and for the evaluation of the segmentation algorithms. In section 6, we present an extensive experimental comparison between our new metric and existing ones, then we analyze the discriminative power of this new metric using subjective tests. In section 7, we demonstrate the usability of our whole evaluation protocol through the evaluation of six recent segmentation methods. Section 8 concludes the paper.

[1] `http://www-rech.telecom-lille1.eu/3dsegbenchmark/`

2 Related Work

In this section, we firstly provide a review of the state-of-the-art of 2D-image and 3D-mesh segmentation evaluation. Indeed, the most significant works for the 3D-mesh segmentation evaluation [4, 7] are based on the same methodology as that proposed in the 2D-image domain [21]. Secondly, we review the measures that have been proposed in the context of 3D-mesh segmentation evaluation, while analyzing their properties.

2.1 State-of-the-art of 2D-image and 3D-mesh segmentation evaluation

Several advanced works exist for the quality assessment of *2D-image* segmentation. Zhang et al. [30] offer a study on the different methods proposed for this task. According to them, the different methods can be classified into five groups:

- **Analytical methods.** They directly treat the segmentation algorithms themselves by taking into account principles, requirements, utilities, complexity, etc. of algorithms. Using analytical methods to evaluate segmentation algorithm avoids a concrete implementation of the algorithms. However, the real quality of these algorithms cannot be obtained by a simple analytical study.
- **Subjective methods.** They evaluate the segmentation algorithms in a subjective way in which the segmentation results are judged by a human operator. Therefore, the evaluation scores can vary significantly from one human evaluator to another since they do not have necessarily the same standards for assessing the quality of a segmentation. Furthermore, the results can depend on the order in which the human operator observes them. To minimize bias, such a method requires a large set of objects and a large group of humans. Unfortunately, this kind of methods cannot be integrated in an automatic system.
- **System level evaluation methods.** This kind of methods indicates if the characteristics of the results obtained by a segmentation algorithm are suited for the over-all system which uses this segmentation algorithm. However, this evaluation method is indirect. If the process which follows the segmentation generates better results, it does not necessarily mean that the segmentation results were superior, and vice-versa.

- **Empirical goodness or unsupervised methods.** They evaluate the performance of the algorithms by judging the quality of the segmented images themselves. To achieve this task, a set of quality criteria has to be defined. These criteria are established according to human intuition about what conditions should be satisfied by an ideal segmentation. However it seems difficult to establish quantitatively the quality of a segmentation only by using such an a priori criteria.
- **Empirical discrepancy or supervised methods.** A set of reference images presenting the ideal segmentation is first of all built. This set of images, which can be manually segmented by experts of the domain, constitutes a ground-truth. The purpose is to measure the discrepancy between these reference segmentations and that obtained by an algorithm to evaluate. So, these methods try to determine how far a segmented image obtained by an algorithm is from one or several segmented images. A large discrepancy involves a large segmentation error and thus this indicates a low performance of the considered segmentation algorithm.

The empirical discrepancy methods are the most popular for 2D-image segmentation evaluation [21,28]. Indeed they seem to be the most suited for a quantitative evaluation as the measures of quality can be numerically computed, and for an objective evaluation thanks to the ground-truth.

Martin et al. [21] have proposed such a method to evaluate image segmentation algorithms. They built a public corpus containing ground-truth segmentations produced by human volunteers for images of a wide variety of natural scenes. They also defined a measure of segmentation similarity based on the computation of refinement error of a pixel between two segments (i.e. regions) containing this pixel.

In the 3D-domain, there exist some works proposing the quality assessment of segmentation in a specific context. In the MRI (Magnetic Resonance Imaging) field for example, Gerig et al. [13] propose a tool that quantifies the segmentation quality of 3D-images (volumetric images) including different shape distance metrics such as maximum Hausdorff distance, and mean/median absolute distance between object surfaces. For texture mapping, Sander et al. [23] introduce a metric based on the texture stretch induced by the parametrization of the segmented regions and allowing the evaluation of the segmentation quality.

More recently, Attene et al. [3] have proposed some criteria like the aspect of the boundaries (smoothness, length), the hierarchical / multi-scale properties, the robustness, the complexity and the number of parameters. However these criteria rather judge some technical points than the *real* quality of the techniques themselves, they rather fall in the empirical goodness methods. As raised by the authors, the main problem is that the objective quality of a segmentation of a given model is quite difficult to define, since it depends on the viewer's point of view and knowledge.

Berretti et al. [5] have presented some experimental results which are based on a ground-truth to validate their own segmentation algorithm. However, the ground-truth is not available on-line and according to the authors it contains very simple 3D-models (surfaces of revolution, vases, etc.).

Lastly, we proposed a framework to study the assessment problem of 3D-mesh segmentation [4]. Another work proposed by Chen et al. [7] addresses the same task. Both of these works propose a benchmark which is based on a ground-truth corpus of human segmented 3D-models, so they both constitute empirical discrepancy methods; the evaluation of a segmentation algorithm is realized by quantifying the consistency between the reference segmentations of the ground-truth corpus and those obtained by this algorithm on the same models using a set of similarity metrics that we will detail in the next subsection.

2.2 Review and analytic study of mesh segmentation similarity metrics

In the following, we summarize the existing metrics used to evaluate 3D-mesh segmentation and check if they are really reliable in the context of 3D-mesh segmentation evaluation. A reliable measure of mesh segmentation similarity has to possess the following set of properties:

- **No degenerative cases.** The score's measure must be proportional to the similarity degree between an automatic segmentation and the ground-truth segmentations of the same model. For example, an oversegmentation where each vertex (or face) is represented by a segment must give a very low value of similarity, since no ground-truth segmentation can be represented in such a way.
- **Tolerance to refinement.** The segmentation performed by some human observers can be coarse while the segmentation performed by others can be finer. However they basically remain consistent; the difference just lies in the level of refinement. Hence, a reliable segmentation measure has to accommodate and to be invariant to these segmentation granularity differences.

- **Cardinality independence.** The measure must neither assume equal cardinality nor depend on this attribute. This means that two segmentations to be compared can have different numbers of segments and different sizes of segments.
- **Tolerance to cut boundary imprecision.** The segment boundaries are defined in a subjective way. Indeed, it is possible that two volunteers define the same segment on a model with a slight difference between boundaries, however, from a semantic point of view, the segments remain similar. Hence, a reliable measure has to accommodate this imprecision of cut boundaries.
- **Multiple ground-truth.** The measure has to be able to compare one automatic segmentation with multiple ground-truth (reference segmentations) for a given model, otherwise, providing multiple ground-truth in a benchmark is useless. An alternative solution is to simply average the similarity scores obtained between an automatic segmentation and each manual segmentation (reference segmentation), however, this may bias the result and not really reflect how much an automatic segmentation agrees with the multiple ground-truth.
- **Meaningful comparison.** The scores obtained by the measure have to allow a meaningful comparison between different segmentations of the same model and between segmentations of different models. For the first case (segmentations of the same model), the scores have to vary according to the segmentation quality, then, more the automatic segmentation is similar to the ground-truth segmentations of the same model, and better the score is. For the second case (segmentations of different models), the scores have to indicate which kind of 3D-models is the most convenient to segment by an automatic algorithm.

Essentially, the measures used to evaluate 3D-mesh segmentation can be classified into three categories: boundary matching, region differencing and non-parametric tests based measures.

In order to be able to formulate the above measures, we need to define what is a mesh segmentation. We will use this definition (according to Shamir [24]) for the remainder of this article.

Definition 1 Let M be a 3D-mesh, and R the set of mesh elements which are the vertices v_i or the faces f_i of M. A segmentation S of M is the set of sub-meshes $S = \{M_0, ..., M_{k-1}\}$ induced by the partitioning of R into k disjoint sub-sets of vertices or faces.

The three categories of measure are:

1. **Boundary matching.** This kind of measures compute the mapping degree between the extracted region boundaries of two segmentations. Chen et al. [7] proposed to use such a measure called *Cut discrepancy*. It measures the distances between cuts, where each cut represents an extracted region boundary. Let S_1 and S_2 be two segmentations of a 3D-mesh M and C_1, C_2, their respective sets of points on the segment boundaries. Let $d_G(p_1, C_2) = min\{d_G(p_1, p_2), \forall p_2 \in C_2\}$ be the geodesic distance from a point $p_1 \in C_1$ to a set of cuts C_2.
   The Cut discrepancy between S_1 and S_2 is then:

$$CD(S_1, S_2) = \frac{DCD(S_1 \Rightarrow S_2) + DCD(S_2 \Rightarrow S_1)}{avgRadius}$$

   where, *avgRadius* is the average Euclidean distance from a point on the surface to centroid of the mesh, and DCD is a directional function defined as $DCD(S_1 \Rightarrow S_2) = mean\{d_G(p_1, C_2), \forall p_1 \in C_1\}$.
   A value of 0 will indicates a perfect matching between S_1 and S_2. As observed by Chen et al. [7] the measure is undefined when the model has no cuts and decreases to zero as more cuts are added to a segmentation. Hence it suffers from a degenerative case (see section 2.2). In addition, it is not tolerant to refinement since for two segmentations that are perfect mutual refinements of each other, it can provide a large value. Moreover, for the unmatched points, it is possible to change their locations randomly and the measure will keep the same value. It is also not tolerant to imprecision of cut boundaries since it is based on a geodesic distance. Finally, it allows to compare an automatic segmentation to only one ground-truth segmentation.
2. **Region differencing.** These measures compute the consistency degree between the regions produced by two segmentations S_1 and S_2. Berretti et al. [5] have proposed an overlap index representing the extent to which a region R_i of an automatic segmentation overlaps to closest region R_j of a ground-truth segmentation. The overlap index O_{index} of R_i is defined as:

$$O_{index} = max_j \frac{A(R_i \cap R_j)}{A(R_i)}$$

   with $A(.)$ the operator that returns the area of a region. If we suppose that S_1 is the automatic segmentation and S_2 is the ground-truth segmentation, then the distance between them is the average of the Overlap index over-all regions of S_1. This measure falls in a degenerative case when R_i is represented

by one face. Then the over-partitioning is not captured and it also does not allow a comparison to multiple ground-truth.
We (Benhabiles et al. [4]) and Chen et al. [7] proposed to use the consistency error measure. It is based on the computing of a local refinement error L_{3D} of a vertex (or face) v_i between S_1 and S_2 and is defined as:

$$L_{3D}(S_1, S_2, v_i) = \frac{|R(S_1, v_i) \backslash R(S_2, v_i)|}{|R(S_1, v_i)|}$$

where the operator $\backslash$ denotes the set differencing, $|x|$ the cardinality of the set x, and $R(S, v_i)$ the region in segmentation S that contains the vertex v_i, i.e. the subset of vertices corresponding to a sub-mesh M_j of S containing v_i.
This local refinement error produces a positive real valued output that presents the ratio of the number of vertices not shared between the first segment and the second one.
Given this L_{3D}, there exist two ways to combine it for all vertices into a global measure for the entire 3D-mesh: the Global Consistency Error (GCE) and the Local Consistency Error (LCE).
The Global Consistency Error (GCE) forces all local refinements to be in the same direction and is defined as:

$$GCE(S_1, S_2) = \frac{1}{N} min\{\sum_i L_{3D}(S_1, S_2, v_i), \sum_i L_{3D}(S_2, S_1, v_i)\}$$

The Local Consistency Error (LCE) allows for different directions of refinement in different segments of the 3D-mesh:

$$LCE(S_1, S_2) = \frac{1}{N} \sum_i min\{L_{3D}(S_1, S_2, v_i), L_{3D}(S_2, S_1, v_i)\}$$

where N is the number of vertices. For both the GCE and the LCE, a value of 0 indicates a complete similarity, whereas a value of 1 indicates a maximum deviation between the two segmentations being compared. There are two degenerative segmentations that achieve a GCE and a LCE score of zero: one vertex per segment, and one segment for the entire mesh. We can also notice that the measure does not allow a comparison to multiple ground-truth.

Chen et al. [7] proposed to use another measure namely Hamming distance. The Hamming distance between two segmentations S_1 and S_2 measures the region differencing between their respective set of segments. The directional Hamming distance is defined as:

$$D_H(S_1 \Rightarrow S_2) = \sum_i \left\| R_2^i \backslash R_1^{i_t} \right\|$$

where the operator $\backslash$ denotes the set differencing, $\|x\|$ the cardinality of the set x, and $i_t = argmax_k \left\| R_2^i \cap R_1^k \right\|$ which allows to find the closest segment in S_1 to the region (or segment) R_2^i in S_2. Given this D_H, and considering S_2 as the ground-truth, the authors of [7] defined the missing rate M_r and the false alarm rate F_r as follow:

$$M_r(S_1, S_2) = \frac{D_H(S_1 \Rightarrow S_2)}{\|S\|}$$

$$F_r(S_1, S_2) = \frac{D_H(S_2 \Rightarrow S_1)}{\|S\|}$$

and the Hamming distance as the average of missing rate and false alarm rate:

$$HD(S_1, S_2) = \frac{1}{2}(M_r(S_1, S_2) + F_r(S_1, S_2))$$

As observed by the authors [7] the measure has a good behavior when the correspondences between segments are correct but it fails when they are not. Another limit is the comparison to only one ground-truth.

3. **Non-parametric tests.** In the statistical literature there exists a lot of non-parametric measures. We can cite for example Cohen's Kappa [8], Jaccard's index [11], Fowlkes and Mallow's index [11]. The latter two are variants of Rand index [22]. Chen et al. [7] proposed to use Rand index for 3D-mesh segmentation evaluation. This index converts the problem of comparing two segmentations S_1 and S_2 with different numbers of segments into a problem of computing pairwise label relationships. If we denote $l_{S_1}^i$ the corresponding label of all elements (vertices or faces) contained in region R_i of S_1 and similarly $l_{S_2}^i$ the corresponding label of all elements contained in region R_i of S_2, the Rand index can be computed as the ratio of the number of pairs of vertices or faces having the compatible label relationship in S_1 and S_2 and can be defined as:

$$RI(S_1, S_2) = \frac{1}{\binom{N}{2}} \sum_{\substack{i,j \\ i<j}} \mathbf{I}(l^i_{S_1} = l^j_{S_1})(l^i_{S_2} = l^j_{S_2}) + \mathbf{I}(l^i_{S_1} \neq l^j_{S_1})(l^i_{S_2} \neq l^j_{S_2})$$

where $\mathbf{I}$ is the identity function, and the denominator is the number of possible unique pairs among N vertices or faces. This gives a measure of similarity ranging from 1, when the two segmentations are identical, to 0 otherwise. This measure does not allow comparison to multiple ground-truth segmentations.

We can notice that all existing measures suffer from either degenerative cases and/or sensitivity to refinement and/or sensitivity to cut boundaries imprecision and/or limitation in term of comparison to multiple reference (i.e. ground-truth) segmentations. Therefore none of these measures satisfies the whole set of defined criteria.

3 The 3D Probabilistic Rand Index (3D-PRI)

The goal of this measure is to perform a quantitative comparison between a mesh segmentation algorithm and a set of ground-truth segmentations (of the same model). In the field of 2D-image, Unnikrishnan et al. [28] proposed a probabilistic interpretation of Rand Index to evaluate the performance of 2D-image segmentation algorithms and shown the relevance of the obtained results. Hence we have generalized this measure for 3D-mesh segmentation evaluation.

Let S_a be the automatic segmentation to be compared to a set of manual segmentations (ground-truth) $\{S_1, S_2, ..., S_K\}$ of a 3D-mesh M. We denote the corresponding label of a vertex v_i (label of the segment to which belongs vertex v_i) by $l^i_{S_a}$ in segmentation S_a and by $l^i_{S_k}$ in the ground-truth segmentation S_k. It is assumed that the label $l^i_{S_k}$ takes a value ranged between 1 and the number of segments of S_k and similarly $l^i_{S_a}$ takes a value ranged between 1 and the number of segments of S_a. The label relationships for each vertex pair is modeled by an unknown underlying distribution. This can be considered as a process where each human segmenter provides information about the segmentation S_k of the 3D-mesh in the form of binary numbers $\mathbf{I}(l^i_{s_k} = l^j_{s_k})$ for each pair of vertices (x_i, x_j). The set of all perceptually correct segmentations defines a Bernoulli distribution over this number, giving a random variable with expected value denoted as p_{ij}. Hence, the set $\{p_{ij}\}$ for all unordered pairs (i, j) defines a generative model of correct segmentations for the 3D-mesh M. The 3D Probabilistic Rand Index is then defined as:

$$\text{3DPRI}(S_a, \{S_k\}) = \frac{1}{\binom{N}{2}} \sum_{\substack{i,j \\ i<j}} e_{ij}p_{ij} + (1-e_{ij})(1-p_{ij}) \quad (1)$$

where e_{ij} denotes the event of a pair of vertices i and j belonging to the same segment (or region) in the automatic segmentation:

$$e_{ij} = \mathbf{I}(l^i_{S_a} = l^j_{S_a})$$

and p_{ij} denotes the probability of the vertices i and j belonging to the same segment in the ground-truth set $\{S_k\}$ and is given by the sample mean of the corresponding Bernoulli distribution as suggested by Unnikrishnan et al. [28]:

$$p_{ij} = \frac{1}{K} \sum_k \mathbf{I}(l^i_{S_k} = l^j_{S_k})$$

The 3D-PRI takes a value ranged between 0 and 1, where 0 indicates no similarity between S_a and $\{S_1, S_2, ..., S_k\}$, and 1 indicates a perfect similarity.

Note that with this formulation for p_{ij}, computing the 3D-PRI is equivalent to averaging the RI over the multiple ground-truths. However the 3D-PRI formulation is generic and we can imagine a different and more efficient way to compute the p_{ij}. The main advantage of the simple mean estimator is its fast computation.

We have noticed in practice, however, that the 3D-PRI suffers from lack of discriminative power in its values. Indeed, the values obtained by the index do not allow to clearly decide if a segmentation obtained by an automatic algorithm is relevant or not. This is due to the limited effective range of 3D-PRI in term of maximum and minimum value. To address this drawback, we present in the next section, the 3D normalized probabilistic Rand index (3D-NPRI).

4 3D Normalized Probabilistic Rand Index (3D-NPRI)

Our objective is to normalize the 3D-PRI, in order to increase its dynamic range and thus its discriminative power. Hence we need to define a baseline to which the index can be expressed. For 3D-mesh segmentations, the baseline may be interpreted as the expected value of the index under some particular segmentations of the input 3D-model. A popular strategy [11,28] of index normalization with respect to its baseline is:

$$Normalized\ index = (Index - Expected\ index)/ \\ (Maximum\ index - Expected\ index) \quad (2)$$

As observed by Unnikrishnan et al. [28] there is a little agreement in the statistics community regarding whether the value of "Maximum index" should be estimated from the data or set constant. We choose to follow what was done by Unnikrishnan et al. [28] and set the value to be 1 (the maximum possible value of the 3D-PRI). Thus, we avoid the practical difficulty of estimating this quantity for complex data sets.

Another parameter to define is the expected probabilistic Rand index E(3D-PRI). One may draw an analogy between the E(3D-PRI) and the 3D-PRI in equation 1 as follow:

$$E\left[\text{3DPRI}(S_a, \{S_k\})\right] = \frac{1}{\binom{N}{2}} \sum_{\substack{i,j \\ i<j}} e'_{ij} p_{ij} + (1 - e'_{ij})(1 - p_{ij}) \quad (3)$$

where $e'_{ij} = E\left[\mathbf{I}(l^i_{S_a} = l^j_{S_a})\right]$. This latter quantity has to be computed in a meaningful way. Unnikrishnan et al. [28] proposed to estimate it from segmentations of all images of the database for all unordered pairs (i, j). Let Φ be a number of images in a data set and K_ϕ the number of ground-truth segmentations of image ϕ. Then, e'_{ij} is expressed as:

$$e'_{ij} = \frac{1}{\Phi} \sum_{\phi} \frac{1}{K_\phi} \sum_{k=1}^{K_\phi} \mathbf{I}(l^i_{S^k_\phi} = l^j_{S^k_\phi})$$

However, this estimation can only be used in a database of 2D-images having equal sizes (where each pixel has its correspondent over all the other segmented images). In the 3D case, it is not possible, since the different models of the corpus have different number of vertices and different connectivities. One possible way to compute the E(3D-PRI) while keeping a correct baseline and without having any constraint on the corpus, is to use random segmentations S_r:

$$E\left[\text{3DPRI}(S_a, \{S_k\})\right] = \frac{1}{N} \sum_{r=1}^{N} \text{3DPRI}(S_r, \{S_{K_r}\}) \quad (4)$$

where N is the number of 3D-models in our corpus and $\{S_{k_r}\}$ are ground-truths of the model concerned by S_r. We then define the 3D-NPRI of an automatic segmentation of a given 3D-model as follow:

$$\text{3DNPRI}(S_a) = \frac{\text{3DPRI}(S_a, \{S_K\}) - E\left[\text{3DPRI}(S_a, \{S_k\})\right]}{1 - E\left[\text{3DPRI}(S_a, \{S_k\})\right]} \quad (5)$$

The random segmentations were generated using a simple algorithm: L seed vertices were randomly chosen on the object, then L connected regions were obtained by a simple region growing mechanism. The number of segments (or regions) takes a value ranged between 2 and the number of vertices of the concerned model. Figure 1 shows some 3D-models of the corpus on which the random segmentation algorithm was applied. We have to precise here that the 3D-NPRI is not affected by the choice of these random segmentations. Indeed we will show later (see figure 3) that the 3D-PRI provides very stable values when comparing ground-truth segmentations to random segmentations (even with very different granularities) hence the normalization constant E(3D-PRI) (see equation 4) is almost invariant to the choice of the random segmentations S_r.

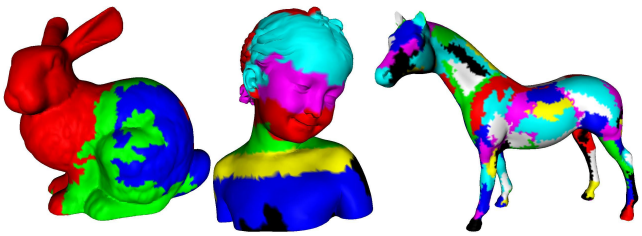


Fig. 1 Random segmentations of some 3D-models of the corpus.

Hence, the 3D-NPRI will take a value with a lower bound of -1 and an upper bound of 1, where -1 indicates no similarity between the automatic segmentation and the ground-truth segmentations of the same model, and 1 indicates a perfect match. The lower bound of -1 is explained by the fact that the expected Index can not exceed 0.5 since we compare a set of random segmentations to a set of ground-truth segmentations (see subsection 6.1). Therefore, the worst case will be:

$$\text{3DNPRI}(S_a) = \frac{0 - 0.5}{1 - 0.5} = -1$$

where the automatic segmentation has no similarity with its corresponding ground-truths.

Note that the metric's definition does not take into account model with different sampling. Moreover, the score of the metric changes by changing the order of vertices on the automatic segmentation and the ground-truths of the same model. However, in our case, it is not really a drawback since we compare segmentations of the same model while keeping the same sampling and the same order of vertices.

5 Ground-Truth Corpus

The current version of our corpus is an improved version of [4] in term of number of models and ground-truth segmentations per model. The corpus is available on-line[2] and contains twenty-eight 3D-models (as triangle meshes) grouped in five classes, namely *animal*, *furniture*, *hand*, *human* and *bust*. Each 3D-model of the corpus is associated with 4 manual segmentations which give a total of 112 ground-truth segmentations done by 36 volunteers. Figure 2 illustrates the models of the corpus with one manual segmentation per model. We have selected a small number of varied models with respect to a set of properties. All the selected models are manifold, connected, and do not have intersecting faces. Hence they are supported as an input by any segmentation algorithm. In order to collect precise manual segmentations, we have assisted the volunteers in tracing the vertex-boundaries through the different models. Note that the volunteers have freely segmented the models and no condition was imposed on the manner with which they have segmented them. For this task, we used MeshLab[3] application allowing an explicit vertex-per-vertex segmentation of models using colors.

Chen et al. [7] proposed another corpus that seems complementary to ours: they present more objects (380 3D-models of the Watertight Track of the 2007 SHREC Shape-based Retrieval Contest [14]) when we selected a small representative set (it allows to rapidly evaluate a segmentation algorithm without running it on 380 objects). They chose to use the web application Amazon's Mechanical Turk[4] to collect the manual – i.e. ground-truth – segmentations without any supervision when we chose to supervise our volunteers to obtain more precise manual segmentations. Finally, their ground-truth presents face-based segmentations whereas ours contains vertex-based segmentations.

6 Experimental comparison of properties of existing segmentation similarity metrics

In what follows, we provide an experimental study of the 3D-PRI/3D-NPRI properties and we compare them to the existing metrics for assessing 3D-mesh segmentation quality. For this end, we use our corpus models and their corresponding ground-truths.

Most of the measures introduced in section 2.2 quantify *dissimilarity* (the lower is the number, the best is the segmentation result) between segmentations rather than *similarity*. In order to have a meaningful comparison between these measures and the 3D-PRI/3D-NPRI, we define the quantities $CDI(S_1, S_2) = 1 - CD(S_1, S_2)$, $GCI(S_1, S_2) = 1 - GCE(S_1, S_2)$, $LCI(S_1, S_2) = 1 - LCE(S_1, S_2)$, and $HDI(S_1, S_2) = 1 - HD(S_1, S_2)$. The "I" in the acronyms stands for "Index", complying with the popular usage of the term in statistics when quantifying similarity. Hence, except the CDI, all of the other indexes are in the range $[0, 1]$ with a value of 0 indicating no similarity between segmentations of the same model and a value of 1 indicating a perfect match. The CDI is in the range $]-\infty, 1]$.

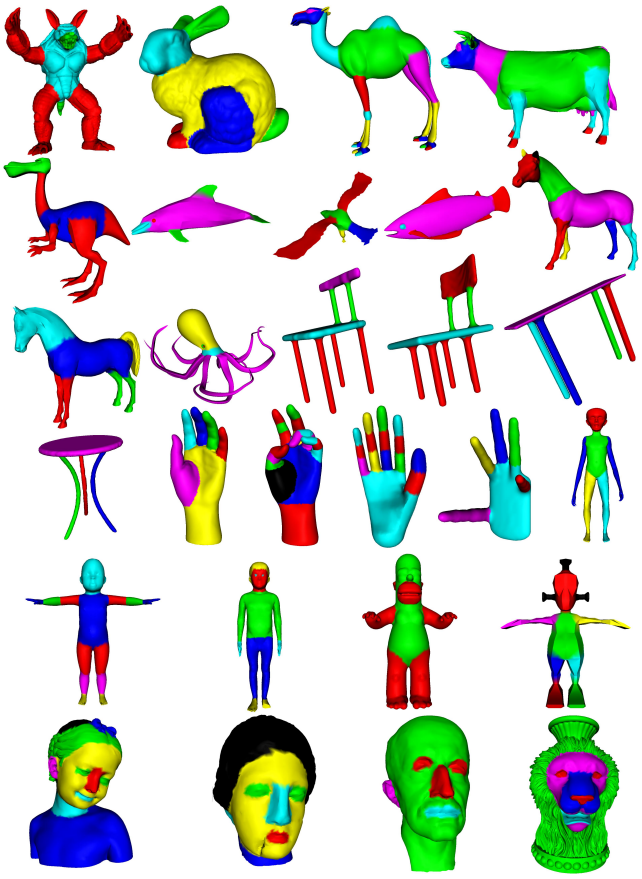


Fig. 2 Models of our corpus associated with one ground-truth.

6.1 Sensitivity to degenerative cases

The first property to study is the sensitivity of each index regarding degenerative cases. For this end, we compare our Probabilistic Rand Index (3D-PRI) with the Cut Discrepancy Index (CDI), the Hamming Distance Index (HDI), the Global and Local Consistency Index (GCI/LCI), and the Overlap Index (OI) for three kinds of random segmentations namely *extreme-low segmentation* (segmentation composed of a 2 or 3 segments), middle-segmentation (segmentation composed of a num-

[2] http://www-rech.telecom-lille1.eu/3dsegbenchmark/
[3] http://meshlab.sourceforge.net/
[4] http://www.mturk.com/

ber of segments which is similar to that of ground-truths of the corresponding model), and *extreme-high segmentation* (segmentation composed of more than 50 segments). They were generated using a random segmentation algorithm. Figure 3 presents the results obtained by the comparison of these random segmentations to the set of the ground-truths for each model of the corpus. Each index of the figure is computed for the three kinds of segmentation (extreme-high segmentation, middle-segmentation, and extreme-low segmentation) and averaged across the entire data set. Since the segmentations are random, the scores obtained by the metrics are expected to be low for the three kinds of segmentation, and it is the case for the 3D-PRI. We can notice, however that although the random segmentations are totally different from the ground-truths, the scores of the other metrics are very high (very good) for certain segmentations with degenerative granularity (extreme-high and/or extreme-low). Hence the 3D-PRI is the most stable regarding degenerative cases considering its scores which are less than 0.32.

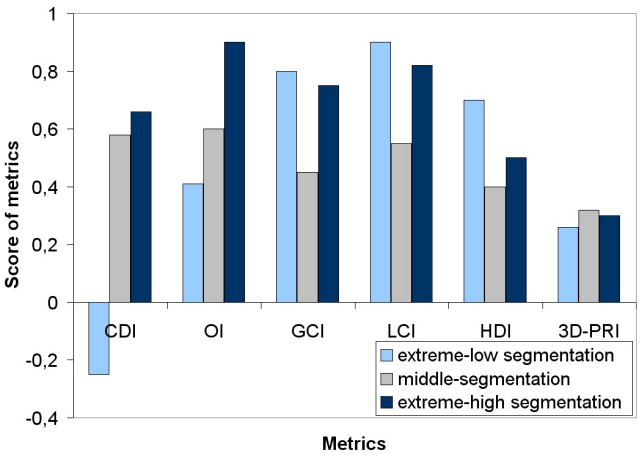


Fig. 3 Comparison of three levels of random segmentation (extreme-low, middle, and extreme-high) to the ground-truths for the whole corpus using different indexes.

6.2 Tolerance to refinement

The second property to study is the tolerance of each index to refinement. For this end, we perform two kinds of experiments. The first one uses segmentations with mutual refinements, and the second one uses segmentations with hierarchical refinements. The obtained results for the first experiment are presented in figure 4.

It shows two segmentations of the *dinopet* model which are perfect mutual refinements of each other, and

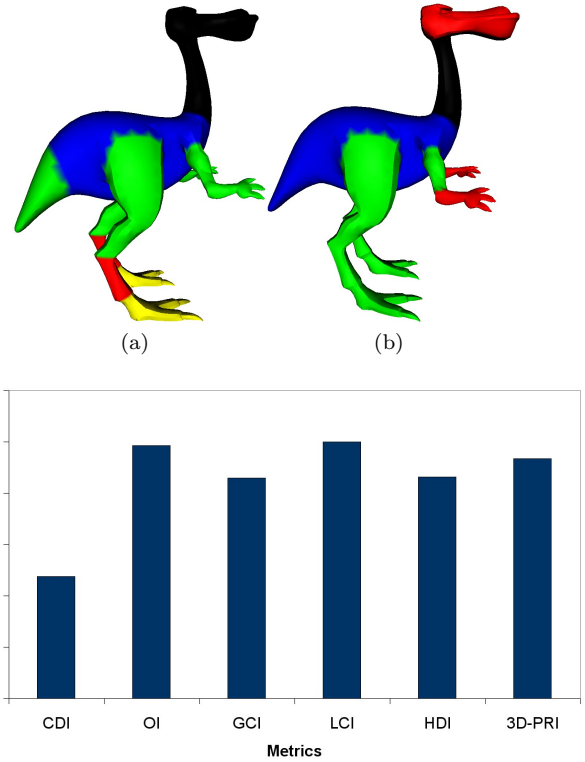


Fig. 4 Tolerance to mutual refinement of different indexes, by comparing two segmentations (a,b) with perfect mutual refinement for the dinopet model.

a plot in which is computed the similarity between the two segmentations using different metrics. The plot of figure 4 clearly shows that the CDI fails to capture the similarity between the two segmentations (a) and (b). Although the two segmentations are similar (the difference just lies in the level of refinement). However, the other metrics have a good behavior toward this kind of refinement since all of them give scores which are close to 1.

The second experiment was performed using the hierarchical segmentation algorithm of Attene et al. [2]. We generated several levels of segmentation (from 4 segments to 15 segments) on the *horse* model of our corpus then we compared these 12 versions to the ground-truths. Figure 5 illustrates the obtained results using different indexes. The OI and the GCI does not appear on the figure since they have the same behavior as the LCI. The figure clearly shows that the CDI is less stable toward hierarchical refinement than the other indexes. The LCI seems completely invariant while the 3D-PRI and the HDI present a slight variation; they are not fully invariant but present a good tolerance to refinement.

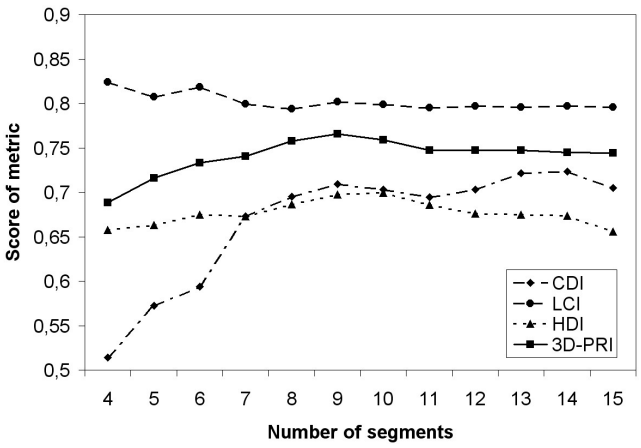


Fig. 5 Tolerance to hierarchical refinement of different indexes, by comparing several levels of segmentation of the horse model to its corresponding ground-truths.

6.3 Independence from cardinality

The third property to study is the independence of each index toward segmentation cardinality. According to the previous performed experiments about the first two properties (degenerative cases and refinement), the CDI seems to be the only metric which depends on the cardinality, in a critical way. Indeed, the comparison between two segmentations with different number of segments will give a bad score using this metric whatever the quality.

6.4 Tolerance to imprecision of cut boundaries

The fourth property to study is the tolerance of each index to the imprecision of cut boundaries. For this end, we manually segmented a simple model (bitorus) into 2 segments. We proposed 5 segmentations (figure 6 (a to e)) where each one of them has a slight difference in the boundary position with comparison to the others, then we computed the similarity between segmentation (c) and the other segmentations. The plot in figure 6 shows the obtained results using different indexes. Contrary to the other indexes, the CDI gives low values of similarity between segmentations. Although the CDI is not in the same range as the other metrics, the plot still allows to illustrate the qualitative behavior of this latter index toward the imprecision of cut boundaries. We can notice also that except the 3D-PRI which presents a slight variation but a good tolerance, the other indexes are almost invariant.

At this point, we have shown that the 3D-PRI satisfies the five properties: ability to compare one automatic segmentation with multiple ground-truth segmentations, no degenerative cases, tolerance to refinement, independence from segmentation cardinality, and tolerance to imprecision of cut boundaries. We also have shown that the 3D-PRI outperforms the other indexes in terms of the first two properties. We show in the next experiments that the normalization of this index (into 3D-NPRI) improves its discriminative power and give better results in term of meaningful comparison.

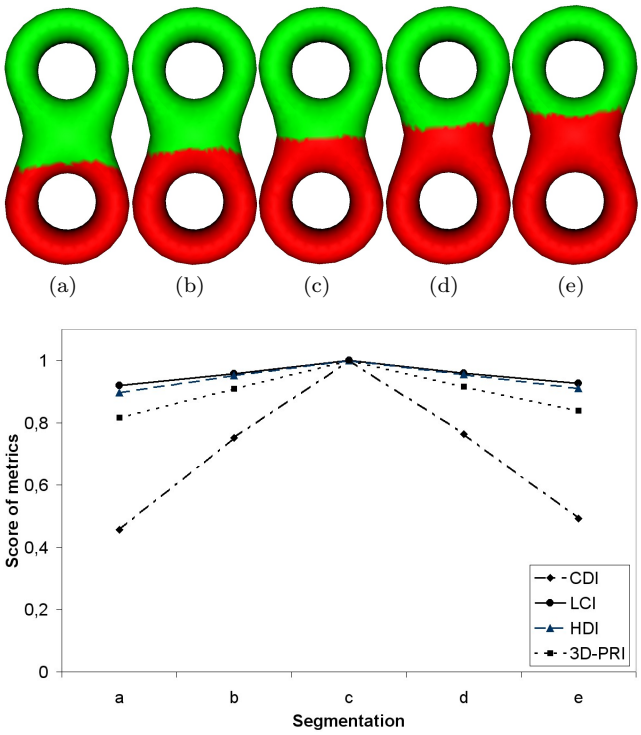


Fig. 6 Tolerance to imprecision of cut boundaries of different indexes, by comparing segmentation (c) to segmentations (a to e) for the bitorus model.

6.5 Meaningful comparison

The main advantage of the 3D-NPRI is the ability to provide values that allow a meaningful comparison between segmentations of different 3D-models. Figure 7 demonstrates this behavior. The top two rows show different 3D-models of our corpus segmented at different granularity using the hierarchical algorithm of Tierny et al. [27]. These automatic segmentations are compared to the ground-truth corpus (see figure 2) using the previous indexes and our 3D-NPRI. Visually, regarding the ground-truth, segmentations a and b (figure 7) seem very poor, segmentations c, d, and f are correct, and segmentation e is perfect. One can notice that the OI similarity is high for all of the 3D-models. Hence, it cannot indicate which segmentation is the

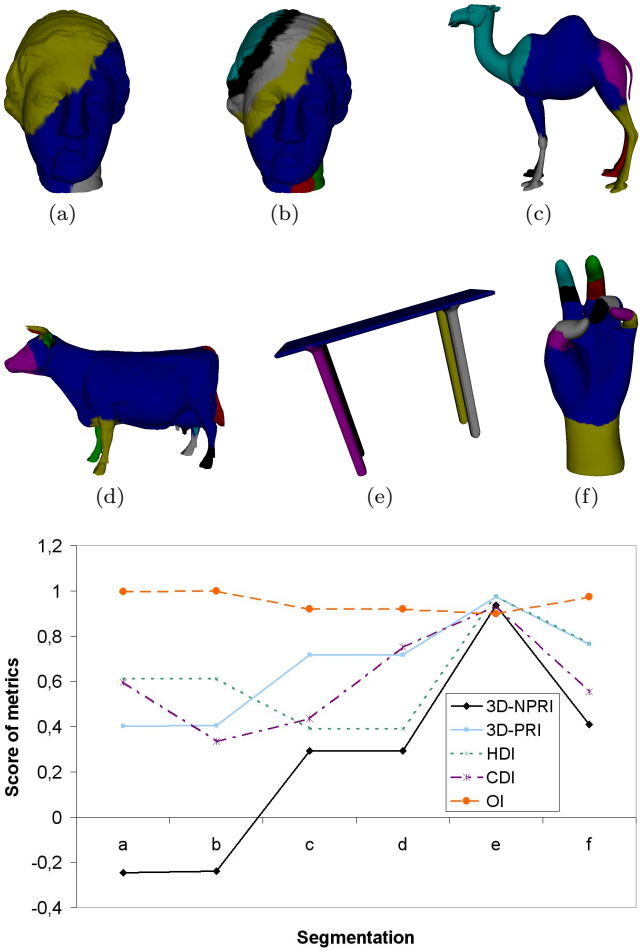


Fig. 7 Example of comparing segmentations of different models: From a to f segmentations using algorithm from [27]. The plot (g) shows the scores of different indexes for each segmentation (a to f).

best. Note that although the HDI gives lower scores than the OI, it also fails to distinguish between correct and poor segmentations since it gives high values for poor ones (figure 7.a and 7.b) and low values for correct ones (figure 7.c and 7.d). The GCI/LCI does not appear in the plot in order to keep a clear display. This latter metric has the same behavior than HDI. The CDI has slightly a better behavior than HDI but still to fail distinguishing between correct and poor segmentations. The 3D-PRI reflects the correct relationship among the segmentations. However, its range is small, and the expected value is unknown, hence it is difficult to determine which segmentation is really good. The 3D-NPRI fixes all of these drawbacks. It reflects the desired relationship among the segmentations with no degenerate cases. Besides, any segmentation which gives a score significantly above 0 can be considered as relevant (since it provides results significantly better than random segmentations).

6.6 Discriminative power

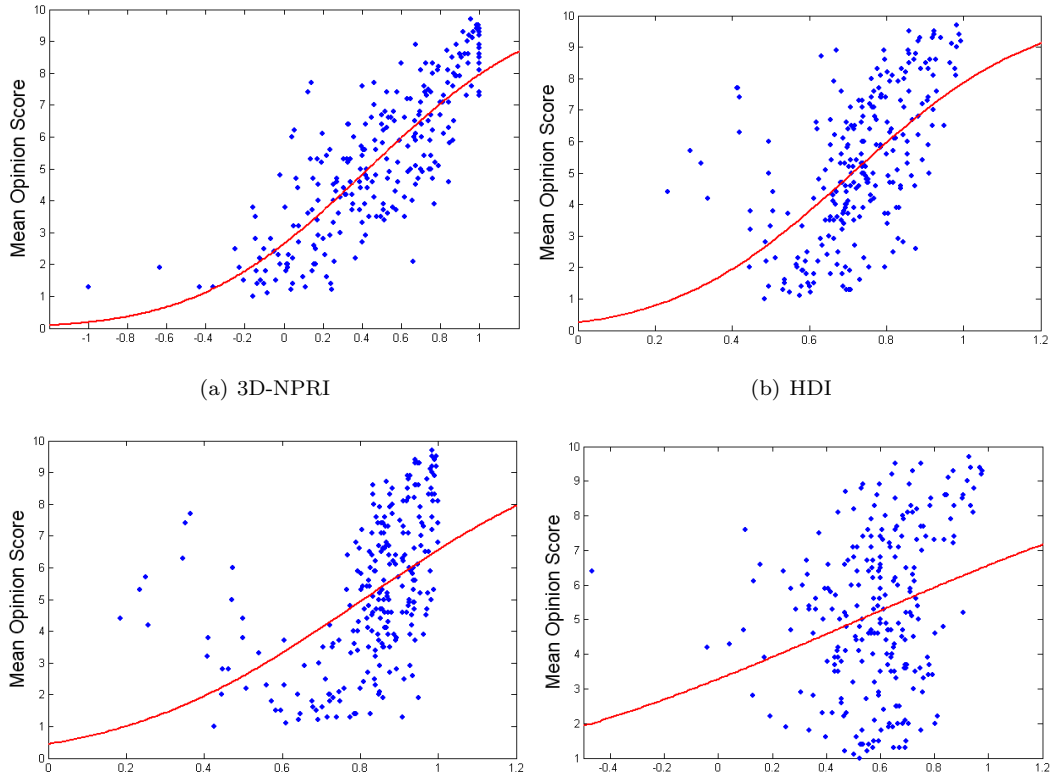
The best way to attest the discriminative power of our 3D-NPRI, is to show that its values are well correlated with the rates given by users for a set of segmentations.

In the following experiment, we study the correlation between the metrics' values and the observers' rates of a set of segmentations obtained from our corpus models. For this end, we used the following algorithms: Attene et al. [2], Lavoué et al. [20], Shapira et al. [25], and Tierny et al [27]. We collected 250 segmentations including 28 ground-truth segmentations and 28 random segmentations. Except the algorithm of Lavoué et al. [20], the others are hierarchical. Hence, we generated for each one of them two levels of segmentation, namely coarse and fine, which gives 28×2 segmentations per algorithm and 28 segmentations from Lavoué's et al. [20] algorithm. We computed the quality index of these 250 segmentations (using our ground-truth) using the different metrics. We then asked several observers to give a rate reflecting the perceived quality of each segmentation between 1 (bad segmentation) and 10 (perfect segmentation). Hence each segmentation was associated with quality index values from the different metrics and a subjective Mean Opinion Score (MOS) from human observers. This latter score reflects the opinion of observers toward the quality of a segmentation. The MOS of the segmentation i is computed as follow:

$$MOS_i = \frac{1}{N} \sum_{j=1}^{N} m_{ij}$$

where N is the number of observers (10 in our experiment), and m_{ij} is the rate (in the range $[1, 10]$, 10 for a very good segmentation) of the jth observer given to the ith segmentation. For the correlation, we considered a statistical indicator namely the Pearson Product Moment Correlation [10]. This indicator measures the linear dependence between two variables X and Y. In order to optimize the matching between the values of the different metrics and the MOS of observers, we performed a psychometric curve fitting using the Gaussian psychometric function (recommended by [9]).

Table 1 shows the results of correlation between the values of different metrics and the MOS of observers for Pearson indicator. The results in the table clearly shows that the 3D-NPRI outperforms the other metrics in term of correlation for each category and for the whole corpus. Moreover, the Pearson correlation value



(a) 3D-NPRI (b) HDI

(c) LCI (d) CDI

Fig. 8 Subjective MOS vs metric values for the whole corpus models and for different metrics. Each circle represents a segmentation. The Gaussian fitted curve is displayed in red.

Table 1 Pearson correlation values (%) between the Mean Opinion Scores and the values of different metrics for each model category of our corpus.

	CDI	**OI**	**GCI**	**LCI**	**HDI**	**3D-NPRI**
animal	2.6	2.3	9.3	8.3	16.9	58.7
bust	10.9	0	45.9	61.1	54.8	77.4
furniture	5.8	14.8	49.9	50.5	63	73.2
hand	21.2	1	54.1	54.4	57.5	70.2
human	1.5	5.5	32.1	32.6	39	51.6
whole	**7.1**	**2.6**	**23.7**	**20.9**	**32.9**	**66.1**

of the 3D-NPRI for the whole corpus is high (66.1%), when those of the other metrics are quite bad (less than 33%). This means that except the 3D-NPRI, the other metrics fail to distinguish between good and bad segmentations. Figure 8 presents the psychometric curve fitting between the objective and subjective scores for 3D-NPRI, HDI, LCI and CDI for 250 segmentations of the corpus models. It visually illustrates the superiority of the 3D-NPRI for predicting the subjective opinion, and leads to conclude that the 3D-NPRI has the best discriminative power. These results clearly validate the 3D-NPRI, since they are in agreement with the human opinion.

The properties of each metric are summarized in table 2 according to the performed experiments in this section.

7 Application for the evaluation of recent segmentation algorithms

In this section, we apply the 3D-NPRI together with the Chen's et al. [7] corpus and our corpus (described in section 5) to evaluate a set of recent automatic segmentation algorithms, then we compare the obtained results by the two corpuses. We have considered the six recent automatic segmentation algorithms used in Chen et al. [7]: Attene et al. [2], Lai et al. [19], Golovinskiy et al. [15], Katz et al. [17], and Shapira et al. [25]. The six algorithms are respectively based on: fitting primitives, random walks, normalized cuts/randomized cuts, core extraction, and shape diameter function. The segmentations using these algorithms for the Chen's corpus are available on-line. On the other hand, we used Attene's et al. [2], and Shapira's et al. [25] algorithms (the only

Table 2 Properties of existing similarity metrics.

	CDI	OI	GCI	LCI	HDI	3D-NPRI
Degenerative cases	yes	yes	yes	yes	yes	no
Tolerance to refinement	no	yes	yes	yes	yes	yes
Cardinality independence	no	yes	yes	yes	yes	yes
Tolerance to cut imprecision	no	yes	yes	yes	yes	yes
Multiple ground-truth	no	no	no	no	no	yes
Meaningful comparison	no	no	no	no	no	yes
Strong discriminative power	no	no	no	no	no	yes

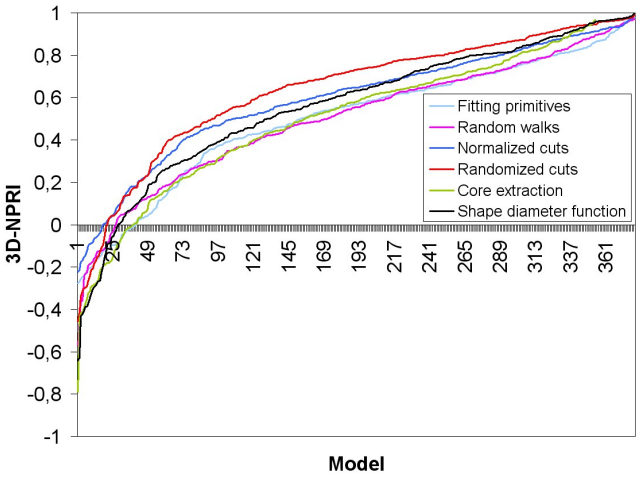
algorithms available on-line among the previous six) to generate automatic segmentations on our corpus models. The reader can refer to the original papers for more details about the six algorithms.

Note that all the algorithms cited above are part-type hierarchical segmentation methods. Hence for each one of them we can generate several levels of segmentation. Chen et al. [7] provided only one level of segmentation for each algorithm applied on their corpus. For this end, they used the parameter settings recommended by the authors of the algorithms. To keep a valid comparison between the two corpuses, we also used the parameter settings recommended by the authors of the algorithms to generate segmentations on our corpus models. Note that the level of segmentation will not influence the evaluation results since we proved that the 3D-NPRI is tolerant to hierarchical refinement (see figure 5).

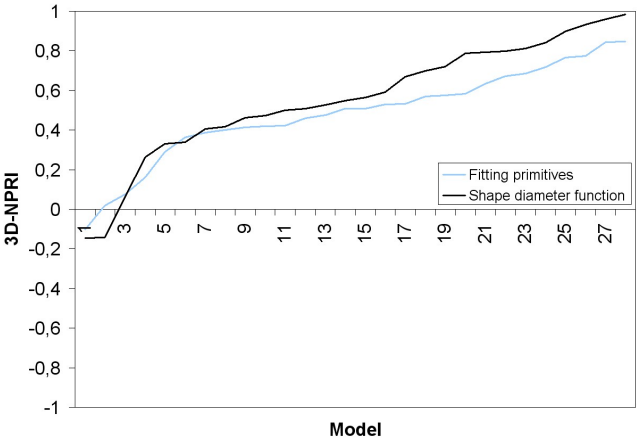
To ensure a relevant comparison between the algorithms, we compute the 3D-NPRI for every 3D-model of the Chen's corpus and of our corpus. Figure 9 shows the 3D-NPRI for each model of the two corpuses and for each algorithm. The values are sorted in increasing order for each algorithm, hence the jth model may not be the same across algorithms. This kind of graph was already applied for segmentation evaluation in the field of 2D-image [28].

Table 3 presents the rank of each algorithm together with the 3D-NPRI mean value over all the two corpuses.

Table 3 and figure 9 demonstrate, as expected, that the segmentations obtained by the six algorithms are relevant since most of the values of the 3D-NPRI are greater than zero. The Randomized Cut algorithm seems to provide the best results. It is very interesting to notice that the Fitting Primitives and Shape Diameter keep similar behavior for the two corpuses although these two corpuses are very different: the profiles of the 3D-NPRI distribution (see figure 9) and the mean 3D-NPRI values (see table 3) for these algorithms are almost exactly the same for both corpuses. Hence it validates the fact that our corpus, since it presents high quality manual segmentation and heterogeneous models, is clearly efficient for segmentation evaluation de-



(a) Results on Chen's et al. [7] corpus



(b) Results on our corpus

Fig. 9 Scores of 3D-NPRI sorted in increasing order over all the two corpus models.

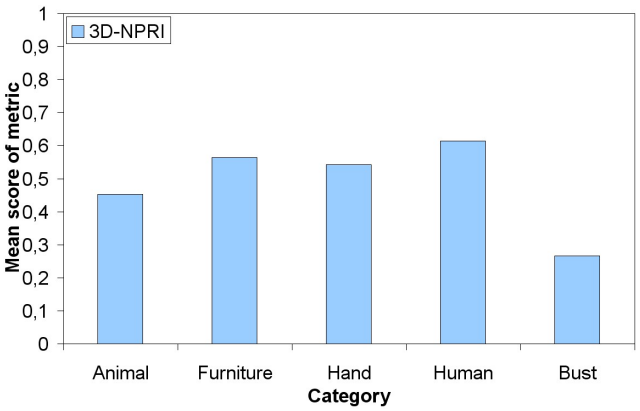
spite its small size. Another interesting experiment is to see which category models the algorithms fail to segment well. For this end, we average the 3D-NPRI for each category of the two corpuses. Figure 10 and 11 illustrate the obtained results for the six algorithms. One can notice that whatever the corpus is, there is no algorithm that is reaching the highest scores for all categories. Moreover, each algorithm has at least one cat-

Table 3 Algorithms ranking applied on respectively the Chen's corpus and our corpus.

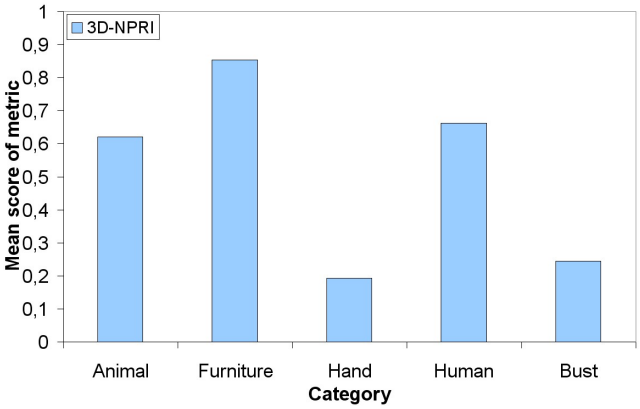
Algorithm	3D-NPRI mean	Rank
Fitting primitives	0.49/0.49	5/2
Random walks	0.50/-	4/-
Normalized cuts	0.59/-	2/-
Randomized cuts	0.63/-	1/-
Core extraction	0.46/-	6/-
Shape diameter function	0.56/0.55	3/1

egory inadequately segmented since its mean 3D-NPRI value is very low (close to 0 or less). The core extraction algorithm for instance fails to adequately segment the *Bearing* and *Mech* categories (see figure 11(e)). This result is straight since the concerned algorithm is a part-based one. Indeed, it tries to detect the core of a model which from a semantic point of view is hard to define in such categories. As observed by Chen et al. [7], some algorithms do not necessarily segment the best (with comparison to others) categories for which they were designed. We can notice this behavior on our corpus too. For instance, the algorithm based on Fitting Primitives gives greater 3D-NPRI score (better) for the *hand* category than the algorithm based on Shape Diameter Function and vice versa for the *furniture* category. As raised by Chen et al. [7], this means that either the human observers do not segment models in the expected way, or the part structures of these models are revealed by other properties.

Our results and those of Chen et al [7] are coherent. This is straight since our metric is a probabilistic interpretation of the Rand Index (metric used by Chen et al. [7] to analyze and evaluate the algorithms) to which we added a normalization allowing a better results analysis.



(a) Fitting primitives



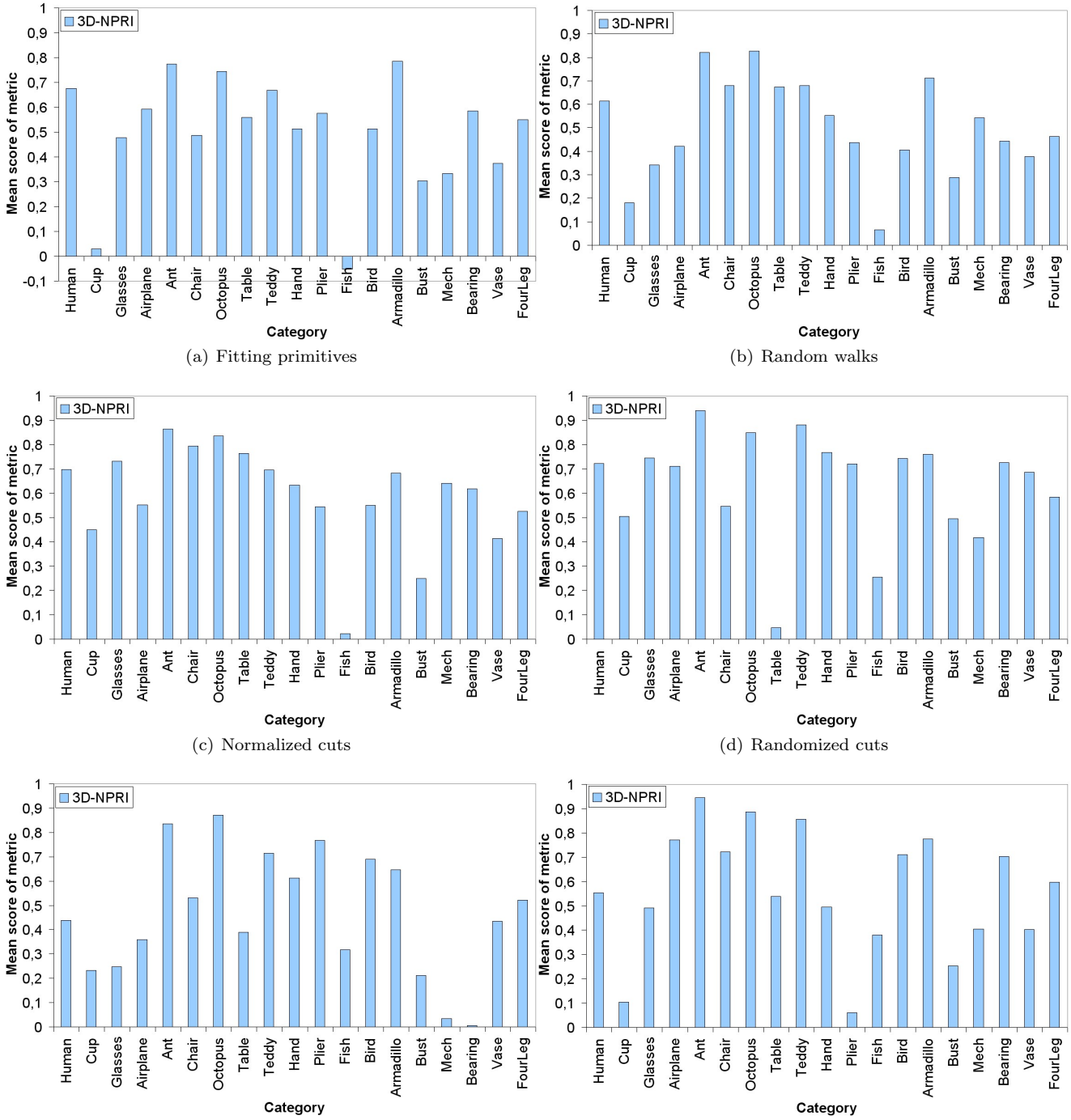
(b) Shape diameter function

Fig. 10 Scores of 3D-NPRI averaged for each category models of our corpus.

8 Conclusion

This paper presents a thorough comparison between existing similarity metrics and a new one addressing the assessment problem of mesh segmentation. For this end we use a corpus of manually segmented models. This corpus is an improved version of our previous one [4] and is available on-line. The new 3D-NPRI metric is a probabilistic interpretation of the Rand Index which allows to quantify the consistency between multiple segmentations of a 3D-mesh model. The paper shows that this new metric outperforms existing ones in terms of properties and discriminative power. The results are validated by comparing subjective scores of human observers to the objective metric scores. Finally the measure is applied together with the Chen's et al. [7] corpus and our corpus to evaluate six recent 3D-mesh segmentation algorithms. This evaluation allowed to compare the obtained results depending on the corpus and showed their coherence.

For future work, we plan to explore other kind of estimator to compute the p_{ij} (see equation 1 in section 3) in order to improve the correlation between metric's scores and the observes' scores, we also plan to enrich our subjective tests by integrating more experiments allowing to compare algorithms. Finally, we plan to exploit our ground-truths to design a learning segmentation algorithm.



(a) Fitting primitives (b) Random walks

(c) Normalized cuts (d) Randomized cuts

(e) Core extraction (f) Shape diameter function

Fig. 11 Scores of 3D-NPRI averaged for each category models of the Chen's corpus.

Acknowledgements We would like to thank Marco Attene, Ariel Shamir, Shy Shalom, and Julien Tierny for providing us the source code or the binary of the segmentation algorithms and Fatan Souhaib for implementing the 3dsegbenchmark site. We also would like to thank Xiaobai Chen for providing on-line the manual and automatic segmentations of different algorithms. We thank AIM@SHAPE, GAMMA-INRIA, and Princeton Shape Benchmark databases for providing 3D-mesh models.

This work is supported by the ANR (Agence Nationale de la Recherche, France) through MADRAS project (ANR-07-MDCO-015).

References

1. Antini, G., Berretti, S., Pala, P.: 3d mesh partitioning for retrieval by parts application. In: IEEE International Conference on Multimedia & Expo (ICME05) (2005)
2. Attene, M., Falcidieno, B., Spagnuolo, M.: Hierarchical mesh segmentation based on fitting primitives. Vis. Comput. **22**(3), 181–193 (2006)
3. Attene, M., Katz, S., Mortara, M., Patané, G., Spagnuolo, M., Tal, A.: Mesh segmentation, a comparative study. IEEE International Conference on Shape Modeling and Applications pp. 7–7 (2006)
4. 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: IEEE International Conference On Shape Modeling And Application (SMI) (2009)
5. Berretti, S., Bimbo, A.D., Pala, P.: 3d mesh decomposition using reeb graphs. Image Vision Comput. **27**(10), 1540–1554 (2009)
6. Biederman, I.: Recognition-by-compenents: A theory of human image understanding. Psychological Review **94**, 115–147 (1987)
7. Chen, X., Golovinskiy, A., Funkhouser, T.: A benchmark for 3d mesh segmentation. ACM Transactions on Graphics (SIGGRAPH) **28(3)** (2009)
8. Cohen, J.: A coefficient of agreement for nominal scales. Educ. and psychological Measurment pp. 37–46 (1960)
9. Corsini, M., Gelasca, E.D., Ebrahimi, T., Barni, M.: Watermarked 3d mesh quality assessment. IEEE Transaction on Multimedia **9**, 247–256 (2007)
10. Daniel, W.W.: A Foundation For Analysis In The Health Sciences Books. 7th edition.John Wiley and sons. (1999)
11. Fowlkes, E.B., Mallows, C.L.: A method for comparing two hierarchical clusterings. Journal of the American statistical association **78(383)**, 553–569 (1983)
12. Funkhouser, T., Kazhdan, M., Shilane, P., Min, P., Kiefer, W., Tal, A., Rusinkiewicz, S., Dobkin, D.: Modeling by example. ACM Transactions on Graphics (Proc. SIGGRAPH) (2004)
13. Gerig, G., Jomier, M., Chakos, A.: Valmet: A new validation tool for assessing and improving 3d object segmentation. In: MICCAI 2001: Fourth International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 516–523. Springer (2001)
14. Giorgi, D., Biasotti, S., Paraboschi, L.: Shrec: Shape retieval contest: Watertight models track. In: http://watertight.ge.imati.cnr.it/ (2007)
15. Golovinskiy, A., Funkhouser, T.: Randomized cuts for 3d mesh analysis. ACM Trans. Graph. **27**(5) (2008)
16. Ji, Z., Liu, L., Chen, Z., Wang, G.: Easy mesh cutting. Computer Graphics Forum **25**(3), 283–291 (2006)
17. Katz, S., Leifman, G., Tal, A.: Mesh segmentation using feature point and core extraction. The Visual Computer **21**(8-10), 649–658 (2005)
18. Katz, S., Tal, A.: Hierarchical mesh decomposition using fuzzy clustering and cuts. ACM Transactions on Graphics (SIGGRAPH) **22(3)**, 954–961 (2003)
19. Lai, Y.K., Hu, S.M., Martin, R.R., Rosin, P.L.: Fast mesh segmentation using random walks. In: SPM '08: Proceedings of the 2008 ACM symposium on Solid and physical modeling (2008)
20. Lavoué, G., Dupont, F., Baskurt, A.: A new cad mesh segmentation method, based on curvature tensor analysis. Computer Aided Design **37(10)**, 975–987 (2005)
21. Martin, D., Fowlkes, C., Tal, D., Malik, J.: A database of human segmented natural images and its application to evaluating algorithms and measuring ecological statistics. International Conference On Computer Vision **2**, 416–423 (2001)
22. Rand, W.M.: Objective criteria for the evaluation of clustering methods. Journal of the American statistical association **66(336)**, 846–850 (1971)
23. Sander, P.V., Snyder, J., Gortler, S.J., Hoppe, H.: Texture mapping progressive meshes. In: SIGGRAPH 2001, pp. 409–416 (2001)
24. Shamir, A.: A survey on mesh segmentation techniques. Computer Graphics Forum **27**(6), 1539–1556 (2008)
25. Shapira, L., Shamir, A., Cohen-Or, D.: Consistent mesh partitioning and skeletonisation using the shape diameter function. Vis. Comput. **24**(4), 249–259 (2008)
26. Sheffer, A., Praun, E., Rose, K.: Mesh parameterization methods and their applications. Foundations and Trends in Computer Graphics and Vision (FTCGV) **2(2)**, 64 (2007)
27. Tierny, J., Vandeborre, J.P., Daoudi, M.: Topology driven 3D mesh hierarchical segmentation. In: IEEE International Conference On Shape Modeling And Application (SMI) (2007)
28. Unnikrishnan, R., Pantofaru, C., Hebert, M.: Toward objective evaluation of image segmentation algorithms. IEEE Transaction on pattern analysis and machine intelligence **29(6)**, 929–944 (2007)
29. Zeckerberger, E., Tal, A., Shlafman, S.: Polyhedral surface decomposition with applications. Computers and Graphics **26(5)**, 733–743 (2002)
30. Zhang, H., Fritts, J., Goldman, S.: Image segmentation evaluation: A survey of unsupervised methods. Computer Vision and Image Understanding **110**, 260–280 (2008)