VOL. 12, NO. 23, DECEMBER 2017 ISSN 1819-6608
ARPN Journal of Engineering and Applied Sciences

www.arpnjournals.com

AN OVERVIEW OF EXISTING EVALUATION METRICS FOR 3D MESH SEGMENTATION

Khadija Arhid[1], Mohcine Bouksim[1], Fatima Rafii Zakani[1], Mohamed Aboulfatah[2] and Taoufiq Gadi[1]
[1]Laboratory of Informatics, Imaging and Modeling of Complex Systems (LIIMSC),Faculty of Sciences and Techniques, Hassan 1st University Settat, Morocco
[2]Laboratory of Analysis of Systems and Treatment of Information (LASTI),Faculty of Sciences and Techniques, Hassan 1st University Settat, Morocco
E-Mail: khadija.arhid@gmail.com

ABSTRACT

The evaluation of mesh segmentation has received a great deal of attention since 3D mesh segmentation is an essential step in many mesh operations. For this reason, notable efforts have been made towards a better evaluation of mesh segmentation methods, and one of the most popular works is the benchmark of Chen et al., which allows a quantitative evaluation of mesh segmentation algorithms. Based on the given data sets, which comprise manual and automatic segmentations, many evaluation metrics have been proposed recently. In this context, we present in this study an overview of the existing similarity metrics and new ones proposed in our previous works addressing the problem of evaluating 3D mesh segmentation by describing each method and giving an extensive study and experimental comparison of them.

Keywords: 3D mesh, 3D mesh segmentation, evaluation metric.

1. INTRODUCTION

3D mesh segmentation is a common pre-processing step in many applications in 3D shape analysis, such as compression [1], skeleton extraction [2], deformation and many others. The reason why several segmentation algorithms have been presented in the literature [3]-[5]. Nevertheless, it still difficult to evaluate whether one method generates more accurate segmentations than another, whether it be for a particular 3D model or a set of models, or more generally, for a whole class of models. The selection of the adequate segmentation method for a particular problem is often based on tests and evaluation the reason why the assessment of 3D mesh segmentation algorithms becomes an important subject which attracts more and more researchers.

The key idea behind an assessment method is to be able to evaluate and classify segmentation algorithms according to their quality. This ability is crucial while the study of segmentation problem since it let the user choose among many possible segmentation algorithm which one is more suitable to use for a particular case, and also allows the test and the evaluation of new segmentation method and compare them with existing ones.

Lots of progress have been made during the past few years in the assessment of segmentation methods; we can mention mainly the works of Chen et al. [6] and Benhabiles *et al*. [7], [8] who have proposed two pioneer works to evaluate the performance of segmentation algorithms. Both works offer a benchmark to study the quality of a segmentation algorithm, by comparing each automatic segmentation with a reference one called the ground truth segmentation. The reference segmentation is the human perception (or how would a human segment the tested 3D model). The authors also propose a set of similarity measures to evaluate the segmentation algorithms by measuring the consistency between the reference segmentations and those obtained by automatic algorithms on the same models using a set of distances or metrics. Other recent works have been published with new evaluation methods will be discussed in the next section of this paper.

The aim of the present paper is to provide a comparative study between well-known evaluation methods and new ones. We try to enhance previous work by comparing the performance of the assessment methods through different tests. Last but not least, we conclude with a discussion where we expose the advantages of using evaluation method and future challenges in the assessment fields.

This paper is organized as follows. In section 2, we present the most common evaluation methods used in the literature along with the newest ones. Section 3, is reserved to the experimental tests where we will compare the performance of the exposed assessment methods and test their discriminative power. Finally, a conclusion with some perspectives will end this paper.

2. A STUDY OF WELL-KNOWN AND NEW EVALUATION METHODS

Mesh segmentation and its performance evaluation are very challenging tasks to achieve. Due to its importance, a significant number of works can be found in the literature treating these subjects. In this section, we will define the properties that define a reliable evaluation method, and then we will review the measures that have been proposed to address the assessment methods of segmentation algorithms. The authors in [8] expose a set of requirements properties for a reliable measure of mesh segmentation similarity which are:

- **No degenerative cases:** it means that the resulting score must be proportional to the similarity degree

between an automatic and a ground-truth segmentation of the same model

- **Tolerance to refinement:** a reliable segmentation measure has to be invariant to the granularity differences in segmentation because we can have two segmentations, one is coarse, and the other is finer and yet still consistent segmentations.
- **Cardinality independence:** this means that the two compared segmentations can have different numbers of segments and a different number of faces/vertices in each segment.
- **Tolerance to cut boundary imprecision:** when segmenting a 3D model, we can have a lot of segmentations for the same model that are the same, but with a slight difference between boundaries, hence a good evaluation measure has to accommodate this imprecision of cut boundaries.
- **Multiple ground-truth:** since each operator can give different ground truth segmentations for the same 3D model, the used evaluation measure should be able to compare each automatic segmentation with all the ground truth segmentations available for the tested 3D model.
- **Meaningful comparison:** the obtained score should represent the quality of the segmentation by giving a score of similarity/dissimilarity between the automatic and a set of ground truth segmentation. Finally, based on this score, it should be easy for us to know which segmentation algorithm is suitable to use with which kind of 3D model.

We can classify the existing evaluation method to four categories:

- **Boundary matching:** boundary matching metrics compute the difference between cuts of different segmentation (e.g., Cut Discrepancy)
- **Region differencing metrics:** region differencing metrics measure the consistency degree between two different regions (e.g., Hamming Distance, Consistency Error, Overlap index).
- **Non-parametric tests:** this kind of method computes the consistency of labels of the same face (or vertex) in two different compared segmentation (e.g., Rand Index, 3D Probabilistic Rand Index).
- **Information theory:** as the name of this category mention this kind of methods are based on the information theory (e.g., Adaptive Entropy Increment)

In the remainder of this section, we will describe and detail the well-known and the new proposed evaluation methods.

- **Hamming Distance** (HD) [6]: The Hamming distance computes the difference between two segments one from the automatic segmentation S_a, which we want to evaluate, and its closest segment from the ground truth segmentation S_g. The Hamming distance is defined as follows:

$$D_H(S_a, S_g) = \frac{1}{2}\left(\frac{D_H(S_a \to S_g)}{\|S\|} + \frac{D_H(S_g \to S_a)}{\|S\|}\right) \quad (1)$$

Where $\|S\|$ denote the cardinality of the whole mesh (number of faces or vertices) and $D_H(S_a \to S_g)$ is the directional function of the Hamming distance defined as follows:

$$D_H(S_a \to S_g) = \Sigma_i \| R_a^i \setminus R_g^{i_t} \| \quad (2)$$

"\" denote the set difference operator. R_a^i is the *i*-th segment from the segmentation S_a and $R_g^{i_t}$ is the closest to R_a^i from S_g obtained by:

$$i_t = \max_k \left\| R_a^i \cap R_g^k \right\| \quad (3)$$

- **Consistency Error** (CE) [6], [9]: this measure is divided into the Global Consistency error (GCE) and the Local Consistency error (LCE), defined as:

$$GCE(S_a, S_g) = \frac{1}{N} \min\left\{\begin{matrix}\sum_i L_{3D}(S_a, S_g, f_i), \\ \sum L_{3D}(S_g, S_a, f_i)\end{matrix}\right\} \quad (4)$$

$$LCE(S_a, S_g) = \frac{1}{N} \sum_i \min\left\{\begin{matrix}L_{3D}(S_a, S_g, f_i), \\ L_{3D}(S_g, S_a, f_i)\end{matrix}\right\} \quad (5)$$

Where

$$L_{3D}(S_a, S_g, f_i) = \frac{\|R(S_a, f_i) \setminus R(S_g, f_i)\|}{\|R(S_a, f_i)\|} \quad (6)$$

S_a and S_g are two segmentations, N is the number of faces in the mesh $L_{3D}(S_a, S_g, f_i)$ is the local refinement error, $R(S, f_i)$ is the segment in segmentation

S that contains a face f_i and the operator "\" define the set difference operator.

- **Cut Discrepancy** (CD) [6]: The Cut Discrepancy is a metric that measures how the segment boundaries are close one to another, this is done by summing the geodesic distances of the cuts of one segmentation against another.

Let's consider S_a, and S_g two segmentations and C_1, C_2 are their respective sets of points on the segment boundaries. The Cut Discrepancy is defined as follows:

$$CD(S_1,S_2)=\frac{mean\{d_G(p_1,C_2),\forall p_1\in C_1\}+mean\{d_G(p_2,C_1),\forall p_2\in C_2\}}{avgRadius}, \quad (7)$$

$$d_G(p_1,C_2)=\min\{d_G(p_1,p_2),\forall p_2\in C_2\}, \quad (8)$$

- **Rand index** (RI) [6], [10]: The Rand Index measures the likelihood that a pair of faces are in the same segment or in different segments in the two compared segmentations. Let consider Sa, S_g as two segmentations and N be the number of faces in the 3D mesh; the Rand Index is defined as follows:

$$RI(S_a,S_g)=\binom{2}{N}^{-1}\sum_{i,j,i<j}\left[C_{ij}P_{ij}+(1-C_{ij})(1-P_{ij})\right] \quad (9)$$

Where $C_{ij}=1$ if the faces i and j belong to the same segment in the segmentation S_a, $P_{ij}=1$ if the faces i and j belong to the same segment in the segmentation S_g. Consequently $C_{ij}P_{ij}=1$ if i and j belong to the same segment in both segmentations.

- **The 3D Normalized Probabilistic Rand Index** (3DNPRI) [8]: the 3DNPRI is a measure based on the Rand Index, but it adds the ability to compare an automatic segmentation with a set of ground truth segmentation for the same 3D objects. Let S_1 be the automatic segmentation and $\{S_k\}$ the set of ground truth segmentations; the 3DPRI is defined as:

$$3DPRI(S_a,\{S_k\})=\frac{1}{\binom{N}{2}}\sum e_{ij}p_{ij}+(1-e_{ij})(1-p_{ij}) \quad (10)$$

$e_{ij}=1$ means that the vertex *i* and *j* belong to the same segment of the automatic segmentation S_a and p_{ij} is defined as the probability of two vertices belonging to the same segment in the ground truth set. The authors also proposed a normalized version of the 3DPRI to increase the discriminative power of the method; the 3DNPRI is defined as:

$$3DNPRI(S_a,\{S_k\})=\frac{3DPRI(S_a,\{S_k\})-E[3DPRI]}{1-E[3DPRI]} \quad (11)$$

Where $E[3DPRI]$ computes the 3DPRI on the whole dataset using random segmentations.

- **Adaptive Entropy Increment** (AEI) [11]: This method is based on the entropy concept from information theory. The method starts by calculating a baseline, which corresponds to the entropy of all the different ground-truth, then the automatic segmentation is added, and the entropy is recalculated. The increment from the baseline to the new value is adopted to evaluate the automatic segmentation. The entropy $H(G_1,...,G_n)$ of n ground truth segmentation G_i is defined as:

$$H(G_1,...,G_n)=-\sum P(G_1,...,G_n)\log(P(G_1,...,G_n)) \quad (12)$$

Let consider A as an automatic segmentation; the normalized entropy increment is defined as follows:

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

Finally, the Adaptive Entropy Increment (AEI) is defined on *N* random segmentations A_r:

$$AEI=\frac{\Delta H}{E(\Delta H)} \quad (14)$$

$$E(\Delta H)=\frac{1}{N}\sum_{r=1}^{N}\Delta H(G_1,...,G_n,A_r) \quad (15)$$

- **Similarity Hamming Distance** (SHD) [11]: The SHD is based on the first Hamming Distance. It adds the ability to compare one automatic segmentation to a set of ground truth segmentation, by associating each segment from the automatic segmentation with the closest segment from the set of the ground truth segmentations. The SHD is defined by the following formula:

$$SHD=\beta.EMD_{D2}(R_a^i,R_{S_k}^i)+(1-\beta).dist(R_a^i,R_{S_k}^i) \quad (16)$$

Where EMD_{D2} is the earthmover's distance between the D2 distributions of the two segments $dist()$ denotes the Euclidean distance between the centers of the two segments and β is a weight.

- **Ultimate Measurement Accuracy** (UMA) [12]: the UMA aims to evaluate the quality of segmentation by means of measuring and calculating the target feature. Let S be a 3D boundary mesh, $S_1 = \{S_1^1, S_1^2, ..., S_1^m\}$, the set of sub-grid generated by manual interactive segmentation and $S_2 = \{S_2^1, S_2^2, ..., S_2^n\}$, the automatic segmentation. The author's define the Form Factor of arbitrary sub-grid S_1^i in S_1 as follows:

$$FF(S_1^i) = \sum_{a=1}^{N_e} L(S_1^i, e_a) \Big/ \sum_{b=1}^{N_f} A(S_1^i, f_b) \quad (17)$$

Where e_a and f_b are the arbitrary boundary edge and face in sub-grid S_1^i, $L(S_1^i, e_a)$ is the length of edge e_a, $A(S_1^i, f_b)$ is the area of f_b, N_e and N_f are respectively the number of boundary edges and faces of S_1^i.

Similarly, the Form Factor of arbitrary sub-grid S_2^i in S_2 as follows:

$$FF(S_2^i) = \sum_{a=1}^{N_e} L(S_2^i, e_a) \Big/ \sum_{b=1}^{N_f} A(S_2^i, f_b) \quad (18)$$

Finally, the UMA of the segmentation S_2 is calculated as follows:

$$UMA(S_1, S_2) = \frac{1}{m} \sum_{i=1, j=p(i)}^{m} \frac{FF(S_1^i) - FF(S_2^i)}{\|S_1^i\|} \quad (19)$$

Where $\|S_1^i\|$ is the area of S_1^i, $p(i)$ represents the correspondence sub-grid of S_1^i.

- **The Weighted Dice Coefficient** (WDC) [13]: Zakani et al. in [13] propose the use of a new evaluation metric based on the Dice's coefficient [14] also known as Sorensen–Dice's [15] coefficient. This coefficient has been extensively used to compute the similarity degree between two samples in many fields like brain images [16] and string similarity [17]. Due to its performance, the authors proposed an adaptation of this coefficient to be able to provide a relevant evaluation of 3D mesh segmentation algorithms. Let's consider $S_a = \{R_a^1, .., R_a^k\}$, and $S_g = \{R_g^1, .., R_g^l\}$ two segmentations, the Weighted Dice Coefficient between two segments is defined as follows:

$$WDC(R_a^i, R_g^{i_t}) = \frac{2 \times Surface(R_a^i \bigcap R_g^{i_t})}{Surface(R_a^i) + Surface(R_g^{i_t})} \quad (20)$$

Where:

- i_t is the index of the closest segment from S_g to R_a^i got by:

$$i_t = argmax_k \| R_a^i \cap R_g^i \| \quad (21)$$

- $Surface(R)$ is the surface of the segment R composed by n faces calculated by:

$$Surface(R) = \sum_{i=1}^{n} surface(f_i) \quad (22)$$

Finally, the WDC between the segmentations S_1 and S_2 is computed as follows:

$$WDC(S_a, S_g) = \frac{\sum_{i=1}^{k} WDC(R_a^i, R_g^{i_t})}{m} \quad (23)$$

The main advantage of this metric is its ability to treat irregular as well as regular meshes and give a relevant evaluation of both kinds of meshes taking into account their regularity.

- **The Weighted Kulczynski dissimilarity index** (WKD) [18]: the WKD index was recently proposed by Zakani *et al*[18]. This metric is based on the well-known Kulczynski index [19], which is a fundamental concept in the analysis of presence-absence data, it is used in many fields such as ecology, biology and image similarity. The WKD has the particularity to be able to compare an automatic segmentation with a set of ground truth segmentations and takes into account the regularity of 3D segmented meshes. The WKD between an automatic segmentation S_a and a set of ground truth $S_n = \{S_{g_1}, S_{g_2}, ..., S_{g_n}\}$ is defined by:

$$WKD(R_a^i, R_s^{i_t}) = 1 - \frac{1}{2}\left(\frac{surf(R_a^i \cap R_s^{i_t})}{surf(R_a^i \cap R_s^{i_t}) + surf(R_a^i \setminus R_s^{i_t})} + \frac{surf(R_a^i \cap R_s^{i_t})}{surf(R_a^i \cap R_s^{i_t}) + surf(R_s^{i_t} \setminus R_a^i)}\right) \quad (24)$$

Where:

- $Surf(R)$ is the surface of the segment R defined in formulas number (22).
- i_t is the index of the closest segment from the set $\{S_n\}$ to R_a^i got by:

$$i_t = max_j(max_k \parallel R_a^i \cap R_{g_j}^k \parallel) \quad (25)$$

Finally, the WKD between the automatic segmentation and the set of ground truth is computed as follows:

$$WKD(S_a, \{S_n\}) = \frac{\sum_{i=1}^{m} WKD(R_a^i, R_{g_j}^{i_t})}{m} \quad (26)$$

With m is the number of segments in the automatic.

- **Weighted Sokal-Sneath Distance** (WSSD [20]): Arhid *et al.*[20] proposed to use the Sokal-Sneath [21] distance communally known in many fields like biological systematics and numerical taxonomy, to compute the dissimilarity between two segmentations. This distance can evaluate an automatic segmentation by comparing it to more than one ground truth segmentations. The WSD is defined as follows:

$$WSSD(R_a^i, R_g^j) = 1 - \frac{S(R_a^i \cap R_g^j)}{S(R_a^i \cap R_g^j) + 2 \times S(R_a^i \setminus R_g^j) + 2 \times S(R_g^j \setminus R_a^i)} \quad (27)$$

$$WSSD(S_a, \{S_n\}) = \frac{\sum_i WSSD(R_a^i, R_g^{i_t})}{m} \quad (28)$$

Where:

- $S(R)$ is the surface of the segment R defined in formulas number (22).
- i_t is the index used to associate each segment from the automatic segmentation with segments from the set of ground truth segmentation (formulas. (25))

- **Sampling theory**[22]: a recent work presented by Arhid et al. [22] proposed to use the sampling theory [23] to evaluate two segmentations. The main advantage of this method is that it focus on the execution time, which is reduced considerably compared to other evaluation methods. It is defined as follows:

$$Diff(Samp^i, R_a^i) = \frac{\left\| Samp^i \setminus R_g^{i_t} \right\|}{\left\| Samp^i \right\|} \quad (29)$$

Where:

- "\" is the set difference operator;
- $\|R\|$ represent the size of the set R, or the total area of all faces in a set R;
- $Samp^i$ is an extracted sample of vertices or faces from the segment R.

- **The Jaro distance** (d_j) [24]**:** another distance was proposed by Bouksim et al. [24] this measure is based on the Jaro distance [25], [26], which is a well-known distance in many fields. The authors proposed an adaptation of this distance to be able to provide a dissimilarity score between an automatic segmentation and many ground truth segmentations. The Jaro distance is defined as follows:

$$d_{j3D}(R_g^{i_t}, R_a^i) = 1 - \left\{ \begin{array}{ll} 0 & if\ m = 0 \\ \frac{1}{4}\left(\frac{m}{|R_g^{i_t}|} + \frac{m}{|R_a^i|} + \frac{S(R_g^{i_t} \cap R_a^i)}{S(R_g^{i_t})} + \frac{S(R_a^{i_t} \cap R_g^i)}{S(R_g^i)} \right) & or \end{array} \right\} \quad (30)$$

Where:

- $S(R)$ is the surface of the segment R defined in formulas number (22).
- m is the number of matching faces between $R_g^{i_t}$ and R_g^i;
- i_t denotes the index of the closest segment from R_a to $\{R_n\}$ it is defined by the following equation $i_t = \min_i(\min_j(d_{j3D}(R_a^i, R_g^j)))$;
- $|R|$ is the size of the segment R (number of faces).

- **The weighted Ochiai index** (WI_O) [27]**:**proposed by Bouksim *et al.*[27], based on Ochiai index also known as Ochiai-Barkman, Otsuka-Ochiai or Driver–Kroeber index. This index is one of the most used indexes in many fields and can easily be used to compare or extract the similarity between two samples of data. Considering that a segment is composed of faces or vertices are a data sample, the authors proposed an adaptation of this index to be applied in the evaluation of 3D mesh segmentation algorithms. Let us consider G as a ground truth segmentation, where R_G^j is the *j*-th segment in the segmentation G and S an automatic segmentation, where R_S^l is the *l*-th segment in the segmentation S .The weighted Ochiai index (WI_O) is formulated as follows:

$$WI_O(R_S^l, R_G^j) = 1 - \frac{Surf(R_S^l \bigcap R_G^j)}{\sqrt{Surf(R_S^l) * Surf(R_G^j)}} \quad (31)$$

whith $Surface(R)$ is the surface of the segment R and it is defined as the sum of surfaces of all the faces composing the segment R.

The WI_O between an automatic segmentation S_a and a set of ground truth $S_n = \{S_{g_1}, S_{g_2}, ..., S_{g_n}\}$ is defined as:

$$WI_O(S, \{S_n\}) = \frac{\sum_l WI_O(R_S^l, R_G^{l_t})}{N} \quad (32)$$

Where:

- l_t represent the index of the closest segment from the set of ground truth segmentations $\{S_n\}$ to R_S^l from S, this is obtained using the formula:

$$l_t = \max_i(ma\,x_j(\| R_S^l \bigcap R_{G_i}^j \|)) \quad (33)$$

- N is the number of segments in the automatic segmentation.

This measure allows two significant advantages; it is appropriate for comparing automatic segmentation provided by the algorithm to evaluate with a set of ground truth segmentations. And, it is also designed to actually capture the real segmentation quality of both regular and irregular meshes.

- **The Normalized Weighted Levenshtein Distance (NWLD)** [28]**:** Zakani *et al*. [28] has recently proposed a new evaluation method based on the Levenshtein Distance to be used for the assessment of 3D mesh segmentation algorithms. This method aims to compute the optimal operations needed to convert a segment to another. Let's consider R_A^i as the ith segment of the automatic segmentation A and $R_{S_j}^{i_t}$ is the best correspondent segment got from all the available set of ground truth segmentation $S = \{S_1, ..., S_n\}$ obtained by $i_t = max_j(max_k \left\| R_A^i \bigcap R_{S_j}^k \right\|)$. The authors define the Weighted Levenshtein Distance (WLD) in the equation (37). Where:

- $\delta = \max(A(f_{m-1}), A(f_{k-1}))_{(R_A^i(k) \neq R_{S_j}^{i_t}(m))}$ is an indicator function equal to: $\max(A(f_{m-1}), A(f_{k-1}))$ and 0 otherwise.

- k is the face index from the segment R_A^i.
- m is the face index from the segment $R_{S_j}^{i_t}$
- $A(f_i)$ represent the area of the face f_i.

$$WLD_{R_A^i, R_{S_j}^{i_t}}(k,m) = \begin{Bmatrix} 0 & if & k = m = 0 \\ k & if & m = 0 \ and \ k > 0 \\ m & if & k = 0 \ and \ m > 0 \\ \min \begin{Bmatrix} WLD_{R_A^i, R_{S_j}^{i_t}}(k-1, m) + A(f_{k-1}) \\ WLD_{R_A^i, R_{S_j}^{i_t}}(k, m-1) + A(f_{m-1}) \\ WLD_{R_A^i, R_{S_j}^{i_t}}(k-1, m-1) + \delta \end{Bmatrix} \end{Bmatrix} \quad (34)$$

Finally, the authors define The Normalized Weighted Levenshtein Distance to be the results of the WLD divided by the maximum score gotten by considering that the two compared segment are entirely different:

$$NWLD_{R_A^i, R_{S_j}^{i_t}} = \frac{WLD_{R_A^i, R_{S_j}^{i_t}}}{\max(WLD_{R_A^i, R_{S_j}^{i_t}})} \quad (35)$$

The main advantage offered by this metric is its discriminative power on regular as well as irregular meshes and its ability to take into consideration the regularity of the 3D mesh and its capability to compare an automatic segmentation to set of ground truth segmentations.

3. EXPERIMENTAL STUDY

To discuss the behavior of the studied metrics, we perform five experiments. The first test examines the ability of these measures to treat irregular as well as regular meshes. The second test underlines the importance of multiple ground truth comparison by showing the different behaviors of the studied metrics regarding single ground truth segmentation and a set of reference segmentations. The next study highlights the discriminating power of these metrics in detecting different segmentation qualities. While the fourth experiment, we compare the discriminative strength of these metrics on simple and complex models. As a final test, we evaluate the robustness of the tested measures against hierarchical segmentations. All the 3D used models were taken for the Princeton Benchmark for 3D Mesh Segmentation made by Chen *et al*. [6].

The first property to test is the ability of each evaluation metric to evaluate objects with both regular and irregular meshes, to achieve this task we took the object 26 (Figure-2) from the Princeton's benchmark which possesses irregular meshes. In Figure-2 (a) present the 3D mesh, while (c) and (d) are two manual segmentations of

this object. These two segmentations (c, d) are constructed manually and differs from the ground truth segmentation with the same number of error faces, but the difference relies on the face's surface which leads to a significant difference between them in term of segmentation quality. By applying all the studied evaluation metrics, we got the results exposed in Table-1. We can see from the obtained results that the AEI, RI, and GCE/LCE give the same error for both segmentations and consequently didn't succeed to catch the real quality of the two tested segmentations while the measures CD, HD, WKD, WIo, WDC, WSSD, NWLD, and Dj3D give a proportional error scores to these segmentation qualities. As the regularity of the mesh affects the quality of the segmentation result, a relevant evaluation measure should take into account this aspect of the 3D mesh.

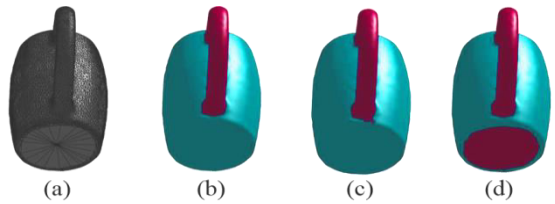

Figure-1. Example of an object taken from the benchmark of Chen *et al*. [6] Possessing an irregular mesh.

Table-1. Evaluation of both segmentations (c) and (d) figured in Figure-2.

	(c)	**(d)**
CD	0,0280	0,0959
HD	0,0003	0,0426
RI	0,0015	0,0015
GCE	0,0015	0,0015
LCE	0,0007	0,0007
WSSD	0,0045	0,2853
WDC	0,0011	0,1060
WIo	0,0015	0,1185
WKD	0,0018	0,1209
D_{j3D}	0,0012	0,0961
NWLD	0,0022	0,1817
AEI	0.0155	0.0155

In this experiment, we underline the utility of comparing an automatic segmentation to a set of ground truth segmentations instead of doing a one to one comparison. For this task, we choose the object cup shown in Figure-3, (a) is the automatic segmentation done with the Randomized Cuts [29] method while (b), (c) and (d) are the set of ground truth segmentations, we can see that none of the ground truth segmentations matches the automatic segmentation. However, we can notice that the automatic segmentation presents a combination of all the available reference segmentations and also offers a relatively good segmentation quality. Firstly we apply the evaluation metrics designed to compare an automatic segmentation with a set of ground truth segmentation to evaluate the quality of the given automatic segmentation by comparing it to only one ground truth and secondly we do a multiple ground-truth evaluation; all the results are shown in Table-2. As mentioned before the automatic segmentation presents a relatively good segmentation, so the obtained scores are expected to be low. By analysing the scores obtained with a one to one comparison we can notice that they didn't reflect the real quality of the automatic segmentation since we got a higher error for this segmentation, while the results got by multiple comparisons give a lower error for the automatic segmentation which illustrates the good quality of the given segmentation. Consequently, we can deduce that an efficient evaluation metric should quantify the consistency between an automatic and multiple ground truth segmentations.

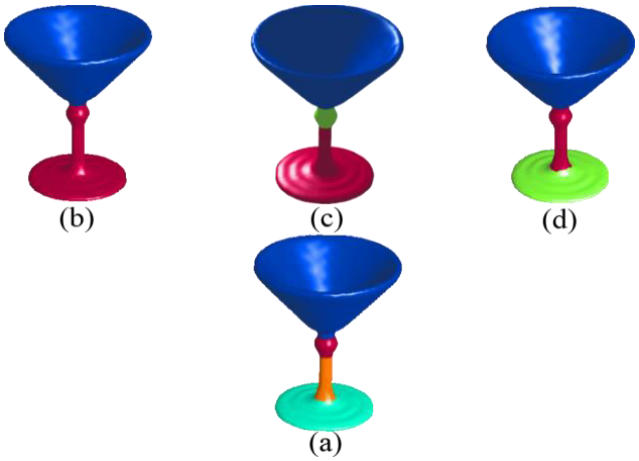

Figure-2. The object cup with an automatic segmentation (a) and three ground truth segmentations(b,c,d).

Table-2. The obtained scores for a one to one evaluation and a multi-ground truth evaluation.

	All ground truth	**(b)**	**(c)**	**(d)**
WDC	0,1014	0,2118	0,2583	0,4393
WKD	0,0659	0,1394	0,1532	0,2512
WIo	0,0938	0,1669	0,2050	0,3712
WSSD	0,2030	0,3573	0,3764	0,5601
D_{j3D}	0,0911	0,1547	0,2055	0,2936
NWLD	0,1309	0,2501	0,2678	0,4762
AEI	0,0526	0,1322	0,2175	0,3571

The third experiment aims to compare the discriminative power of the studied methods by evaluating different segmentations quality. Figure-3, shows the six chosen segmentations, the segmentations (a) and (b) represent an excellent segmentation quality (c) and (d) medium one and finally (e) and (f) have a very bad segmentation quality. 4igure-5shows the error values

obtained using eleven evaluation methods. As can be seen from the results AEI, WIo, D_{j3D}, WKD, WSSD, WDC and NWLD succeed in providing scores which reflect the quality of the segmentation, since they give a high error score for the two bad segmentation (e) and (f) reasonable error score for the medium quality (c) and (d) and a very low score for the good segmentation (a) and (b). Other methods like RI, GCE, LCE or CD didn't succeed in detecting these qualities since they provide inconsistent scores with the quality of the segmentation.

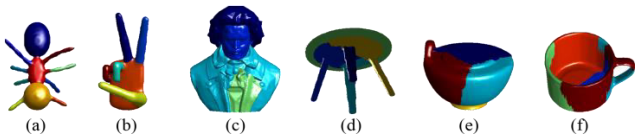

Figure-3. Tree groups of segmentation qualities (a) and (b) are good, (c) and (d) are relatively good and (e) and (f) present a bad segmentation quality.

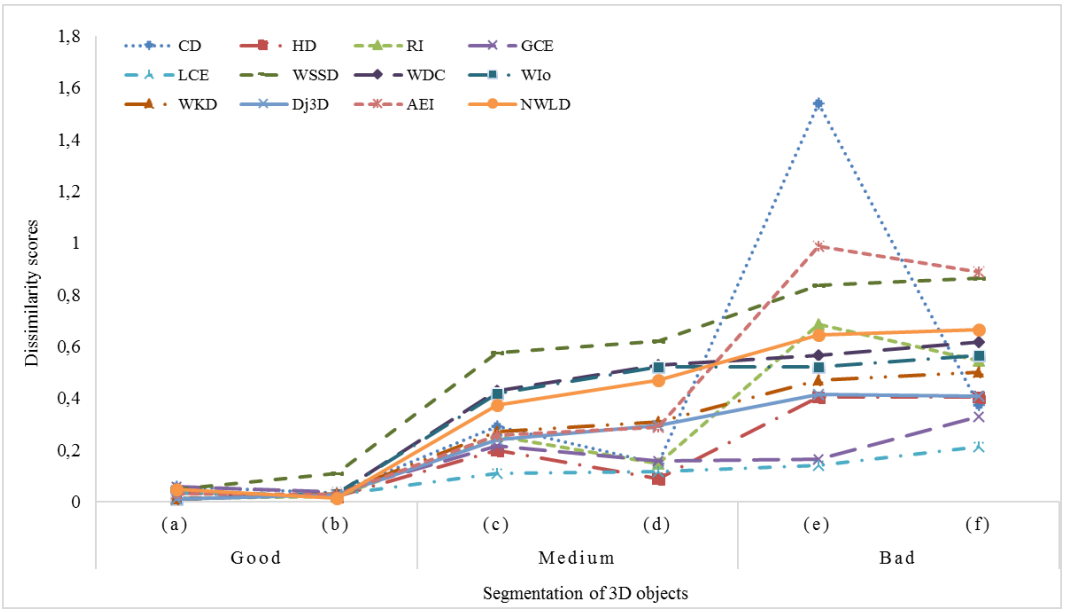

Figure-4. The dissimilarity scores using the studied measures for the objects shown in Fig. 4.

In this next test, we study the discriminative power of the different metrics on simple and complex models. To achieve this task, we perform two experiments; the first one aims to explore the behavior of each metric on simple models by applying the different metrics to evaluate six segmentations of a simple model, we choose to do it with a cup model. As can be seen in Figure-6 the four first segmentations are consistent (using the segmentation method proposed by our research team [30]) while the two last represent unreachable segmentations. In the second experiment, we choose a dog model wich is an example of a complex model and seven automatic segmentations representing five segmentations with a high quality and two inconsistent segmentations (Figure-6), and then we compute the dissimilarity error using different measures. The obtained results are shown in Figure-7and Figure-8. The methods including AEI, WSD, WKD, WIo, WDC and DJ3D have a good behavior not only on simple models but also on complex models. They give a low scores for the consistent segmentations and high scores for the unreachable segmentations. While the metrics CD, RI, HD, GCE, and LCE fail to give logical scores for some segmentation regardless of their quality.

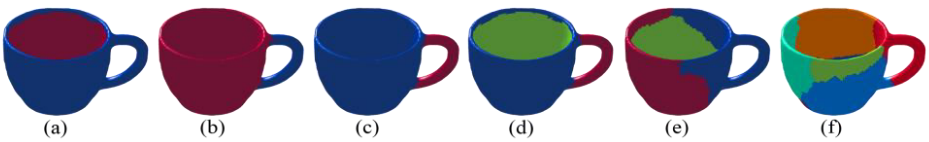

Figure-5. A simple model cup with four consistent segmentations (a-d) and two bad segmentations (e-f).

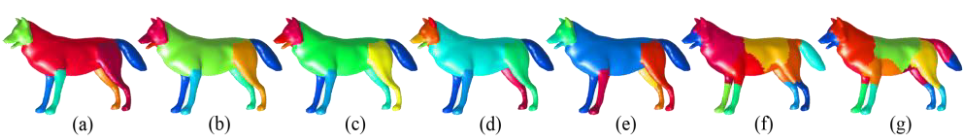

Figure-6. Example of a complexe object taken from the Princeton benchmark accompanied

with five good segmentations (a-e) and two inconsistent segmentations (f-g).

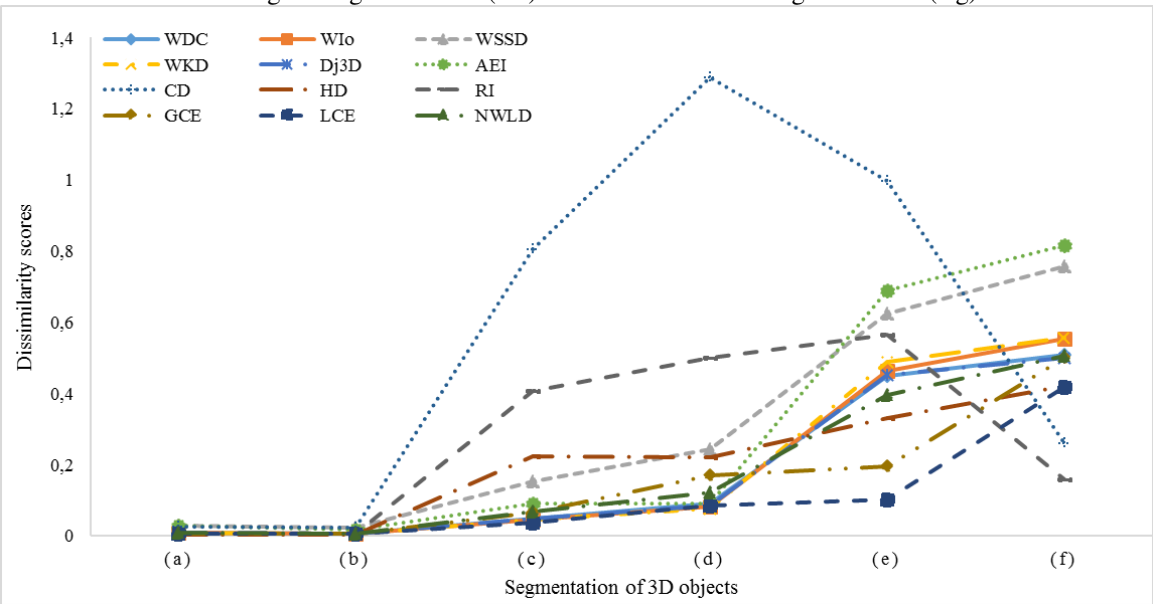

Figure-7. The obtained results using several metrics on simple model presented in Figure-6.

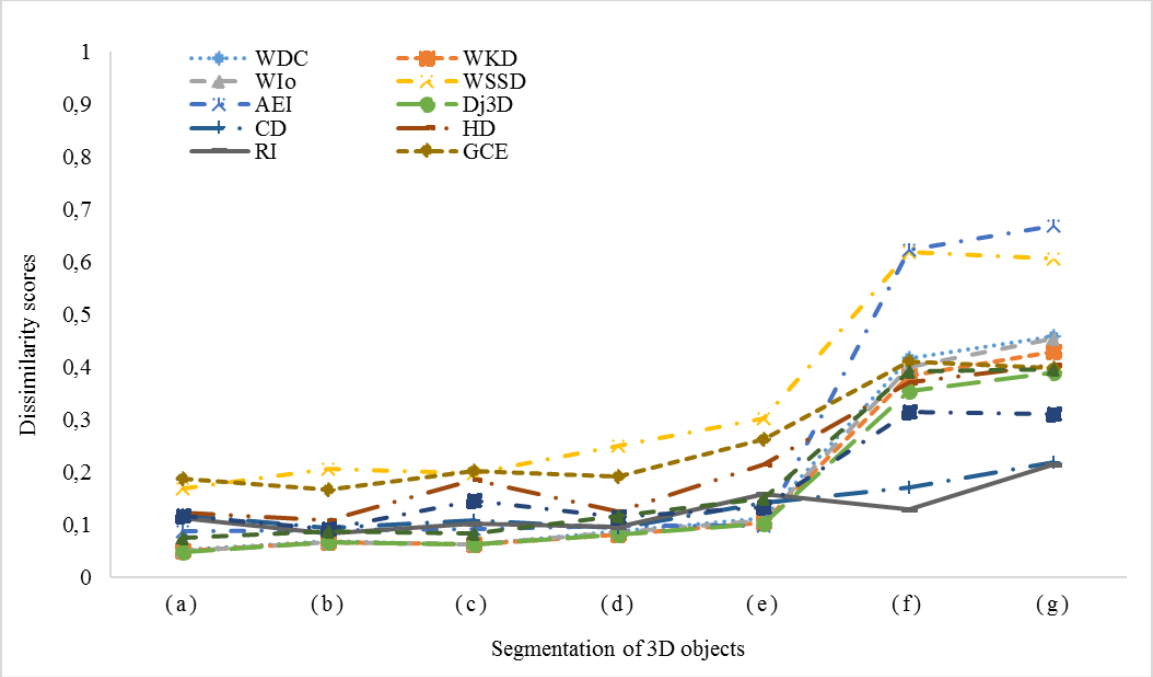

Figure-8. The obtained results using several metrics on the complex model shown in Figure-7.

The fifth and the last experiment studies the sensitivity of each index regarding hierarchical segmentations. To this end, we take the object horse got from the benchmark of Chen *et al*. [6] with nine levels of segmentation gotten using Randomized Cut [29] algorithm Figure-9, and we apply all the evaluation metrics cited in this work to evaluate these segmentations. As can be deduced from the obtained results in Table-3, the AEI, WSD, WIo, WDC, WKD, D_{j3D},and the NWLD seem to be relatively invariant to hierarchical refinement since they give a very close scores for all level of the evaluated segmentations wich means that these metrics can tolerate these refinements and give consistent results while the other evaluation metrics failed to do the same task since there are less stable against segmentations with different hierarchical structures.

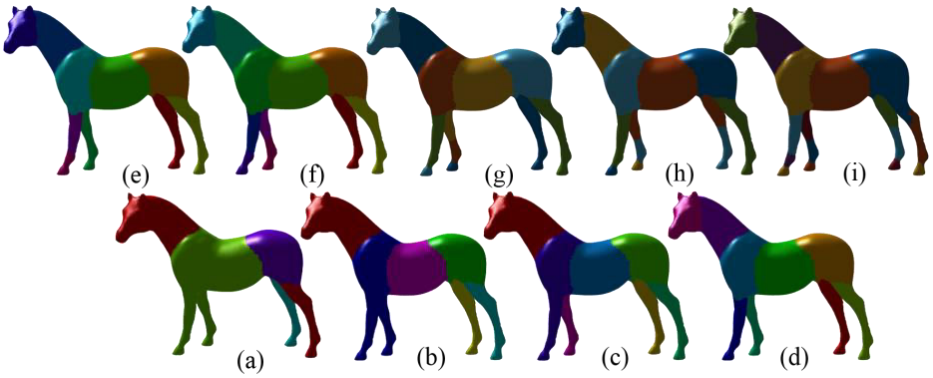

Figure-9. An object is taken from the benchmark of Chen et al. [6] with nine levels of segmentation.

Table-3. The obtained results using the cited metrics for the different levels of segmentations.

	AEI	WSD	WIo	WDC	WKD	Dj3D	CD	HD	RI	GCE	LCE	NWLD
(a)	0,09	0,18	0,11	0,11	0,10	0,10	0,36	0,25	0,32	0,32	0,17	0,11
(b)	0,12	0,23	0,17	0,18	0,10	0,10	0,31	0,32	0,33	0,29	0,15	0,18
(c)	0,12	0,25	0,17	0,17	0,11	0,10	0,21	0,27	0,25	0,23	0,09	0,19
(d)	0,12	0,23	0,14	0,14	0,09	0,09	0,10	0,20	0,08	0,27	0,18	0,15
(e)	0,14	0,21	0,17	0,17	0,12	0,12	0,06	0,08	0,04	0,13	0,09	0,18
(f)	0,17	0,25	0,21	0,22	0,16	0,15	0,15	0,28	0,19	0,25	0,19	0,19
(g)	0,15	0,22	0,18	0,20	0,14	0,13	0,11	0,24	0,10	0,31	0,19	0,19
(h)	0,17	0,21	0,21	0,22	0,16	0,16	0,11	0,25	0,10	0,30	0,19	0,19
(i)	0,19	0,25	0,20	0,21	0,15	0,15	0,10	0,25	0,10	0,28	0,18	0,18

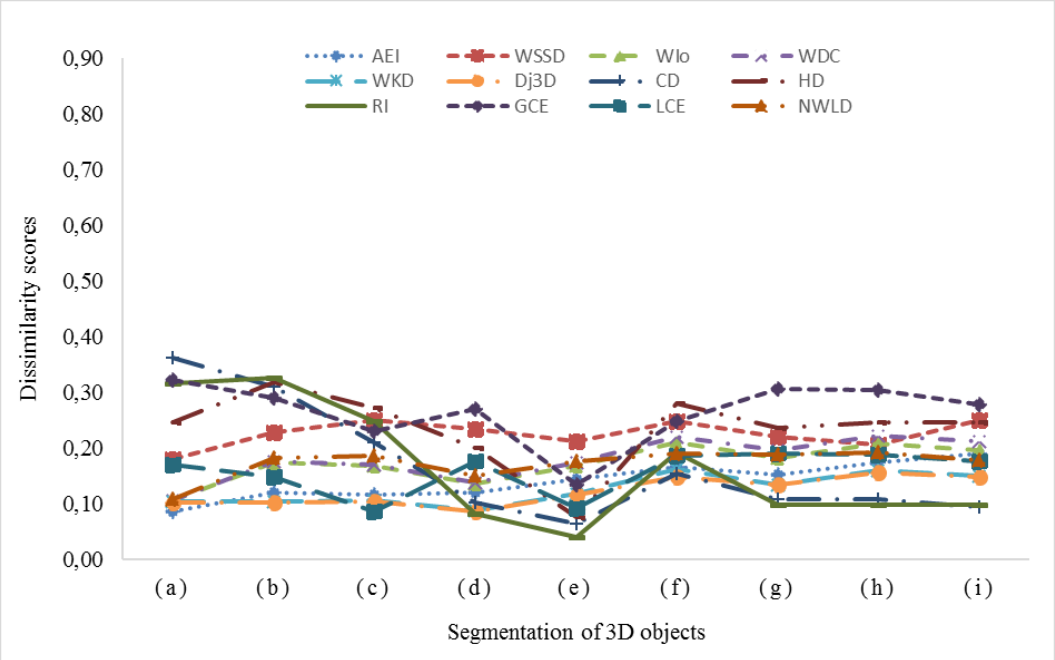

Figure-10. A graph showing the obtained results using the cited metrics for the different levels of segmentations.

CONCLUSIONS

In this paper, we have studied some of the existing evaluation metrics addressing the problem of 3D mesh segmentation performance assessment by discussing their contributions and comparing their performances. This study is done using the benchmark of Chen *et al.*[6] to evaluate the performance of the studied evaluation methods. After experiments, we can deduce that an evaluation method should take into consideration the regularity of the mesh in addition of doing a multi-ground truth comparison instead of doing one to one comparison to provide best results and a robust evaluation. As a perspective, we plan to explore unsupervised assessment methods where no reference segmentation is needed since it is not always available.

ACKNOWLEDGEMENTS

We would like to show our gratitude to Xiaobai Chen *et al*. for providing online (http://segeval.cs.princeton.edu/), 3D-mesh models along with their automatic and ground truth segmentations in addition to the source code of his evaluation methods. We are also immensely grateful to Liu and Hao Zhang *et al.* for providing us the source code of the AEI evaluation method

REFERENCES

[1] A. Maglo, G. Lavoué, F. Dupont, and C. Hudelot. 2015. 3D Mesh Compression: Survey, Comparisons, and Emerging Trends. ACM Computing Surveys. 47(3): 1-41.

[2] P. K. Saha and G. Sanniti di Baja. 2016. A survey on skeletonization algorithms and their applications.

[3] M. Attene, S. Katz, M. Mortara, G. Patane, M. Spagnuolo and A. Tal. 2006. Mesh Segmentation - A Comparative Study. IEEE International Conference on Shape Modeling and Applications 2006 (SMI'06). pp. 7-7.

[4] A. Shamir. 2008. A survey on mesh segmentation techniques. Computer Graphics Forum. 27(6): 1539-1556.

[5] P. Theologou, I. Pratikakis and T. Theoharis. 2015. A comprehensive overview of methodologies and performance evaluation frameworks in 3D mesh segmentation. Computer Vision and Image Understanding. 135: 49-82.

[6] X. Chen, A. Golovinskiy and T. Funkhouser. 2009. A benchmark for 3D mesh segmentation. ACM Transactions on Graphics. 28(3): 1.

[7] H. Benhabiles, G. Lavoué, J. P. Vandeborre, and M. Daoudi. 2010. A subjective experiment for 3D-mesh segmentation evaluation. In: 2010 IEEE International Workshop on Multimedia Signal Processing, MMSP2010. pp. 356-360.

[8] H. Benhabiles, J.-P. Vandeborre, G. Lavoué, and M. Daoudi. 2010. A comparative study of existing metrics for 3D-mesh segmentation evaluation. The Visual Computer. 26(12): 1451-1466.

[9] H. Benhabiles, J.-P. Vandeborre, G. Lavoue, and M. Daoudi. 2009. A framework for the objective evaluation of segmentation algorithms using a ground-truth of human segmented 3D-models. 2009 IEEE International Conference on Shape Modeling and Applications. pp. 36-43.

[10] W. M. Rand. 1971. Objective Criteria for the Evaluation of Clustering Methods. Journal of the American Statistical Association. 66: 846-850.

[11] Z. Liu, S. Tang, S. Bu and H. Zhang. 2013. New evaluation metrics for mesh segmentation. Computers and Graphics (Pergamon). 37(6): 553-564.

[12] X. Sun, L. Wang, X. Wang and X. Zhao. 2015. A Novel Quantitative Evaluation Metric of 3D Mesh Segmentation. pp. 621-628.

[13] F. R. Zakani, K. Arhid, M. Bouksim, M. Aboulfatah, and T. Gadi. 2016. New measure for objective evaluation of mesh segmentation algorithms. in 2016 4th IEEE International Colloquium on Information Science and Technology (CiSt). pp. 416-421.

[14] L. R. Dice. 1945. Measures of the Amount of Ecologic Association between Species. Ecology. 26(3): 297-302.

[15] T. J. Sørensen. 1948. A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on Danish commons, København.

[16] A. P. Zijdenbos, B. M. Dawant, R. A. Margolin and A. C. Palmer. 1994. Morphometric analysis of white matter lesions in MR images: method and validation. IEEE transactions on medical imaging. 13(4): 716-24.

[17] G. Kondrak, D. Marcu and K. Knight. 2003. Cognates can improve statistical translation models. in Proceedings of the 2003 Conference of the North

American Chapter of the Association for Computational Linguistics on Human Language Technology companion volume of the Proceedings of HLT-NAACL 2003-short papers - NAACL '03. 2: 46-48.

[18] F. R. Zakani, K. Arhid, M. Bouksim, T. Gadi and M. Aboulfatah. 2016. Kulczynski similarity index for objective evaluation of mesh segmentation algorithms. In: 2016 5th International Conference on Multimedia Computing and Systems (ICMCS). pp. 12-17.

[19] S. Kulczynski. 1927. Die Pflanzenassoziationen der Pieninen. Bulletin International de l'Academie Polonaise des Sciences et des Lettres, Classe des Sciences.

[20] K. Arhid, M. Bouksim, F. R. Zakani, T. Gadi and M. Aboulfatah. 2016. An objective 3D mesh segmentation evaluation using Sokal-Sneath metric. In: 2016 5th International Conference on Multimedia Computing and Systems (ICMCS). pp. 29-34.

[21] R. R. Sokal, P. H. A. Sneath, and others. 1963. Principles of numerical taxonomy. Principles of numerical taxonomy.

[22] K. Arhid, M. Bouksim, F. R. Zakani, M. Aboulfatah and T. Gadi. 2016. New evaluation method using sampling theory to evaluate 3D segmentation algorithms. In: 2016 4th IEEE International Colloquium on Information Science and Technology (CiSt). pp. 410-415.

[23] P. Minkkinen. 2004. Practical applications of sampling theory. Chemometrics and Intelligent Laboratory Systems. 74(1): 85-94.

[24] M. Bouksim, F. R. Zakani, K. Arhid, M. Aboulfatah and T. Gadi. 2016. New evaluation method for 3D mesh segmentation. In: 2016 4th IEEE International Colloquium on Information Science and Technology (CiSt). pp. 438-443.

[25] M. A. Jaro. 1989. Advances in record-linkage methodology as applied to matching the 1985 census of Tampa Florida.JASA: Journal of the American Statistical Society.84(406).

[26] M. A. Jaro. 1995. Probabilistic linkage of large public health data files. Statistics in Medicine. 14(5-7): 491-498.

[27] M. Bouksim, F. Rafii Zakani, K. Arhid, M. Aboulfatah and T. Gadi. 2016. Evaluation of 3D Mesh Segmentation Using a Weighted Version of the Ochiai Index. In: 2016 IEEE/ACS 13th International Conference of Computer Systems and Applications (AICCSA).

[28] F. Rafii Zakani, K. Arhid, M. Bouksim, M. Aboulfatah and T. Gadi. 2016. A New Evaluation Method for Mesh Segmentation Based on the Levenshtein Distance. International Review on Computers and Software (IRECOS).

[29] A. Golovinskiy and T. Funkhouser. 2008. Randomized cuts for 3D mesh analysis. ACM Transactions on Graphics. 27(5): 1.

[30] K. Arhid, F. R. Zakani, M. Bouksim,M. Aboulfatah and T. Gadi, "An Efficient Hierarchical 3D Mesh Segmentation Using Negative Curvature and Dihedral Angle," International Journal of Intelligent Engineering and Systems. vol. 10, no. 5, pp. 143–152, 2017.