A subjective experiment for 3D-mesh segmentation evaluation

Halim Benhabiles #1, Guillaume Lavoué *2, Jean-Philippe Vandeborre #−3, Mohamed Daoudi #−4

# LIFL (UMR USTL/CNRS 8022), University of Lille
France
− Institut TELECOM ; TELECOM Lille 1
France
[1] halim.benhabiles@lifl.fr
[3] jean-philippe.vandeborre@lifl.fr
[4] mohamed.daoudi@lifl.fr

** University of Lyon, CNRS, INSA-Lyon, LIRIS, UMR5205, F-69621*
France
[2] glavoue@liris.cnrs.fr

***Abstract*—In this paper we present a subjective quality assessment experiment for 3D-mesh segmentation. For this end, we carefully designed a protocol with respect to several factors namely the rendering conditions, the possible interactions, the rating range, and the number of human subjects. To carry out the subjective experiment, more than 40 human observers have rated a set of 250 segmentation results issued from various algorithms.**

The obtained Mean Opinion Scores, which represent the human subjects' point of view toward the quality of each segmentation, have then been used to evaluate both the quality of automatic segmentation algorithms and the quality of similarity metrics used in recent mesh segmentation benchmarking systems.

I. Introduction

3D-mesh segmentation is an active research topic with important applications such as indexing, compression, etc. The performance of these applications depends on a prior efficient segmentation algorithm. Hence, the *evaluation* of the 3D-mesh segmentation quality is a critical step. A natural approach to achieve this goal is subjective tests which are based on the quantitative human judgment.

In this context, the objective of the present work is to perform a subjective quality assessment experiment for 3D-mesh *semantic* segmentation. This latter kind of segmentation aims at distinguishing segments that correspond to relevant features of the shape, by following higher level notions such as defined in human perception theory [1], and gives a meaningful decomposition of the shape. To establish the subjective experiment, a protocol is designed with respect to many factors such as the number of human subjects and the rating range. The protocol is an effort to make subjective evaluation for 3D-mesh segmentation more relevant and standardized. In this subjective experiment, human observers have rated a set of segmentations obtained from different automatic algorithms. The results of the subjective experiment are useful for the quantitative evaluation of automatic segmentation algorithms, and for the evaluation of segmentation similarity metrics used in recent mesh segmentation benchmarking systems [2], [3].

The paper is organized as follows. Section 2 summarizes existing works addressing 3D-mesh segmentation evaluation, while section 3 details our experiment. Section 4 makes clear the usefulness of the subjective experiment results through a quantitative evaluation of four recent automatic segmentation algorithms and also the evaluation of objective segmentation similarity metrics proposed by [2], [3]. A conclusion is drawn in section 5.

II. Related work

Contrary to the important number of proposed algorithms addressing 3D-mesh segmentation [4], a little attention has been paid, by the computer graphics community, to the *quality evaluation* of the segmentations produced by these algorithms. Very recently, two main works [2], [3] have been proposed to accomplish this latter task. Both works are based on a benchmarking system including a ground-truth corpus and a set of similarity metrics. The ground-truth corpus comprises a set of 3D-models of different categories (human, animal, etc.), and each model is associated with multiple manual segmentations (ground-truths) done by human observers. The evaluation of a segmentation algorithm is then realized by measuring the similarity, using similarity metrics, between the automatic segmentation generated by this algorithm for a given model and its corresponding ground-truths. The closer is the automatic segmentation to ground-truths, the better its quality is.

Although, these solutions allow an objective and a quantitative evaluation thanks to the ground-truths and the similarity metrics, the ideal way to evaluate segmentation algorithms remains an explicit subjective experiment where observers directly rate segmentation results. Moreover, such subjective experiment will allow to quantify the efficiency of existing

MMSP'10, October 4-6, 2010, Saint-Malo, France.
978-1-4244-8112-5/10/$26.00 ©2010 IEEE.

ground-truth based benchmarks and to evaluate the quality of the introduced similarity metrics.

III. Our subjective experiment

A. The corpus of segmentations

The design of stimulus is a critical step in the subjective protocol. In our case, we need to select a set of 3D-models that will be segmented by different algorithms and then rated by human subjects. For this end, we use our corpus [2] of 3D-models which is available on-line[1] and is dedicated for the segmentation evaluation task. The size of the corpus is reasonable (28 3D-models), and its content is representative since it contains different categories of 3D-models. Figure 1 illustrates the models of the corpus with one manual segmentation per model.

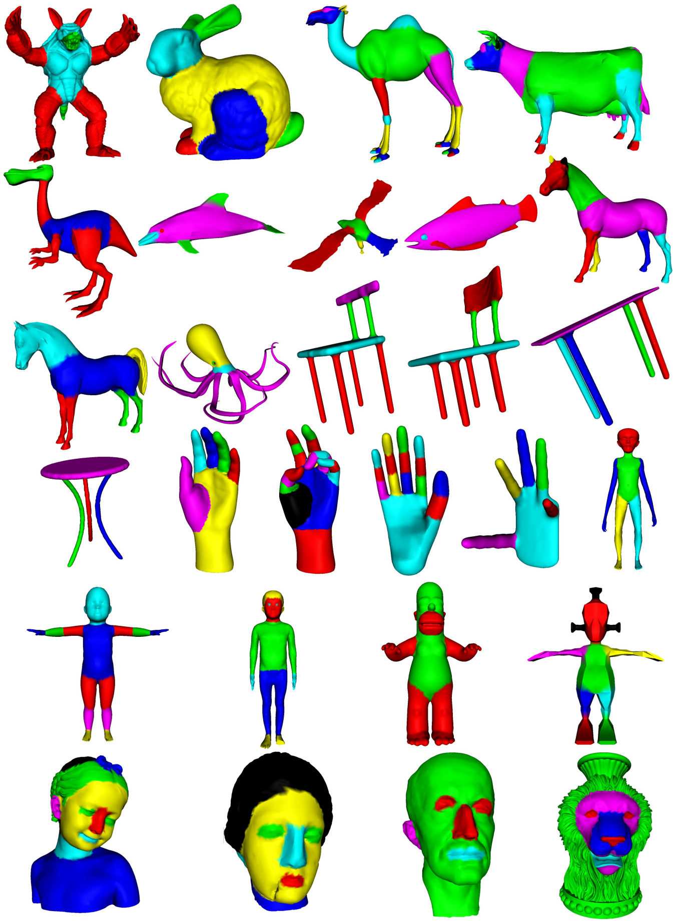


Fig. 1. Models of our corpus associated with one ground-truth.

In our experiment, we asked human subjects to rate segmentations of these objects coming from different automatic algorithms. We have created a set of 250 segmentations based on the corpus of 28 models. For this task, we have considered four automatic segmentation algorithms: Attene et al. [5], Lavoué et al. [6], Shapira et al. [7] and Tierny et al. [8]. The source code and/or the binary were provided by the authors for each automatic segmentation algorithm. Except the algorithm of Lavoué et al. [6], the others are hierarchical. Hence, for each algorithm, we generated two levels of segmentation per model namely coarse and fine, which gave 28 × 2 segmentations per algorithm and 28 segmentations for the Lavoué's et al. algorithm. Figure 2 illustrates an example of coarse and fine segmentation of the *hand* model using the algorithm from Tierny et al. [8].

Note that the number of segments of a given level of segmentation (coarse for example) is not the same through the different models and through the different algorithms. For the algorithms from Shapira et al. [7] and Tierny et al. [8], the number of segments is automatically computed. We just need to fix the level of detail of the desired segmentation. For the algorithm from Attene et al. [5], the number of segments is manually fixed, we then select two numbers (a small one and a big one). These two numbers vary according to the complexity of the model and to the consistency of the segmentation. For the algorithm from Lavoué et al. [6] the number of segments was also manually chosen so as to optimize the quality. To these 28 × 7 segmentations were added 28 ground-truth segmentations coming from our corpus [2] and 28 random segmentations generated using a simple algorithm based on a random region growing mechanism. Figure 3 illustrates different segmentations of the *camel* model. Thus we obtained a whole corpus of 250 segmentations to rate.

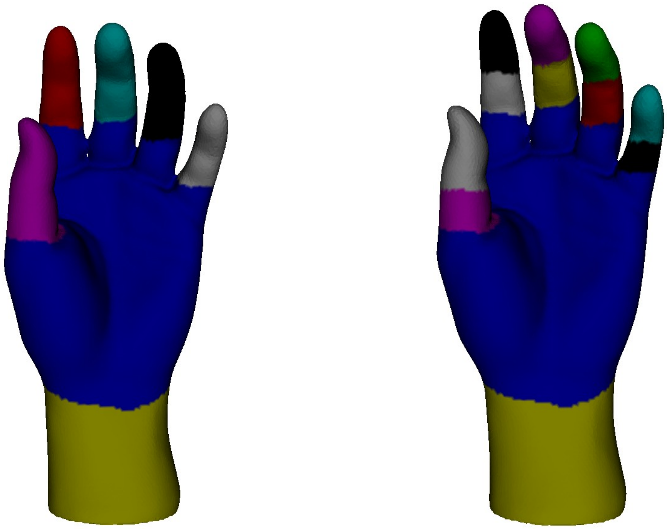
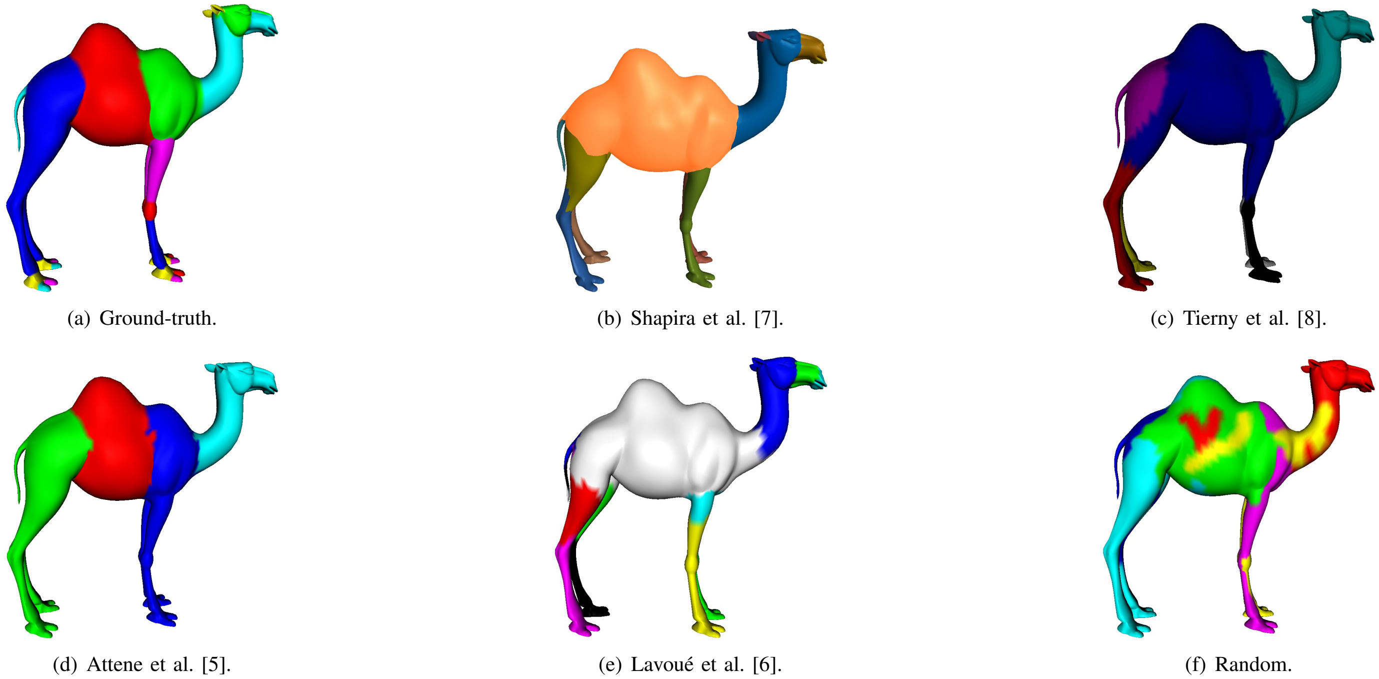


Fig. 2. From left to right, coarse and fine segmentation of the hand model using Tierny's et al. [7] algorithm.

B. Subjective protocol

The protocol that we propose is inspired from existing ones used for video segmentation quality evaluation, 3D-watermarking quality evaluation, and image quality evaluation [9], [10], [11]. They are all based on *Single Stimulus Continuous Quality Scale* (SSCQS) which is a standard technique used to evaluate the quality of video and multimedia content. Our protocol consists of the following stages:

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



(a) Ground-truth.

(b) Shapira et al. [7].

(c) Tierny et al. [8].

(d) Attene et al. [5].

(e) Lavoué et al. [6].

(f) Random.

Fig. 3. Segmentation of the camel model using different algorithms.

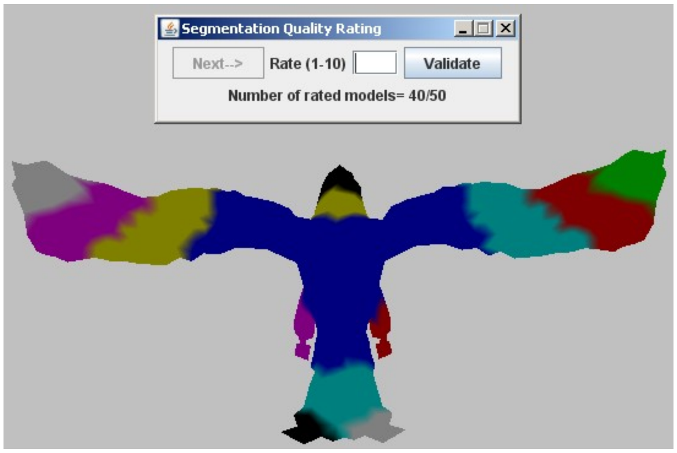
- ***Oral instructions.*** We give instructions to our volunteers and make them familiar with the rating task, the 3D-models, and the available interactions.
- ***Training.*** We show some ground-truth and random (bad) segmentations of several models, in order to clarify the concept of good and bad segmentation for the user and to establish a referential range for him. The goal for the user is not to learn the ground-truth of each model, but to learn what is a good segmentation so as to be able to rate the quality of a given segmentation independently from ground-truths.
- ***Experimental trials.*** For each segmentation from the corpus, we ask the volunteer to give a score between 1 and 10 indicating its quality from a semantic point of view. 10 for a perfect segmentation and 1 for a bad one. This scale range allows the volunteers to distinguish more easily between the quality of segmentations.

During the experiment trials, each segmentation is displayed one by one to the observer on a 22-inch LCD monitor, without the ground-truth. In order to avoid the effect of the temporal sequencing factor, the sequence of segmentations was randomly generated for each participant. Interaction was allowed (rotation, scaling, translation). It is important to notice that rating 250 segmentations represents a too much fastidious task for an observer; hence we only asked each human subject to rate 50 segmentations from the corpus (randomly chosen with a bias so as to obtain enough scores for all the 250 segmentations). The user interface which was developed for this rating task is illustrated in figure 4.

The Mean Opinion Score (MOS) is then computed for each segmentation of the corpus:

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

MOS_i is the mean opinion score of the i^{th} segmentation, n is the number of test subjects, and m_{ij} is the score ($\in [1, 10]$) given by the j^{th} subject to the i^{th} segmentation. This subjective experiment has been conducted on 45 people (students and staff) from the University of Lille, which provided a total of 10 opinion scores per segmentation.



Number of rated models= 40/50

Fig. 4. User interface for rating the segmentations.

TABLE I

ALGORITHMS RANKING ASSOCIATED WITH THE AVERAGE OF MOS FOR CORPUS CATEGORIES.

	Ground-truth	Shapira et al. [7]	Tierny et al. [8]	Attene et al. [5]	Lavoué et al. [6]	Random
animal	1 / 8.26	2 / 7.20	3 / 5.72	5 / 4.83	4 / 5.01	6 / 2.37
bust	1 / 8.03	2 / 4.64	4 / 2.81	3 / 3.64	5 / 2.64	6 / 1.78
furniture	1 / 9.25	3 / 7.74	5 / 3.35	2 / 8.53	4 / 6.21	6 / 1.99
hand	1 / 8.68	5 / 4.82	2 / 7.64	4 / 4.85	3 / 5.53	6 / 1.60
human	1 / 7.77	2 / 6.77	3 / 5.20	5 / 4.54	4 / 4.62	6 / 2.28
whole	**1 / 8.36**	**2 / 6.51**	**3 / 5.27**	**4 / 5.21**	**5 / 4.92**	**6 / 2.10**

IV. RESULTS AND DATA ANALYSIS

A. Consistency of the ratings

Firstly in order to check the suitability of our evaluation protocol and the relevance of the mean opinion scores, we have assessed the variation between the different observers in their subjective ratings of the objects. The value of the intraclass correlation coefficient (ICC) is 0.65, that is a rather good value that means that the observers had a good agreement on their visual estimations; hence we can assert that our protocol was correct since it led to produce meaningful consistent ratings among the observers.

B. Influence of the refinement on the segmentation quality

Some automatic algorithms are hierarchical, i.e. they are able to produce segmentations with different levels of refinement. An interesting experiment is to study whether this level of granularity influences the quality perceived by the observers. For this end, we averaged the MOS of the models for each category, for each algorithm and for both levels of segmentation (coarse and fine), then we compared the results of the two levels. Figure 5 illustrates the obtained results for the three hierarchical algorithms. One can notice that the averages of the two levels of segmentation for a given category or for the whole corpus are close to each other. More exactly, the average variation between the two levels for the whole corpus and for each algorithm: Shapira et al. [7], Attene et al. [5] and Tierny et al. [8] is respectively of 7%, 11%, and 10%. This means that the segmentations remain consistent whatever their level of refinement.

C. Performance comparison of segmentation algorithms

Table I presents the rank, based on the MOS, of each algorithm (fine segmentation for hierarchical algorithms) for each category models of the corpus including random segmentations and ground-truths. The MOS mean values are also displayed. As expected, our ground-truths have the best ranks for each category and for the whole corpus, when random segmentations have the worst ones. This validate the relevance of our ground-truth corpus [2]. The table shows that there is no automatic algorithm which outperforms the others in all categories. It also shows that the models of the *bust* category, seem to be the most difficult to segment by automatic algorithms, since the average of their MOS is the lowest with comparison to other categories. This may be due to the geometrical and topological complexity of these models, but the main reason is probably the fact that these models represent human faces. Human face images are well-known in subjective experiments as a high-level factor attracting human attention, hence some features not relevant from a geometrical point of view can be considered highly relevant for human observers. Globally, the algorithm from Shapira et al. [7] seems to be the best one after ground-truths.

D. Evaluation of similarity metrics

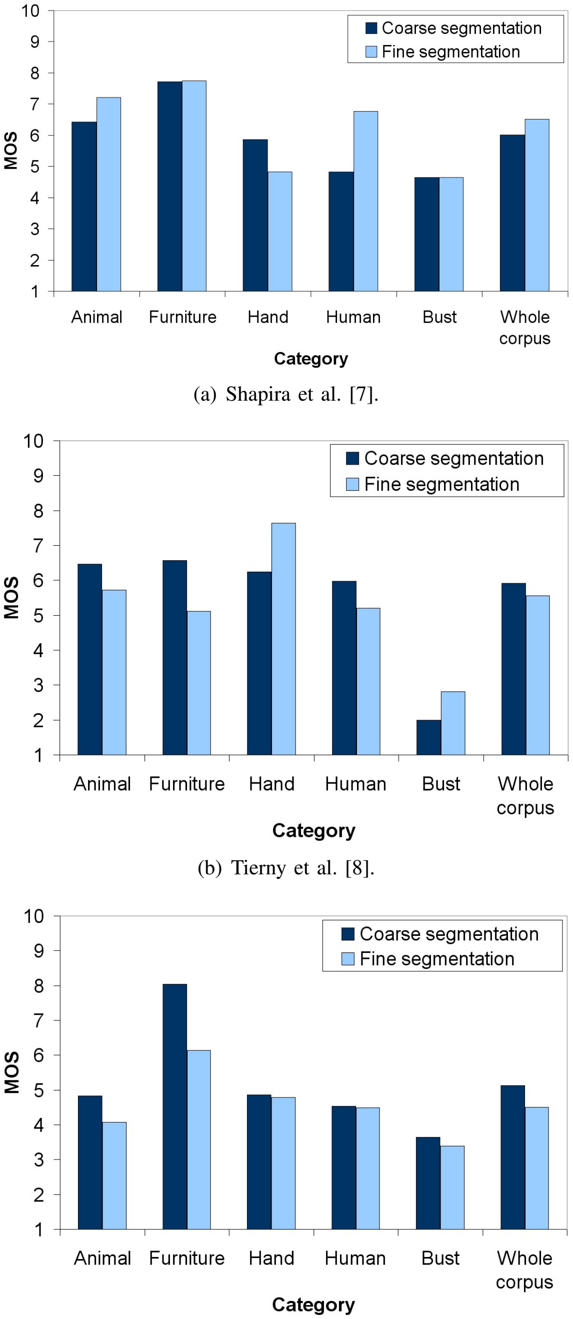
Another interesting experiment is to evaluate the quality of the similarity metrics used in benchmarking systems [2], [3]. For this end, we use our corpus [2] which is based on the same models used in the subjective experiment and comprises 4 ground-truths for each model. We compute the similarity between the 250 segmentations and their corresponding ground-truths using the following metrics: Cut discrepancy (CD), Local Consistency Error (LCE), Hamming distance (HD), and Rand index (RI). Then we compute the correlation (Spearman rank correlation [12]) between the 250 MOS and the 250 values acquired by each metric.

Table II shows these correlation values which are high (more than 80%) using the RI metric, medium (between 50% and 60%) using the LCE and HD metrics, and low (less than 30%) using the CD metric. This means that this latter metric fails to distinguish whether an automatic segmentation is close or not to ground-truths. Hence, a benchmark which is based on it will give irrelevant results. It is clear that the ideal is to have a metric which reflects as much as possible the human opinion on the quality of a segmentation. We can conclude that the RI is the appropriate metric to use in existing benchmarks [2], [3] since it outperforms the other metrics for each category and for the whole corpus.

TABLE II

SPEARMAN CORRELATION VALUES (%) BETWEEN THE MEAN OPINION SCORES AND THE VALUES OF DIFFERENT METRICS.

	CD	LCE	HD	RI
animal	19.0	44.5	50.2	76.9
bust	11.1	76.4	69.1	82.6
furniture	23.4	57.5	76.2	86.4
hand	54.0	79.8	75.15	83.2
human	03.5	65.6	63.9	70.8
whole	**26.5**	**57.9**	**61.8**	**81.1**



(a) Shapira et al. [7].

(b) Tierny et al. [8].

(c) Attene et al. [5].

Fig. 5. Average of MOS of segmentations obtained from different hierarchical algorithms.

V. Conclusion

In this paper a subjective segmentation rating experiment is proposed. The protocol has been carefully designed so as to be able to obtain relevant results. The results are very useful to understand the human perceptual mechanisms. They yield to the quantitative quality evaluation of automatic algorithms and the validation of similarity metrics used in recent benchmarking systems.

In the future, we plan to combine these results with our ground-truth corpus to propose a learning segmentation algorithm. This would help to improve the segmentation quality of automatic algorithms.

Acknowledgment

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

References

[1] I. Biederman, “Recognition-by-compenents: A theory of human image understanding,” *Psychological Review*, vol. 94, pp. 115–147, 1987.

[2] H. Benhabiles, J.-P. Vandeborre, G. Lavoué, and M. Daoudi, “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.

[3] X. Chen, A. Golovinskiy, and T. Funkhouser, “A benchmark for 3d mesh segmentation,” *ACM Transactions on Graphics (SIGGRAPH)*, vol. 28(3), 2009.

[4] A. Shamir, “A survey on mesh segmentation techniques,” *Computer Graphics Forum*, vol. 27, no. 6, pp. 1539–1556, 2008.

[5] M. Attene, B. Falcidieno, and M. Spagnuolo, “Hierarchical mesh segmentation based on fitting primitives,” *Vis. Comput.*, vol. 22, no. 3, pp. 181–193, 2006.

[6] G. Lavoué, F. Dupont, and A. Baskurt, “A new cad mesh segmentation method, based on curvature tensor analysis,” *Computer Aided Design*, vol. 37(10), pp. 975–987, 2005.

[7] L. Shapira, A. Shamir, and D. Cohen-Or, “Consistent mesh partitioning and skeletonisation using the shape diameter function,” *Vis. Comput.*, vol. 24, no. 4, pp. 249–259, 2008.

[8] J. Tierny, J.-P. Vandeborre, and M. Daoudi, “Topology driven 3D mesh hierarchical segmentation,” in *IEEE International Conference On Shape Modeling And Application (SMI)*, 2007.

[9] E. D. Gelasca, T. Ebrahimi, M. Karaman, and T. Sikora, “A framework for evaluating video object segmentation algorithms,” in *CVPRW ’06: Proceedings of the 2006 Conference on Computer Vision and Pattern Recognition Workshop*. IEEE Computer Society, 2006, p. 198.

[10] M. Corsini, E. Drelie Gelasca, T. Ebrahimi, and M. Barni, “Watermarked 3d mesh quality assessment,” *IEEE Transaction on Multimedia*, vol. 9, no. 2, pp. 247–256, February 2007.

[11] B. E. Rogowitz and H. E. Rushmeier, “Are image quality metrics adequate to evaluate the quality of geometric objects,” in *in Human Vision and Electronic Imaging*, 2001, pp. 340–348.

[12] W. W. Daniel, *A Foundation For Analysis In The Health Sciences Books*. 7th edition.John Wiley and sons., 1999.