A subjective experiment for 3D-mesh segmentation evaluation
Halim Benhabiles, Guillaume Lavoué, Jean-Philippe Vandeborre, Mohamed Daoudi

To cite this version:

HAL Id: hal-00666767
https://hal.archives-ouvertes.fr/hal-00666767
Submitted on 6 Feb 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.
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
* University of Lyon, CNRS, INSA-Lyon, LIRIS, UMR5205, F-69621
France
2 glavoue@liris.cnrs.fr
3 mohamed.daoudi@lifl.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
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 segmen-
tation per model.

In our experiment, we asked human subjects to rate seg-
mentations 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 \(28 \times 2\) 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 \(28 \times 7\) 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.

\(^1\)http://www-rech.telecom-lille1.eu/3dsegbenchmark/
• **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:

\[
MOS_i = \frac{1}{n} \sum_{j=1}^{n} m_{ij}
\]

\(MOS_i\) 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.

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

Fig. 4. User interface for rating the segmentations.
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>
<thead>
<tr>
<th></th>
<th>CD</th>
<th>LCE</th>
<th>HD</th>
<th>RI</th>
</tr>
</thead>
<tbody>
<tr>
<td>animal</td>
<td>19.0</td>
<td>44.5</td>
<td>50.2</td>
<td>76.9</td>
</tr>
<tr>
<td>bust</td>
<td>11.1</td>
<td>76.4</td>
<td>69.1</td>
<td>82.6</td>
</tr>
<tr>
<td>furniture</td>
<td>23.4</td>
<td>57.5</td>
<td>76.2</td>
<td>86.4</td>
</tr>
<tr>
<td>hand</td>
<td>54.0</td>
<td>79.8</td>
<td>75.15</td>
<td>83.2</td>
</tr>
<tr>
<td>human</td>
<td>03.5</td>
<td>65.6</td>
<td>63.9</td>
<td>70.8</td>
</tr>
<tr>
<td>whole</td>
<td>26.5</td>
<td>57.9</td>
<td>61.8</td>
<td>81.1</td>
</tr>
</tbody>
</table>

TABLE II Spearman correlation values (%) between the mean opinion scores and the values of different metrics.
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