Une expérimentation subjective pour l'évaluation de segmentations de maillages 3D

H. Benhabiles[1] G. Lavoué[2] J-P. Vandeborre[1,3] M. Daoudi[1,3]

[1] LIFL (UMR USTL/CNRS 8022), Université de Lille, France

[2] Université de Lyon, CNRS, INSA-Lyon, LIRIS, UMR5205, F-69621, France

[3] Institut TELECOM ; TELECOM Lille 1, France

{halim.benhabiles, jean-philippe.vandeborre, mohamed.daoudi}@lifl.fr
{glavoue}@liris.cnrs.fr

Résumé

Dans cet article, nous proposons une expérimentation subjective pour l'évaluation de la qualité de segmentations de maillages 3D. Dans ce but, nous avons conçu un protocole tout en respectant un certain nombre de facteurs : les conditions d'affichage, les interactions possibles, l'intervalle de notation, ainsi que le nombre d'opérateurs humains. Afin de mettre en œuvre cette expérimentation, plus de 30 opérateurs humains ont participé à la notation de 250 segmentations provenant de plusieurs algorithmes. Pour éviter l'effet du facteur de séquencement temporel, les segmentations ont été affichées aux opérateurs de manière aléatoire avec un biais pour obtenir suffisamment de notes pour chacune de ces segmentations. Le score moyen (Mean Opinion Score) est ensuite calculé pour chaque segmentation. Ce score reflète l'opinion des opérateurs vis-à-vis de la qualité de la segmentation.

Les résultats de l'expérimentation subjective sont utilisés pour évaluer la qualité des algorithmes automatiques utilisés ainsi que les métriques existantes de similarité entre segmentations.

Mots clefs

Maillage 3D, évaluation, segmentation, expérimentation subjective.

1 Introduction

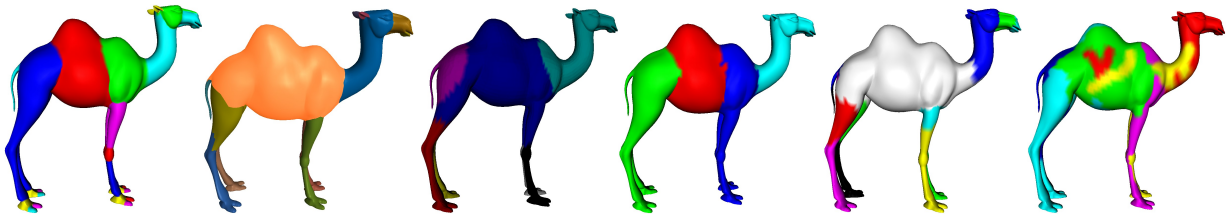
La segmentation de maillages 3D est un domaine de recherche très actif avec de nombreuses applications importantes telles que l'indexation, la compression, etc. La performance de ces applications dépend fortement de l'efficacité de l'algorithme de segmentation. L'*évaluation* de la qualité d'une segmentation de maillage 3D est donc une étape critique. Une approche naturelle pour atteindre cet objectif est d'effectuer des tests subjectifs basés sur le jugement humain quantitatif.

Dans ce contexte, l'objectif du présent travail est de mettre en œuvre une expérimentation subjective pour l'évaluation de la qualité des segmentations de maillages 3D. Pour cela, nous avons conçu un protocole tout en respectant un certain nombre de facteurs tels que l'intervalle de notation, les conditions d'affichage, etc. Ce protocole vise à standardiser l'évaluation subjective et à la rendre plus pertinente. Dans cette expérimentation, les opérateurs humains ont noté un ensemble de segmentations provenant de plusieurs algorithmes. Les résultats de cette expérimentation sont utilisés pour l'évaluation quantitative des algorithmes de segmentation automatique, ainsi que l'évaluation des métriques de similarités entre segmentations, utilisées dans les systèmes de benchmark récents [1, 2].

L'article est organisé comme suit. La section 2 fournit un court état de l'art sur les travaux existants concernant l'évaluation de la segmentation de maillage 3D. La section 3 décrit en détail notre expérimentation. La section 4 met en avant l'utilité des résultats de l'expérimentation subjective à travers l'évaluation quantitative de quatre algorithmes de segmentation récents, ainsi que l'évaluation objective des métriques de similarités proposées dans [1, 2]. La section 5 conclut l'article.

2 Etat-de-l'art sur l'évaluation de la segmentation de maillages 3D

Contrairement aux nombreuses propositions d'algorithmes adressant la segmentation de maillages 3D [7], moins d'attention a été accordée par la communauté graphique envers l'*évaluation* de la *qualité* de la segmentation produite par ces algorithmes. Deux travaux principaux ont été proposés récemment [1, 2] sur ce sujet de l'évaluation. Ils s'appuient sur un système de benchmarking incluant un corpus de vérités-terrains et un ensemble de métriques de similarité. Le corpus de vérités-terrains est composé d'un ensemble de modèles 3D faisant partie de différentes catégories (humain, animal, etc.). Chaque modèle 3D est associé à plusieurs vérités-terrains (segmentations manuelles) effectuées par des opérateurs humains. L'évaluation d'un algorithme de segmentation consiste à mesurer la simila-



(a) Vérité-terrain. (b) Shapira *et al.* [3]. (c) Tierny *et al.* [4]. (d) Attene *et al.* [5]. (e) Lavoué *et al.* [6]. (f) Aléatoire.

Figure 1 – *Segmentation du modèle* camel *en utilisant plusieurs algorithmes.*

rité, en utilisant des métriques de similarité, entre la segmentation automatique produite par cet algorithme pour un modèle donné, et ses vérités-terrains correspondantes. Plus la segmentation automatique est proche des vérités-terrains, meilleure est la qualité de l'algorithme.
Bien que ces solutions permettent une évaluation qui est à la fois objective et quantitative grâce aux vérités-terrains et aux métriques de similarités, le moyen idéal pour évaluer les algorithmes de segmentation reste une expérimentation subjective explicite, où des observateurs notent directement les segmentations résultantes. De plus, une telle expérimentation subjective permettra de quantifier l'efficacité des benchmarks existants ainsi que d'évaluer les métriques de similarité introduites.

3 L'expérimentation subjective

3.1 Corpus des segmentations

La conception du stimulus est une étape clé dans le protocole subjectif. Dans notre cas, nous avons besoin de sélectionner un ensemble de modèles 3D qui seront segmentés par plusieurs algorithmes puis notés par des opérateurs humains. A cette fin, nous utilisons notre corpus [1] de modèles 3D qui est disponible en-ligne [1], et est dédié à l'évaluation de la segmentation. La taille du corpus est raisonnable (28 modèles 3D), et son contenu est représentatif puisqu'il comprend différentes catégories communes de modèles 3D.
Dans notre expérimentation, nous avons demandé aux opérateurs humains de noter les segmentations de ces objets provenant de plusieurs algorithmes automatiques. Nous avons crée un ensemble de 250 segmentations basé sur les 28 modèles 3D du corpus. Pour cette tâche, nous avons pris en considération 4 algorithmes de segmentation automatique : Attene *et al.* [5], Lavoué *et al.* [6], Shapira *et al.* [3], et Tierny *et al.* [4]. Mis à part l'algorithme de Lavoué *et al.* [6], les autres sont hiérarchiques ; nous avons donc généré, pour chacun d'eux, deux niveaux de segmentations : grossière et fine, pour chaque modèle, ce qui donne au total 28×2 segmentations par algorithme hiérarchique et 28 segmentation pour l'algorithme de Lavoué *et al.* [6]. A ces 28×7 segmentations, nous avons ajouté 28 segmentations vérités-terrains provenant de notre corpus ainsi que 28 segmentations aléatoires générées par un algorithme simple basé sur un mécanisme de croissance de région aléatoire. La figure 1 illustre différentes segmentations du modèle *camel*. Le code source et/ou le binaire de chaque algorithme de segmentation automatique sont fournis par leurs auteurs. Ainsi, nous avons obtenu un corpus de 250 segmentations à noter.

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

3.2 Protocole subjectif

Le protocole que nous proposons est inspiré de ceux qui existent déjà pour l'évaluation de la qualité de segmentation de vidéo [8], l'évaluation de la qualité de tatouage 3D [9], et l'évaluation de la qualité d'image [10]. Ces protocoles sont tous basés sur le *Single Stimulus Continuous Quality Scale* (SSCQS) qui est une technique standard de notation utilisée pour évaluer la qualité de vidéo et du contenu multimédia. Notre protocole comprend les étapes suivantes :

- ***Instructions orales.*** Nous expliquons à nos volontaires la tâche qu'il doivent compléter, et nous les familiarisons avec l'opération de notation, les modèles 3D, ainsi que les interactions possibles.
- ***Apprentissage.*** Nous montrons quelques segmentations vérités-terrains et aléatoires de différents modèles à l'utilisateur afin qu'il puisse comprendre le concept de bonne et de mauvaise segmentation, et établir un ordre de référence. Le but n'est pas d'apprendre à l'utilisateur les vérités-terrains de chaque modèle, mais plutôt de lui apprendre à distinguer entre les différentes segmentations pour qu'il soit en mesure de noter la qualité d'une segmentation donnée indépendamment des vérités-terrains.
- ***Tests expérimentaux.*** Pour chaque segmentation du corpus, nous avons demandé au volontaire de donner un score entre 1 et 10, indiquant sa qualité d'un point de vue sémantique, 10 pour une segmentation parfaite, et 1 pour une segmentation très mauvaise. Cet intervalle permet aux volontaires de distinguer plus facilement entre la qualité des segmentations.

Durant les tests expérimentaux, les segmentations sont affichées une par une au volontaire sur un écran LCD 22-pouces, sans les vérités-terrains. Pour éviter l'effet du facteur de séquencement temporel, la séquence de segmentations affichée à chaque participant est générée

aléatoirement. Les interactions de base sont autorisées (zoom, rotation, translation). Il est évident que la notation de 250 segmentations par un volontaire représente une tâche fastidieuse, ce qui nous a incité à restreindre ce nombre à 50 segmentations par volontaire de manière à obtenir suffisamment de notes pour les 250 segmentations. La figure 2 illustre l'interface développée pour l'opération de notation.

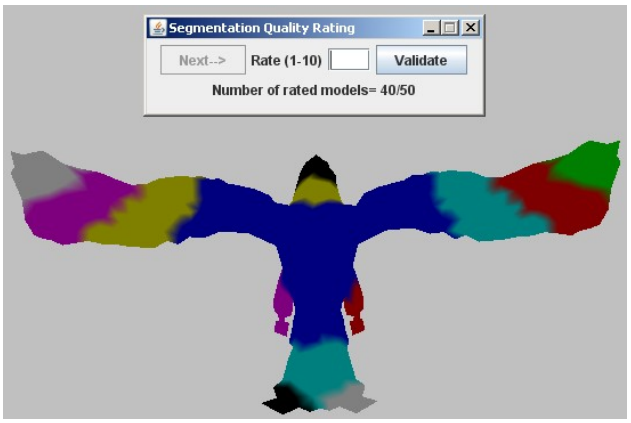


Figure 2 – *Interface utilisateur pour la notation des segmentations.*

Le *Mean Opinion Score* (MOS) est ensuite calculé pour chaque segmentation du corpus :

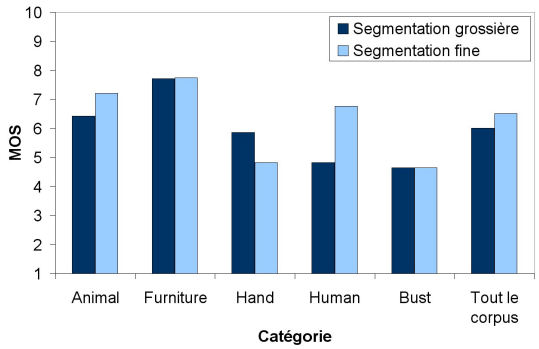
$$MOS_i = \frac{1}{n}\sum_{j=1}^{n} m_{ij} \quad (1)$$

MOS_i est le *mean opinion score* de la i^{eme} segmentation, n est le nombre de sujets, et m_{ij} est le score ($\in [1, 10]$) affecté par le j^{eme} sujet à la i^{eme} segmentation. Cette expérimentation subjective à été menée sur 35 personnes (étudiants et personnels) de l'Université de Lille 1, offrant un total de 7 scores par segmentation.

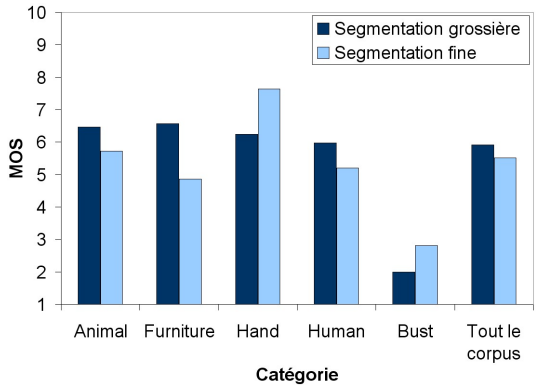
4 Résultats et analyse des données

4.1 Influence du raffinement sur la qualité de la segmentation

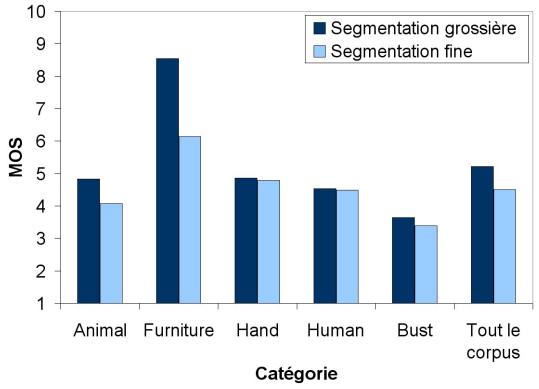
Certains algorithmes automatiques sont hiérarchiques, c'est-à-dire qu'ils sont capables de produire des segmentations avec différents niveaux de raffinement. Une expérience intéressante est d'étudier l'influence du niveau de granularité sur la qualité perçue par les observateurs. Dans ce but, nous avons calculé la moyenne du MOS des modèles de chaque catégorie, pour chaque algorithme, et pour les deux niveaux de segmentation (grossière et fine), ensuite nous avons comparé les résultats des deux niveaux. La figure 3 illustre les résultats obtenus pour les trois algorithmes hiérarchiques. On peut remarquer que les moyennes des deux niveaux de segmentation, pour une catégorie donnée ou bien pour tout le corpus, sont proches l'une de l'autre. Plus précisément, la variation moyenne entre les deux niveaux pour tout le corpus, et pour chacun des algorithmes : Shapira *et al.* [3], Tierny *et al.* [4], et Attene *et al.* [5], est respectivement de 7%, 10%, et 11%. Cela veut dire que les segmentations restent sémantiquement consistantes quelque soit leur niveau de raffinement.



(a) Shapira *et al.* [3].



(b) Tierny *et al.* [4].



(c) Attene *et al.* [5].

Figure 3 – *Moyennes des MOS des segmentations obtenues par des algorithmes hiérarchiques.*

4.2 Comparaison de la performance des algorithmes de segmentation

Le tableau 1 présente le classement, basé sur le MOS, de chaque algorithme (segmentation fine pour les algorithmes

Tableau 1 – *Classement des algorithmes accompagné des moyennes MOS pour chaque catégorie du corpus.*

	Vérités-terrains	**Shapira *et al.* [3]**	**Tierny *et al.* [4]**	**Attene *et al.* [5]**	**Lavoué *et al.* [6]**	**Aléatoire**
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
tout	**1 / 8.36**	**2 / 6.51**	**3 / 5.27**	**4 / 5.21**	**5 / 4.92**	**6 / 2.10**

hiérarchiques) pour chaque catégorie de modèles ainsi que pour tout le corpus, incluant les segmentations aléatoires et les segmentations vérités-terrains. Les moyennes du MOS sont aussi affichées. Comme prévu, les vérités-terrains ont le meilleur classement pour chaque catégorie et pour tout le corpus, alors que les segmentations aléatoires sont les dernières en classement. Ceci valide la pertinence de notre corpus de vérités-terrains. Le tableau montre qu'il n y a aucun algorithme automatique qui atteint les meilleurs scores pour toute les catégories. Il montre aussi que la classe *bust* est la plus difficile à segmenter par les algorithmes automatiques, puisque la moyenne de son MOS est la plus faible en la comparant avec les autres classes. Ceci peut être dû à la complexité géométrique des modèles de cette classe, mais la raison principale est probablement le fait que ces modèles représentent des visages humains. Ce dernier type de modèles est connu dans les expérimentations subjectives comme étant un facteur de haut niveau qui attire l'attention humaine. En effet, certaines caractéristiques qui ne sont pas significatives d'un point de vue géométrique, peuvent être considérées manifestement significatives par les observateurs humains. Globalement, l'algorithme de Shapira *et al.* [3] semble être le meilleur après les vérités-terrains.

4.3 Evaluation des métriques de similarité

Une autre expérience intéressante est d'évaluer la qualité des métriques de similarité utilisées dans les systèmes de benchmark [1, 2]. Pour cela, nous utilisons le corpus présenté dans [1] qui est basé sur les mêmes 28 modèles que ceux utilisés dans l'expérimentation subjective, et comprend 4 vérités-terrains pour chaque modèle. Nous calculons la similarité entre les 250 segmentations et leurs vérités-terrains correspondantes à l'aide des métriques suivantes : Ecart de Frontières (EF), Erreur de Consistance Locale (ECL), Distance de Hamming (DH), et Indice de Rand (IR). Ensuite nous calculons la corrélation (Spearman rank correlation [11]) entre les 250 MOS et les 250 valeurs calculées par chacune des métriques. Si les métriques sont pertinentes, leurs valeurs devraient être corrélées avec les MOS données par les utilisateurs. Les résultats sont affichés dans le tableau 2. On peut distinguer dans ce tableau, 3 classes de corrélation : corrélation élevée (plus de 80%) pour la métrique IR, corrélation moyenne (entre 50% et 60%) pour les métriques ECL et DH, et corrélation faible (moins de 30%) pour la métrique EF. Le faible taux de corrélation de cette dernière métrique indique qu'elle échoue à clairement différencier entre une segmentation proche des vérités-terrains, et une mauvaise segmentation. Ainsi, un benchmark basé sur cette métrique donnera forcement des résultats qui ne sont pas pertinents. La métrique IR est la métrique qui donne les meilleurs résultats ; c'est donc celle qui devrait être utilisée en priorité dans les benchmarks existants [1, 2] puisqu'elle donne le meilleur taux de corrélation que se soit pour une catégorie donnée ou bien pour tout le corpus. Cette forte corrélation de 82% valide non seulement cette métrique mais atteste également de la qualité globale du benchmark présenté dans [1] (métrique et vérités-terrains).

Tableau 2 – *Corrélation de Spearman (%) entre les MOS et les valeurs des différentes métriques.*

	EF	**ECL**	**DH**	**IR**
animal	19.9	42	49.3	78.3
bust	8.9	73.5	68.1	81.6
furniture	21.6	52.9	69.2	85.3
hand	55.4	77.1	72.4	82.8
human	11.5	63.6	64.3	76.7
tout	**28.2**	**55.3**	**60.2**	**82.1**

5 Conclusion

Dans cet article, une expérimentation subjective pour la notation de segmentations de maillages 3D a été proposée. Dans ce but, un protocole a été soigneusement défini afin que les résultats obtenus soient pertinents. Ces résultats se sont avérés très utiles puisqu'ils nous ont permis d'effectuer une évaluation quantitative de la qualité des segmentations d'algorithmes automatiques récents ainsi que d'évaluer les métriques de similarité utilisées dans les systèmes de benchmark actuels. Dans le futur, nous visons à utiliser ces résultats avec notre corpus de vérités-terrains pour proposer un nouvel algorithme de segmentation.

Remerciements

Ce travail a bénéficié d'une aide de l'ANR (Agence Nationale de la Recherche) à travers le projet MADRAS (ANR-07-MDCO-015).

Références

[1] H. Benhabiles, J-P. Vandeborre, G. Lavoué, et M. Daoudi. A framework for the objective evaluation of segmentation algorithms using a ground-truth of human segmented 3d-models. Dans *IEEE International Conference On Shape Modeling And Application (SMI)*, 2009.

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

[3] Lior Shapira, Ariel Shamir, et Daniel Cohen-Or. Consistent mesh partitioning and skeletonisation using the shape diameter function. *Vis. Comput.*, 24(4) :249–259, 2008.

[4] Julien Tierny, Jean-Philippe Vandeborre, et Mohamed Daoudi. Topology driven 3D mesh hierarchical segmentation. Dans *IEEE International Conference On Shape Modeling And Application (SMI)*, 2007.

[5] Marco Attene, Bianca Falcidieno, et Michela Spagnuolo. Hierarchical mesh segmentation based on fitting primitives. *Vis. Comput.*, 22(3) :181–193, 2006.

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

[7] A. Shamir. A survey on mesh segmentation techniques. *Computer Graphics Forum*, 27(6) :1539–1556, 2008.

[8] Elisa Drelie Gelasca, Touradj Ebrahimi, Mustafa Karaman, et Thomas Sikora. A framework for evaluating video object segmentation algorithms. Dans *CVPRW '06 : Proceedings of the 2006 Conference on Computer Vision and Pattern Recognition Workshop*, page 198. IEEE Computer Society, 2006.

[9] Massimiliano Corsini, Elisa Drelie Gelasca, Touradj Ebrahimi, et Mauro Barni. Watermarked 3d mesh quality assessment. *IEEE Transaction on Multimedia*, 9(2) :247–256, February 2007.

[10] Bernice E. Rogowitz et Holly E. Rushmeier. Are image quality metrics adequate to evaluate the quality of geometric objects. Dans *in Human Vision and Electronic Imaging*, pages 340–348, 2001.

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