# SHREC08 Entry: Report of the Stability Track on Watertight Models

Silvia Biasotti*  Marco Attene†

IMATI CNR, Dept. of Genova, Via De Marini, 6 - Genova - Italy

**ABSTRACT**

In this report we present the results of the *Stability on Watertight Models Track*. The aim of this track is to evaluate the stability of algorithms with respect to input perturbations that modify the representation of the object without changing its overall shape significantly. Examples of these perturbations include geometric noise, varying sampling patterns, small shape deformations and topological noise.

**Index Terms:** H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval—Retrieval Models H.5.2 [Information Interfaces and Presentation]: User Interfaces—Benchmarking;

## 1 INTRODUCTION

A major barrier to a widespread adoption of 3D retrieval techniques in both commercial and academic systems is the lack of a standardized evaluation of the methods. What is the best shape characterization or the best similarity measure for a given domain? The answer is not trivial at all and depends on several factors. The aim of SHREC is to evaluate the performance of existing 3D shape retrieval algorithms, by highlighting their strengths and weaknesses, using a common test collection that allows for a direct comparison of methods. After the first successful experience of SHREC 2006, from 2007 the contest has moved towards a multi-track organization, in which different datasets are used to target different retrieval contexts. SHREC 2008 continues with the multi-track organization and this report outlines the results on datasets containing perturbed models.

## 2 DATA COLLECTION AND QUERIES

Two data collections have been provided with this track. Both collections are made of watertight mesh models in which various kinds of perturbations were introduced. Two sets of models A and B were provided, the set B containing the models in A. More in detail, the set B is made of 15 classes of 100 models each, for a total of 1500 models; A contains 1229 models (all the models in B after having excluded the 271 models with self-intersections).

The set B has been generated as follows. Among the 20 classes used in the SHREC07 track *Watertight models* [2], we have selected 15 classes, namely *humans*, *cups*, *glasses*, *airplanes*, *chairs*, *octopuses*, *tables*, *hands*, *fishes*, *birds*, *springs*, *armadillos*, *bustes*, *mechanical parts*, *four leg animals*; then, we perturbed the 20 models in each class with additive Gaussian noise, unven re-sampling, small protrusions, and topological noise (see an example in fig. 1). At the end each class of the dataset B was made of of 100 models.

Each model was used in turn as a query against the remaining part of the database. For a given query, the goal of the track is to retrieve the most similar objects. The relevance, marginal relevance or non-relevance of the models for a given query, i.e. the ground truth, was established a priori by two classification schemes.

*e-mail: silvia.biasotti@ge.imati.cnr.it

†e-mail: marco.attene@ge.imati.cnr.it

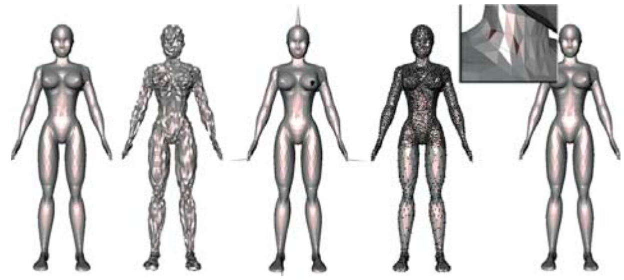



Figure 1: (a) A model of the database [2] and its perturbations: (b) Gaussian noise, (c) small protrusions, (d) uneven re-sampling and (e) adding topological noise.

## 3 PARTICIPANTS

Each participant was asked to submit up to 3 runs of his/her algorithm, in the form of dissimilarity matrices; each run could be for example the result of a different setting of parameters or the use of a different similarity metric. We remind that the entry $(i, j)$ of a dissimilarity matrix represents the distance between models $i$ and $j$. This track included 3 groups of participants:

1. Tony Tung and Francis Schmitt with 3 matrices;
2. Thibault Napolon, Tomasz Adamek, Francis Schmitt and Noel E. OConnor with 2 matrices;
3. Dong Xu, Li Cui, Ping Hu, Weiguo Cao and Hua Li, with 3 matrices.

For details on the algorithms and the different runs proposed by the participants, the reader is referred to their papers, included at the end of this report.

## 4 PERFORMANCE MEASURES

The performance of the methods on the dataset B has been evaluated by considering two different levels of ground truth. The first classification (coarser) considers in the same class the models in the original class and their perturbations, that is, each class is made of the 20 original models plus their four perturbations so that a total of 100 elements per each class was reached. The second classification (finer) considers in the same class every model and its perturbations, that is, each class is made of 5 models: 1 original model plus its four perturbed versions. Then, this classification subdivides the dataset in 300 classes of five elements.

The two schemes correspond to two possible interpretation of the stability of the methods: in the first case we evaluate how much the models and their perturbations are still recognized to belong to the original class while in the second case the attention is on the model and its perturbations rather than to the other models in the same original class.

As performance measures of the method we have adopted the *precision* and *recall*, that are two fundamental measures often used in evaluating search strategies. Recall is the ratio of the number of relevant records retrieved to the total number of relevant records in the database, while precision is the ratio of the number of relevant records retrieved to the size of the return vector [3].

In our contest, for each query the total number of relevant records in the database is 100 for the coarser classification and 5 for the finer one, that is the size of each class. Starting from here, we evaluate the precision-recall measures for each query, and then average it over each class and over the entire database.

Recall and precision are represented in a diagram, where precision has been computed as average of the precision scores after each relevant item in the scope. Finally, we consider the area under the diagrams which is relevant to evaluate the overall performance of a method.

## 5 RESULTS AND DISCUSSIONS

The participants sent two or three dissimilarity matrices on the dataset B that correspond to different choices of the parameters. A general observation is that almost the performances of the same method perform more or less the same. The performance of the participants is evaluated using their single best run, selected in terms of the area of the precision-recall diagram; details on each run and performance on the dataset A are in [1]. Curves shifted upwards and to the right indicate a superior performance which is may be roughly described by the area under the graph.

Fig.2(a) shows the recall precision diagram obtained using the coarse classification of the dataset, i.e., the original models of a class and their perturbations are considered in the same class. Fig.2(b) shows the recall precision diagram obtained using the fine classification of the dataset, i.e., a single class of models is made of the original model and its four perturbations.

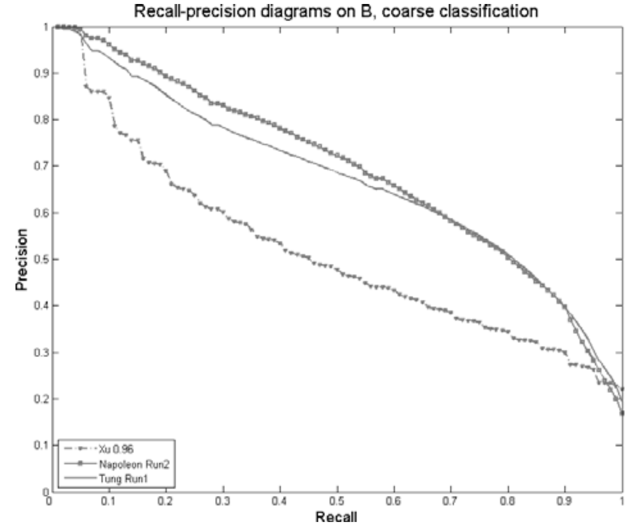
Finally, we have performed experiments on the retrieval performance of the method on the single types of perturbations. In this case, we have evaluated the retrieval performance of the methods when the original models are used as queries against one perturbation at time and when the models obtained using a single perturbation are used as queries against theirselves. For these methods, the degradation is measured in terms of the area of the recall precision graph and is reported in Tables 1 and 2 as a percentage of the performance of the methods on the models of the original dataset.

| **Degradation of the retrieval performance** | | | | |
|---|---|---|---|---|
| *Method* | *GN* | *SP* | *TN* | *UR* |
| Tung et al. | 46.4% | 34.14% | 44.32% | 38.90% |
| Napoleon et al. | 44.96% | 37.50% | 47.83% | 36.54% |
| Xu and Li | 30.48% | 33.86% | 40.29% | 24.67% |

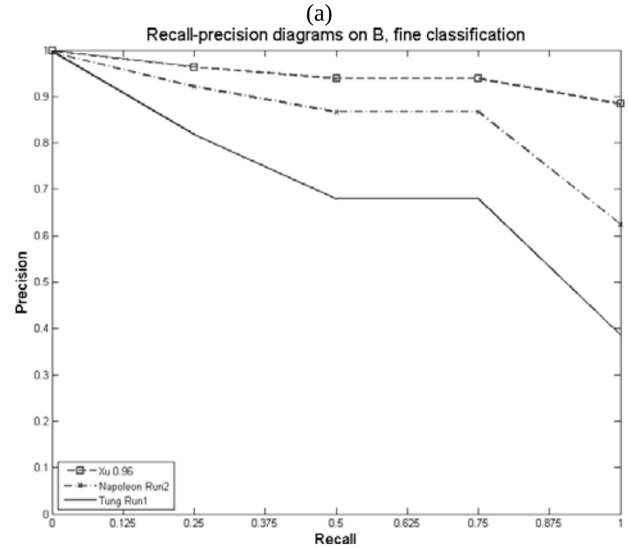
Table 1: The same type of pertubed models are used both as queries and dataset, so each class is made of 20 elements. The percentages represent the minimum on the different runs provided by the participants. Symbols: GN is Gaussian noise, SP is small protrusions, TN is topological noise and UR uneven re-sampling.

| **Degradation of the retrieval performance** | | | | |
|---|---|---|---|---|
| *Method* | *GN* | *SP* | *TN* | *UR* |
| Tung et al. | 53.68% | 50.03% | 48.11% | 39.05% |
| Napoleon et al | 45.20% | 53.96% | 48.56% | 38.04% |
| Xu and Li | 33.48% | 51.78% | 44.28% | 26.63% |

Table 2: The original models are used as queries against the corresponding perturbed models, the classes of both queries and dataset are made of 20 elements. The percentages represent the minimum on the runs provided by the participants. Symbols: GN is Gaussian noise, SP is small protrusions, TN is topological noise and UR uneven re-sampling.




(a)




(b)

Figure 2: Comparison of the best final recall precision graphs of each participant over the coarser (a) and the finer (b) classification.

## ACKNOWLEDGMENTS


The authors would like to thank Daniela Giorgi and Simone Marini for their support during the preparation of the datasets and the evaluation of the results. This work was partially supported by the FP6 IST Network of Excellence 506766 AIM@SHAPE.


## REFERENCES


[1] S. Biasotti and M. Attene. Stability on watertight models track. Technical Report 01, IMATI, Genova, Italy, 2008. available at http://shrec.ge.imati.cnr.it/shrec08_stability.html.

[2] D. Giorgi, S. Biasotti, and L. Paraboschi. Watertight models track. Technical Report 09, IMATI, Genova, Italy, 2007.

[3] G. Salton and M. McGill. *Introduction to modern information retrieval*. McGraw Hill, 1983.