TerraMobilita/iQmulus 3D urban analysis benchmark: evaluation protocol

Mathieu Brédif, Beatriz Marcotegui, Nicolas Paparoditis, Andres Serna, Bruno Vallet

16/04/2014

Introduction

The object of the TerraMobilita/iQmulus 3D urban analysis benchmark is to evaluate the current state of the art in urban scene analysis from point clouds acquired with a mobile mapping system. A very detailed semantic tree for urban scenes is provided in xml format. A representation of this tree is given in Figure 1. We call analysis the capacity of a method to separate the scene into these categories (classification), and to separate the different objects of the same type for object classes (detection). The proposed semantic tree is very detailed and probably no existing method treats the whole problem. This is why, the participants to the benchmark can choose whatever subtree of this tree. In this case, they will simply apply the "other" label to the nodes that they do not wish to detail. The evaluation will be performed accordingly and only the relevant metrics will be given. The benchmark aims at evaluating both the classification and detection quality.

1 Classification quality

The classification quality will be evaluated point-wise. The results of the evaluation will be a confusion matrix for each node of the tree that the evaluated method handles. Rows and lines will be the subclass labels from the ground truth and the evaluated method respectively, and matrix values are the percentage of points with the corresponding labels. All nodes from the semantic tree have an "other" class, so participants can classify into less classes than what is given in the tree. For non root nodes, an additional label "not in class" will be given for point that were not classified correctly at a lower level.

2 Detection quality

The detection quality work measures the capacity of the method to detect the objects present in the scene. Thus it requires to choose a criterion to decide if an object from the ground truth is detected or not. This biases the evaluation as this choice will impact the result. The solution that we propose is to give the evaluation result for a varying threshold m on the minimum object overlap. In this benchmark, an object is defined by the subset of points with the same object identifier. For a such subsets S^{GT} of the ground truth and S^{AR} of the evaluated algorithm result, we will validate S^{AR} as a correct detection of S^{GT} (a match) iff:

$$\frac{|S^{GT}|}{|S^{GT} \cup S^{AR}|} > m \text{ and } \frac{|S^{AR}|}{|S^{GT} \cup S^{AR}|} > m \tag{1}$$

where $|\cdot|$ denotes the cardinal (number of objects) of a set. The standard precision/recall are then functions of m:

$$precision(m) = \frac{\text{number of detected objects matched}}{\text{number of detected objects}}$$

$$recall(m) = \frac{\text{number of detected objects matched}}{\text{number of ground truth objects}}$$

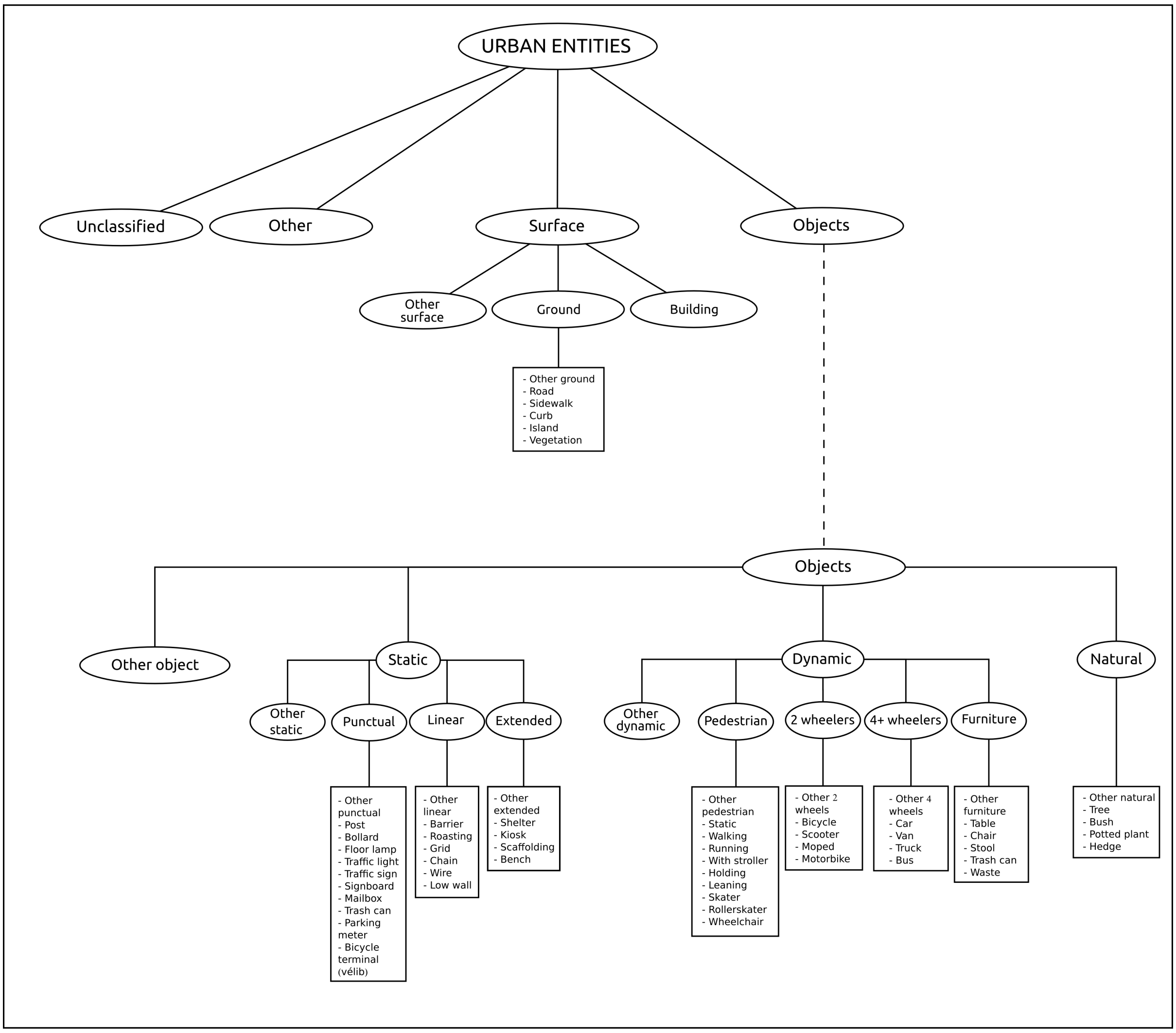

Figure 1: Semantic classes tree

Precision/Recall will be evaluated for each object types at each level of the semantic tree that the participants have handled and results will be presented as two curves. Precision/Recall are decreasing in m and this decay indicates the geometric quality of the detection (good geometry implies slower decay).

3 Segmentation quality

When the threshold m is below 0.5, the criterion (1) does not guarantee that objects are uniquely matched. When $m < 1/n$, n objects from the ground truth (GT) can be matched to a single object of the algorithm result (AR), or the opposite. Thus for $m < 0.5$ we will also give the curves of over-segmentation (1-to-n) and under-segmentation (n-to-1) by averaging n over the matches defined by (1).