SCALABLE EVALUATION OF 3D CITY MODELS

Oussama Ennafii[1,2]*, Arnaud Le Bris*[1]*, Florent Lafarge*[2]*, Clément Mallet*[1]

1: Univ. Paris-Est, LASTIG STRUDEL, IGN, ENSG, France – firstname.lastname@ign.fr
2: INRIA Sophia Antipolis, France – firstname.lastname@inria.fr

ABSTRACT

The generation of 3D building models from Very High Resolution geospatial data is now an automatized procedure. However, urban areas are very complex and practitioners still have to visually assess the correctness of these models and detect reconstruction errors. We proposed an approach for automatically evaluating the quality of 3D building models. It is cast as a supervised classification task based on a hierarchical taxonomy and multimodal handcrafted features (building geometry, optical images, height data). In this paper, we evaluate how the urban area composition impacts prediction transferability and scalability of our framework to unseen scenes. This allows to define minimal feature and training sets for a problem where no benchmark data has been released so far.

Index Terms— 3D, urban, buildings, quality, classification, error detection, geospatial imagery, Very High Resolution, scalability, transferability, representativeness.

1. INTRODUCTION

3D urban models have a wide range of applications covering in particular critical domains with significant societal challenges. 3D city modeling has therefore been deeply explored in the photogrammetric, GIS, computer vision, and computer graphics literature with an emphasis on compactness, full automation, robustness to acquisition constraints, scalability, and, inevitably, quality [1, 2, 3]. The problem remains partly unsolved [4, 5], since current algorithms lack of generalization capacity. They fail handling the significant heterogeneity of urban landscapes [6]. Human intervention is needed as a post-processing refinement and correction step, which is highly time-consuming (individual visual inspection of buildings). Consequently, automatizing the last step remains a critical step. Surprisingly, it has been barely investigated in the literature [7, 8, 9, 10]. In [11], we proposed a *semantic evaluation* framework in which building semantics is taken into account through the detection and categorization of modeling errors at the facet level for each 3D building. Our solution is independent from a given urban area, the Level of Details (LOD) of buildings, and the 3D modeling method. The problem is formulated as a supervised classification problem which predicts all errors affecting the building model. It is based on (i) a taxonomy of errors, hierarchical, adapted to all LODs, and independent from input models, and (ii) a multimodal handcrafted baseline of features that are extracted from the model itself, as well as from Very High Resolution external data (optical images and Digital Surface Models - DSM).

This paper focuses on the evaluation of the scalability of our framework. When dealing with non trivial remote sensing problems with limited training and testing sets, two major pitfalls exist: overfitting for a given area and divergent conclusions for various areas. This prevents from drawing strong conclusions of large applicability on the problem at stake. Here, using three distinct manually annotated urban areas in France, we analyze the *transferability*, *generalization*, and *representativeness* capacities of the proposed feature and training sets for all errors in the taxonomy. Eventually, this allows to define suitable feature and training sets for a problem where no benchmark data has been released so far.

2. SCALABILITY ANALYSIS

2.1. The evaluation framework

Our 3D building model evaluation framework consists in detecting errors, for every building, according to a hierarchical taxonomy of 9 atomic errors [11]. These errors correspond to two distinct levels:

- **The building level**: over segmentation (`BOS`), under segmentation (`BUS`), imprecise footprint borders (`BIB`), inaccurate footprint topology (`BIT`).
- **The facet level**: over segmentation (`FOS`), under segmentation (`FUS`), imprecise borders (`FIB`), inaccurate topology (`FIT`), imprecise geometry (`FIG`).

We disentangle fidelity and modeling errors and can evaluate LOD-0, 1, and 2 models. Error detection is performed with a supervised Random Forest classifier at three different *finesse* levels (0: qualifiable/non qualifiable – 1: valid/erroneous – 2: facet/building error – 3: 9 atomic errors). The process is based on a set of 60 features stemming from 3 different sources (20 each): the geometry of the model itself (*e.g.,* angle between adjacent facets), a VHR optical image (*e.g.,*

matching scores between image and building contours), and a Digital Surface Model (based on the histogram of the height residuals with the model). More details are available in [11], where one can see that all sources contribute equally, even if geometric features suffice to reach high accuracies.

2.2. Strategy

Experiments shown that the scene composition can affect greatly model error detection. This fact motivates studying training the classifier and testing prediction on different scenes. The goal is to prove the resilience of the prediction to unseen urban scenes. As the annotation process requires a lot of effort, this trait is crucial to guarantee the scalability of this method. Different configurations are evaluated (Figure 1).

- **Transferability**: we test how transferable are the learned classifiers from one scene to another. We train on area A_i and test on another one A_j. We will denote each transferability experiment by the couple (A_i, A_j) or by $A_i \rightarrow A_j$. $n(n-1)/2$ transferability couples are therefore possible, n being the number of areas.
- **Generalization**: we try to find out how omitting one area from the training set affects the results on the same area. We also aim to confirm the outcome of the transferability experiments. Experiments that fuse all areas except A_i ($\bigcup_{\forall j \neq i} A_j$) for training and test on A_i are noted by the couple $(\bigcup_{\forall j \neq i} A_j, A_i)$ or by $\bigcup_{\forall j \neq i} A_j \rightarrow A_i$. There are n possibilities.
- **Representativeness**: the objective is to find out, when mixing all labelled buildings from all n datasets, which amount of training data is required for a stable outcome, as well as how such ratio affects the test results if trained only on one type of cities. Different ratios between 20% and 70% are evaluated.

F-score metrics per error are selected for the evaluation of the various configurations. These experiments are also compared with the results obtained for each area (training+testing) with various feature sets [11]: 3D model geometry (Geom.), VHR optical image (Im.), and DSM (Hei.). For each error and problem, we compute the mean and the standard deviation of the F-score according to several possible feature sets and evaluate which contributes the most to the detection of the error. These are: Geom., Geom. $\cup$ Hei., Geom. $\cup$ Im., and All.

3. DATA

3D models from three different cities of France are evaluated (3,235 buildings in total): **Elancourt**, **Nantes** and the XIII$^{\text{th}}$ district of Paris (**Paris-13**). **Elancourt** exhibits a high diversity of building types: residential areas (hipped roof buildings) and industrial districts (flat roofs). **Nantes** represents a denser urban setting with lower building diversity. In Paris-13, high towers, with flat roof, coexist with Haussmann style buildings that typically exhibit highly fragmented roofs. The **Elancourt** (*resp.* **Nantes** and **Paris-13**) scene contains 2,009 (*resp.* 748 and 478) annotated building models. The spatial resolution of the DSM and the orthorectified images is 6 cm while it is 10 cm for the two other sets.

3D models were generated using [12], fed with existing building footprints and aerial VHR multi-view DSMs. The algorithm simulates possible LOD-2 roof structures from a predefined grammar with facets satisfying some geometric constraints. The best configuration is selected using a scoring system on the extrapolated roofs. Finally, orthogonal building façades connect the optimal roof to the ground. This method is adapted to roof types of low complexity and favors symmetrical models (residential areas). It has been selected to ensure a varying error rate for the three areas of interest.

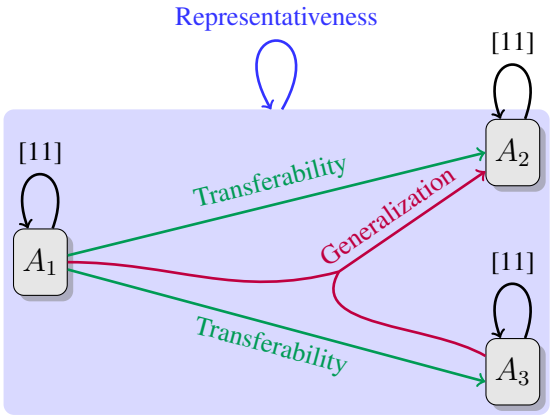


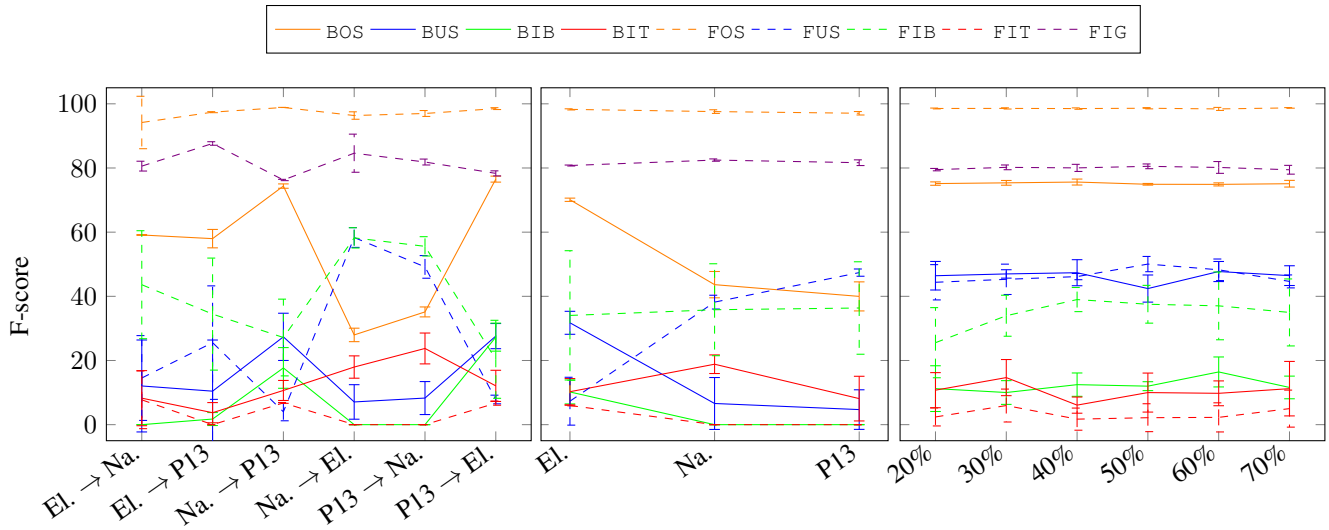
Fig. 1. Three-fold scalability analysis of our 3D building model evaluation framework. A_i indicates the area of interest. Arrow origins and targets give information about which area(s) is(are) considered as training and test sets, respectively. Results are compared with a baseline provided in [11].

4. EXPERIMENTS

Main results are reported in Figure 1 and Table 1.

4.1. Transferability

Three datasets lead to the evaluation of six transferability cases. In Figure 1a, one can see two distinct behaviours: stability across datasets and high fluctuations. In general, facet errors are more resilient and often yield better results than building errors. Fluctuations in F-score can be mainly noticed for the Nantes $\rightarrow$ Elancourt and Paris-13 $\rightarrow$ Nantes cases: `FUS` and `FIB` are better detected while a significant drop in performance exists for `BOS` (-30%). This stems from the limited diversity and size of training sets for Nantes and



(a) Transferability. (b) Generalization. (c) Representativeness.

Fig. 2. F-score mean and standard deviation values for the various tested feature sets for the three case studies. Experiments are performed for the nine atomic errors. B⋆⋆ and F⋆⋆ correspond to building and facet errors, respectively. El., Na., P13 correspond to the Elancourt, Nantes, and Paris-13 datasets, respectively.

Paris-13. This confirms the intuitive idea that dense urban area scenes are more helpful with LOD-1∪0 errors and topological ones (FIT).

Table 1 first shows that training on the testing area obviously leads to the best results, especially for building errors, even if the decrease in accuracy remains satisfactory in a large majority ($\sim$ 10%). Elancourt, with more heterogeneous facets, is pivotal for the facet errors (FUS, FIB). Nantes provides the most effective training set for building errors. Secondly, contrarily to the ablation study conducted in [11], image-based features and, with a lower impact, height attributes are often decisive in this experimental study. The geometry of the 3D model no longer suffices and remote sensing data is mandatory for transfer learning.

4.2. Generalization

Concerning the stability of the F-score values, similar conclusions to the transferability study can be drawn. Again, the framework performs better for the facet errors than the building ones in terms of generalization. Main decrease in accuracy (-25%) can be noticed for BOS and BUS errors when integrating Elancourt training data. Such a discrepancy in test scores between urban scenes proves the complementarity between the different datasets for most labels: dense city centers (Nantes, Paris-13) are really helpful for building errors and LOD1∪0, while residential areas with a large diversity and simplicity in roof types are particularly tailored as training sets for facet errors.

Similarly, Table 1 shows that better results are still obtained when predicting errors on the training area. Few improvements are noticed, namely for detecting the inaccurate footprint topology (BIT), facet imprecise borders (FIB), and under-segmentation (FUS). Again, the composition of the urban scene has a significant impact on such a variation, again promoting the necessity of creating hybrid training sets capturing the geographical diversity of urban environments. Eventually, similarly to the previous experimental setting, remote sensing features have a major role in order to get the best scores. In particular, VHR optical images and their high frequencies (contours) are highly helpful for inner roof discrepancies (BUS, BIB, FUS, FIB). Conversely, the Digital Surface Model is needed for outer and coarse errors (BOS, BIT, FIG).

4.3. Representativeness

Figure 1c shows (i) most of the time superior performance in error detection than the previous studies, and (ii) the extreme stability of all errors across different split ratios. This first indicates that merging training samples for three distinct urban environments is the most suitable solution for a relevant training set, as already proved in Sections 4.1 and 4.2. Better results are noticed for the errors BOS, BUS and FUS. Such errors correspond to topological issues within building roofs. They therefore require a large diversity of training samples in order to be correctly retrieved. Secondly, scalability in such conditions is ensured since no matter how small the training set is. No standard logarithmic behaviour can be noticed: 20% of the full labels are sufficient to retrieve all errors with an accuracy similar to training and predicting on a single area.

		BOS	BUS	BIB	BIT	FOS	FUS	FIB	FIT	FIG
Transferability	Elancourt → Nantes							Im.		
	Elancourt → Paris-13						Im.	Im.		
	Nantes → Paris-13									
	Nantes → Elancourt				All			Im.	Im.	
	Paris-13 → Nantes									Hei.
	Paris-13 → Elancourt			Im.				Im.		
General.	Elancourt		Im.				Im.	Im.	Geom.	Hei
	Nantes	All	Im.	Im.				Im.		
	Paris-13	All			Hei.			Im.		

Table 1. Evolution of the F-score value, for each error, between each tested configuration and the best result per area [11]. Feature sets having a significant impact on the classification results are mentioned. Otherwise, Geom. Im., and Hei. contribute equally. The color indicates the magnitude: ■: $[-45, -35\%[$ – ■: $[-35, -25\%[$ – ■: $[-25, 15\%[$ – ■: $[-15, 5\%[$ – ■: $[-5, 5\%[$ – ■: $[5, 15\%[$ – ■: $[15, 25\%]$ – □: statistics cannot be computed.

We can finally notice less variance in F-score per experiment. The variance vanishes with a more diverse dataset even if it is still larger than the variance in the ablation study [11]. No significant variation can be noticed when changing the feature set (Geom., Geom. ∪ Im., Geom. ∪ Hei., All). This explains why these experiments are not displayed in Table 1.

5. CONCLUSION

In this paper, we satisfactorily evaluated the scalability capacity of our 3D building evaluation framework over three distinct urban landscapes. Three dimensions of the problem were examined, namely transferability, generalization, and representativeness. The results first show that training samples stemming from very diverse urban environments are required to efficiently scale up this classification pipeline and detect errors with confusion close to the best results per area. In such a context, a limited number of samples is sufficient. Such findings allow to define a suitable training sample strategy in the future, especially in a context where fine-grained manual annotation of 3D data is complex. Secondly, experiments proved that multimodal remote sensing data is required for scalability purposes, while 3D building geometry was sufficient for a single area analysis. The next step consists in benchmarking major 3D city modelling techniques over various urban areas.

6. REFERENCES

[1] F. Lafarge and C. Mallet, "Creating large-scale city models from 3D-point clouds: a robust approach with hybrid representation," *IJCV*, vol. 99, no. 1, pp. 69–85, 2012.

[2] R. Cabezas, J. Straub, and J. W. Fisher, "Semantically-aware aerial reconstruction from multi-modal data," in *ICCV*, 2015, pp. 2156–2164.

[3] H. Zeng, J. Wu, and Y. Furukawa, "Neural procedural reconstruction for residential buildings," in *ECCV*, 2018.

[4] P. Musialski, P. Wonka, D. Aliaga, M. Wimmer, L. van Gool, and W. Purgathofer, "A survey of urban reconstruction," *Eurographics State of the Art Reports*, 2012.

[5] F. Rottensteiner, G. Sohn, M. Gerke, J. D. Wegner, U. Breitkopf, and J. Jung, "Results of the ISPRS benchmark on urban object detection and 3D building reconstruction," *ISPRS Journal*, vol. 93, pp. 256–271, 2014.

[6] M. Sester, L. Harrie, and A. Stein, "Theme issue "quality, scale and analysis aspects of urban city models"," *ISPRS Journal*, vol. 66, no. 2, pp. 155 – 156, 2011.

[7] H. Kaartinen, J. Hyyppä, E. Gülch, G. Vosselman, H. Hyyppä, L. Matikainen, A.D. Hofmann, U. Mäder, Å. Persson, U. Söderman, et al., "Accuracy of 3d city models: EuroSDR comparison," *ISPRS Archives*, vol. 36, no. 3/W19, pp. 227–232, 2005.

[8] L. Boudet, N. Paparoditis, F. Jung, G. Martinoty, and M. Pierrot-Deseilligny, "A supervised classification approach towards quality self-diagnosis of 3D building models using digital aerial imagery," *ISPRS Archives*, vol. 36, no. 3, pp. 136–141, 2006.

[9] J.-C. Michelin, J. Tierny, F. Tupin, C. Mallet, and N. Paparoditis, "Quality evaluation of 3D city building models with automatic error diagnosis," *ISPRS Annals*, vol. XL, no. 7/W2, pp. 161–166, 2013.

[10] B. Xiong, S. Oude Elberink, and G. Vosselman, "A graph edit dictionary for correcting errors in roof topology graphs reconstructed from point clouds," *ISPRS Journal*, vol. 93, pp. 227–242, 2014.

[11] O. Ennafii, A. Le Bris, F. Lafarge, and C. Mallet, "The necessary yet complex evaluation of 3D city models : a semantic approach," in *JURSE*, 2019.

[12] M. Durupt and F. Taillandier, "Automatic building reconstruction from a Digital Elevation Model and cadastral data: an operational approach," *ISPRS Archives*, vol. 36, no. 3, pp. 142–147, 2006.