# Land-use characterisation using Google Street View pictures and OpenStreetMap

Shivangi Srivastava, ,
Wageningen University & Research
Wageningen, Netherlands
shivangi.srivastava@wur.nl

Sylvain Lobry
Wageningen University & Research
Wageningen, Netherlands

Devis Tuia
Wageningen University & Research
Wageningen, Netherlands

John E. Vargas-Muñoz
University of Campinas
Campinas, Brazil

**Abstract**

This paper presents a study on the use of freely available, geo-referenced pictures from Google Street View to model and predict land-use at the urban-objects scale. This task is traditionally done manually and via photointerpretation, which is very time consuming. We propose to use a machine learning approach based on deep learning and to model land-use directly from both the pictures available from Google Street View and OpenStreetMap annotations. Because of the large availability of these two data sources, the proposed approach is scalable to cities around the globe and presents the possibility of frequent updates of the map. As base information, we use features extracted from single pictures around the object of interest; these features are issued from pre-trained convolutional neural networks. Then, we train various classifiers (Linear and RBF support vector machines, multi layer perceptron) and compare their performances. We report on a study over the city of Paris, France, where we observed that pictures coming from both inside and outside the urban-objects capture distinct, but complementary features.
*Keywords*: land-use classification, urban-objects, OpenStreetMap, Google Street View pictures, deep learning, convolutional neural networks.

## 1 Introduction and related work

In the last decade, satellite/aerial imagery has been used to perform land-cover classification with great accuracy as it is possible to identify various land-cover types using spectral information contained in the top view. On the contrary, land-use is a more complex task to be achieved using the overhead (aerial or satellite) perspective. For example, different land-uses could be identified with a building, such as a school, a supermarket, or a hospital. Moreover, a land-use class like "university"- often includes several types of land covers, such as buildings, green areas, or a pond. Therefore, to have an interpretation of the use of urban-objects, satellite/aerial imagery alone is insufficient. For this reason, we propose to use ground-based (i.e. terrestrial with side-view) pictures, which capture characteristic feature of urban-objects. In this study, we define an urban-object as an urban spatial construct with a clear boundary of its own, which could be a building (e.g. supermarket, hospital) or an open space (natural and man-made, e.g. forest , garden, park, stadium, cemetery).

Using geolocated ground based pictures to model urban landscapes is a rising trend in geospatial computer vision (Lefèvre et al. 2017). These pictures are generally either available on social media (Flickr, Instagram) or on online repositories for web-based projects like Geograph. Tracewski et al. (2017) used georeferenced pictures from online sources (Flickr, Panoramio, and Geograph) to map the land cover of the cities of London (United Kingdom) and Paris (France). But such sources have their limitations for land-use characterisation and suffer from the following issues: 1) The pictures' content mostly does not pinpoint a particular urban-object, but landscapes, and depict the perception of the user taking that picture (mostly touristic viewpoints, landscapes); 2) Platforms like Instagram provide personalized picture content (selfies, object zooms of flowers, pets, etc.) which is seldom a characteristic of the land-use of the urban-object; 3) Pictures are unevenly distributed in a city, with touristic areas having a great concentration, while few pictures are available for other important urban-objects like hospitals, universities, or industrial areas; 4) The orientation and position of the camera is not known, thus adding ambiguity about the accuracy of the geo-location and the picture content (Produit et al. 2014); 5) The data is not available evenly across major cities around the globe. For example, Geograph though suitable for land-cover analysis but still it cannot be used for land-use characterisation, as it has limited data for urban-objects, and is geographically limited to Great Britain, Ireland, and the Isle of Man.

"Google Street View" (GSV) pictures seem more promising in this respect, as they capture many urban-objects with precise geolocation and are objective, in the sense that they simply picture a street from a vehicle, without any personal touch. GSV pictures are evenly distributed across urban areas of most countries worldwide (in 2012, Google announced that it has covered 39 countries and about 3,000 cities). They also provide the possibility of extracting contextual information by varying parameters like: field-of-view, heading, and pitch (Google Street View Image API). In addition, GSV is updated every year, and historical data dating back to 2007 could be accessed across major cities. The privacy and safety concerns are addressed by blurring out personal content like faces, license plates, homes (on request). Figure 1 provides examples of GSV and other social media generated picture content for a given urban-object. Recent research has used GSV to assess physical changes in an urban area (Naik et al. 2017), to detect urban trees (Wegner et al. 2016) or to map land-cover within a city (Workman et al. 2017). The last work is probably the most significant for our paper: Workman et al. (2017) propose a methodology to use GSV pictures to train a model predicting land-use at the pixel level and with dense (public) ground truth provided by the New York City Department of City Planning. Despite the quality of the results, this method cannot be applied easily to new cities, as most do not have such high quality pre-processed labels. This kind of ground truth is not frequently updated because of economical reasons, availability of expert human resource as well as time involved. To apply the model trained in New York to new cities remains a hard task that requires model adaptation and complex machine learning pipelines (Yokoya et al. in press). To tackle the issue of lack of ground truth, we propose to complement GSV pictures with the labels extracted from OpenStreetMap (OSM). OSM is an open geographic data source, which provides annotations of land-use for various urban-objects in cities worldwide.

Specifically, In this paper, we propose a new way to tackle land-use classification using GSV pictures and OSM data. We first extract visual features from the GSV pictures using deep Convolutional Neural Network (CNNs) models and then use the features to train classifiers using the OSM labels at the urban-object level. This means that we consider multiple pictures for a single urban-object, i.e. all the GSV pictures surrounding the object (acquired by the Google car) and also those available within the object itself (pictures taken by various users). We studied a set of urban-objects from the city center of Paris (France), grouped in 13 fine grained land-use classes. Through this study, we were able to show that we can predict the land-use labels for various urban-objects. The results obtained are promising and open up new possibilities for automating the use of freely available data for land-use mapping and automatic updating.

## 2 Dataset

We chose the city of Paris, France for our case study. We work at the urban-object level, where we consider the task of classifying vector shapes corresponding to urban-objects into a single land-use class. The objects are those defined in OSM, and each object is labeled with one of the 13 land-use classes detailed in Table 1. Some of these 13 classes were defined by merging sub-classes based on their urban utility: for example, class "educational" was created by combining sub-classes like "school", "university" or "college". We then downloaded, for each urban-object, the corresponding GSV pictures using the Google API, for both within and outside locations (we will differentiate them as 'inside' and 'outside' pictures hereafter). For inside location of a given urban-object in our dataset, we extracted pictures where users uploaded GSV pictures. We analyzed the OSM polygon boundaries of the urban-objects for which 'outside' pictures taken by Google were available. For every OSM polygon boundary that was adjacent to a street we downloaded a picture looking at the center of that boundary. In order to determine if a boundary of an urban-object is adjacent to a street we used street vectorial data from the European Environmental Agency's Urban Atlas vector database. By this procedure, we obtained a dataset of 4249 labeled objects from OSM and corresponding 34261 pictures from GSV (Figure 1).

## 3 Method

Our proposed method is composed of two main steps. The first step involves extracting features for each GSV pictures using pre-trained CNNs. CNNs are state-of-the art approaches for many computer vision tasks like segmentation, classification and object detection (Goodfellow et al., 2016). A pre-trained networks is a model that has been trained using a large dataset for another task, but whose features provide very strong visual cues to train land cover / land-use classifiers (Lefèvre et al., 2017). In this work we evaluate two pre-trained CNN models as features extractors, namely Inception-V3 (Szegedy et al. 2016) and VGG-16 (Simonyan and Zisserman 2014), mostly because they were successfully applied to land-cover classification from geo-tagged pictures in (Xu et al. 2017) and (Workman et al. 2017). We retained as feature vectors the activations obtained from the penultimate layer of the CNNs and used them as inputs for the visual recognition task of discerning urban land-use classes. Specifically, we employed the following three pre-trained

Figure 1: a) OSM Layer, b) GSV outside pictures, c) GSV inside pictures, d) Social media pictures (source:Panoramio).

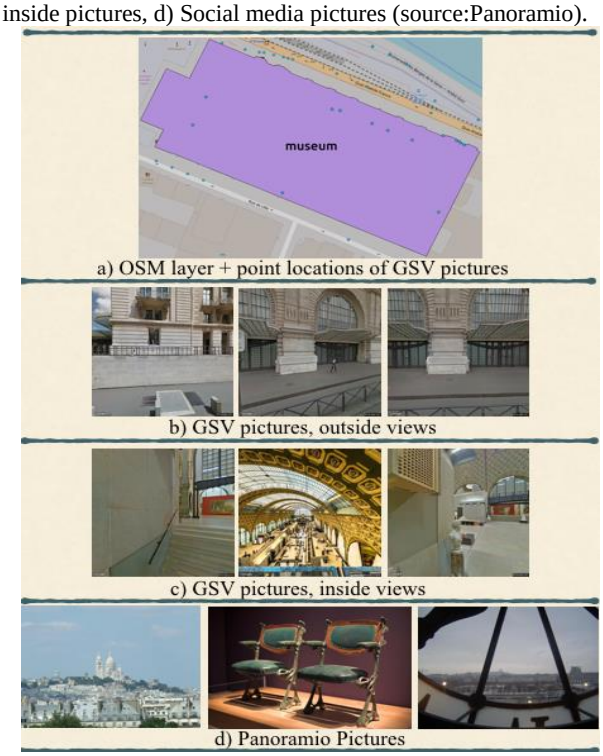



Table 1:Number of urban-objects, pictures per land-use class

| Land-use | # Urban-objects | # Pictures |
|---|---|---|
| Parks | 1892 | 14579 |
| Sports | 311 | 1630 |
| Heritage | 153 | 1791 |
| Forest | 177 | 2937 |
| Educational | 622 | 3868 |
| Medical | 71 | 1062 |
| Government | 279 | 2063 |
| Religious | 212 | 1131 |
| Cemetery | 51 | 460 |
| Parking | 126 | 1358 |
| Industrial | 69 | 954 |
| Hotel | 252 | 1246 |
| Marina | 34 | 1182 |

CNNs for feature extraction:
1) Inception-v3 trained on the ImageNet dataset: this configuration provides a 2'048-dimensional feature (named as Inception in tables 2a, 2b and 2c) per GSV picture.
2) VGG-16 also trained on the ImageNet dataset: this configuration provides a 4'096 dimensional feature vector (named as VGG16I in tables 2a, 2b and 2c) per GSV picture.
3) VGG-16, but trained on the Places365 dataset (Zhou et al. 2017): this also provides 4'096 dimensional feature vector (named as VGG16P in tables 2a, 2b, 2c) per GSV picture.

The reason for exploring models trained on two different datasets in VGG-16 (ImageNet and Places365) is to study whether a model trained on ImageNet (which is an object-centric dataset, i.e. a model trained to recognize objects within the pictures) is more appropriate for our task than using a

model trained on the Places365 dataset (which is a scene-centric dataset, i.e. a model trained to categorize the whole picture into a class).

The second step of our method performs the classification into land-use at the urban-object level. Given a feature vector extracted by the CNN, we train a land-use classifier aiming at predicting the land-use label of each urban-object. Note that we have several GSV pictures associated with each urban-object: though single pictures alone may not be able to identify the land-use, the ensemble of these pictures could provide sufficient richness of information to enable the classifier to learn the correct land-use of the urban-object. For example, both a park and a forest will have pictures of trees, but the park may also have pictures with people having picnics, children playing or people running. A university and a museum may have similar appearance in pictures taken from outside of them, but the inside pictures will help discerning the two. Therefore, we implemented and compared two variants of the aggregation of the feature extracted from single GSV pictures belonging to an urban-object.

Variant-1 (Fig. 2) consists in classifying the features of each GSV picture and then taking the majority prediction over all the pictures for an urban-object as the final prediction;

Variant-2 (Fig. 3) consists in averaging the features coming from all the GSV pictures of an urban-object and then training the classifier with the averaged feature.

In both cases of the variants, we compared three classifiers: 1) Support Vector Machines with Linear kernel (SVM-Linear), 2) Support Vector Machines with radial basis function (SVM-RBF), and 3) Multi-Layer Perceptron (MLP).

Figure 2: Flowchart for Variant-1

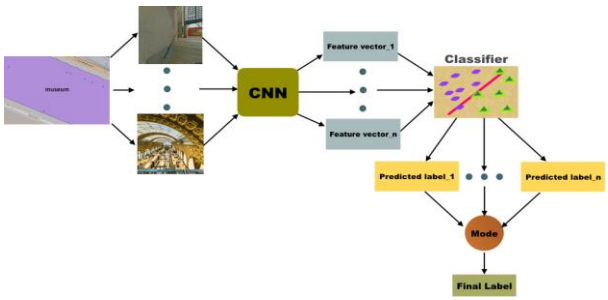
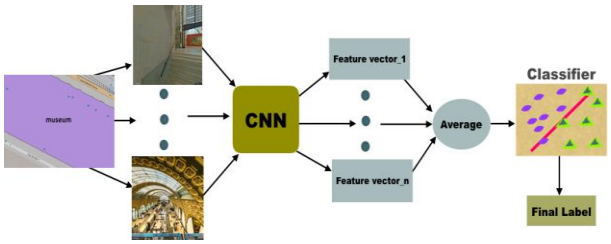



Figure 3: Flowchart for Variant-2




## 4 Experiments and Results

The data available for each class was split in the ratio of 3:1 for train and test, respectively. As it can be seen in Table 1, our dataset suffers from class imbalance. To tackle it effectively, we trained the classifiers with class weights inversely proportional to the number of samples for that class. For all SVM classifiers, a grid search was performed to obtain the best parameters. The MLP (with a hidden layer of 128 neurons) was run during 250 epochs with initial learning rate of 0.0001 and a batch size of 20. Below, we report the average of overall accuracy (OA) and average accuracy (AA) over five different train and test splits of the dataset in the tables. Tables 2a, 2b and 2c show OA and AA for combinations of picture locations (Table 2a: inside pictures only, Table 2b: outside pictures only, Table 2c: inside and outside pictures together), feature type, aggregation methods, and classifiers. OA1 and AA1 are the accuracies using variant-1. OA2 and AA2 are the results for variant-2.

As far as AA is concerned, we observed that averaging features (variant-2) leads to considerably better performances than the approach using features for individual pictures (variant-1), see Table 2b and 2c, AA2 is clearly better than AA1. Variant-1 is suboptimal because different land-uses (e.g., forest and park) may contain common objects (e.g., trees, muddy paths, animals), making it difficult to distinguish between land-use classes, for example if the classifier learns to predict pictures with objects like trees as pertaining to the "forest" class, then a land-use class like park will need to have more pictures of other objects (like benches, fountain, concrete pathways, open spaces) than trees to be able to predict correctly the label "park". Also in variant-1, the objects found in different pictures for an urban-object are not bound together while training, thus the contextual information needed to learn a land-use class is missing: in other words, in variant-1 each picture is classified into a land-use class independently, while land-use can be understood only by looking at the ensemble of pictures per urban-object. The approach of aggregating features (variant-2) is more appropriate, mostly because the averaged feature vector contains simultaneously information about the various objects that characterise a land-use class: for example, an urban-object from the class "religious place" could be represented by objects like statues, benches or candles. For variant-2, OA2 is generally higher (see Table 2b, 2c) than OA1 of variant-1 except for the case of using inside pictures only. When the classifiers are trained with both inside and outside pictures together they perform better than just with outside pictures alone. OA for the 'inside pictures' experiments are close to those of the model using both 'inside and outside pictures' simultaneously (compare OA1 and OA2 of Table 2a with corresponding entry in Table 2c). However, also note that direct comparison is delicate, since the number of samples of 'inside pictures' is smaller than 'inside and outside pictures' (thus size of test set is smaller too). This is due to the fact that not all urban-objects come with inside pictures in GSV.

In general, the classifiers trained with features extracted from Inception-v3 perform better than those with features extracted from other networks, which are VGG16-I and VGG16-P (see bold OA and AA in Tables 2a, 2b,2c). Features extracted from VGG-16 trained on ImageNet and on Places365 are comparable in performance for various classifiers except for MLP where VGG16-P is better than

VGG16-I in AA. We also wanted to experiment with Inception-v3 architecture, trained on Places365: unfortunately, we could not validate this model in this paper, as the pre-trained model for Inception-v3 with Places365 dataset is not available and our dataset is not sufficient to train such a large CNN model from scratch. As an extension of this work in future, we will fine-tune pre-trained CNN models with our land-use characterisation. Regarding the classifiers, MLP tends to obtain a better AA, whereas SVM-RBF classifier performs better in terms of OA but yields poor AA (Table 2a, 2b, 2c). This can be explained by the fact that SVM RBF obtains poor class accuracy when classifying classes "heritage", "forest", "medical" (below 30%), and "parking", "cemetery" and "hotel" (below 40%).

Table 2a: Inside pictures for urban-objects

| Feature type | Classifier | OA1 | AA1 | OA2 | AA2 |
|---|---|---|---|---|---|
| Inception | RBF | 74.15 | 47.19 | **68.37** | 46.70 |
| | Linear | **74.16** | 47.89 | 65.39 | 48.02 |
| | MLP | 63.73 | **49.90** | 51.12 | 49.91 |
| VGG16I | RBF | 69.73 | 41.08 | 65.14 | 42.54 |
| | Linear | 65.35 | 37.43 | 63.21 | 42.94 |
| | MLP | 66.25 | 42.14 | 60.00 | 47.52 |
| VGG16P | RBF | 67.78 | 40.84 | 67.95 | 45.97 |
| | Linear | 66.11 | 39.61 | 64.74 | 44.51 |
| | MLP | 68.69 | 46.81 | 58.29 | **50.35** |

Table 2b: Outside pictures for urban-objects

| Feature type | Classifier | OA1 | AA1 | OA2 | AA2 |
|---|---|---|---|---|---|
| Inception | RBF | **64.46** | 35.95 | **67.64** | 46.55 |
| | Linear | 63.03 | 35.88 | 61.18 | 45.15 |
| | MLP | 57.92 | **46.47** | 55.2 | **54.38** |
| VGG16I | RBF | 59.70 | 26.42 | 64.22 | 43.45 |
| | Linear | 57.79 | 28.20 | 60.43 | 39.36 |
| | MLP | 59.00 | 32.82 | 60.42 | 46.69 |
| VGG16P | RBF | 61.01 | 28.00 | 66.42 | 44.02 |
| | Linear | 58.86 | 32.41 | 61.52 | 41.48 |
| | MLP | 59.78 | 38.88 | 59.60 | 49.35 |

Table 2c: Inside and outside pictures for urban-objects

| Feature type | Classifier | OA1 | AA1 | OA2 | AA2 |
|---|---|---|---|---|---|
| Inception | RBF | 66.65 | 42.24 | **68.79** | 49.10 |
| | Linear | **69.05** | 48.11 | 62.77 | 50.08 |
| | MLP | 58.05 | **50.52** | 55.88 | **56.77** |
| VGG16I | RBF | 62.75 | 34.06 | 65.15 | 43.17 |
| | Linear | 59.20 | 31.22 | 61.31 | 43.50 |
| | MLP | 60.90 | 36.99 | 60.24 | 49.93 |
| VGG16P | RBF | 60.22 | 34.83 | 66.97 | 48.14 |
| | Linear | 60.22 | 34.83 | 62.97 | 47.84 |
| | MLP | 60.88 | 41.12 | 59.61 | 53.56 |

## 5 Conclusion

In this paper, we proposed a machine learning method to predict land-use at the urban-object level using freely available ground based photography (Google Street View) and crowdsourced land-use information (OpenStreetMap). The preliminary results obtained on a case study over Paris, France, permits us to provide a series of guidelines: first, we recommend to exploit the complementarity of outside and inside pictures (when available), to aggregate pictures of a same urban-object at the feature level rather than after single picture classification. Secondly, regarding the classifier, the choice between MLP and SVM-RBF depends on the desired results; although SVM-RBF can correctly predict labels for a larger number of urban-objects, it is inferior in average accuracy, while MLP is more consistent among the different classes. In the future, we will extend this study to more urban areas worldwide using fine-tuned CNN models.

## References


Goodfellow, I., Bengio, Y., & Courville, A. (2016) Deep Learning. *MIT Press.*

Lefèvre, S., Tuia, D., Wegner, J.-D., Produit, Ti. & Nassaar, A. (2016) Toward Seamless Multiview Scene Analysis From Satellite to Street Level. *Proceedings of the IEEE*, 105(10), 1884-1899.

Naik, N., Kominers, S.-D., Raskar, R., Glaeser, E.-L. & Hidalgo, C.-A. (2017) Computer vision uncovers predictors of physical urban change. *Proceedings of the National Academy of Sciences of the United States of America*, 114(29), 7571-7576.

Produit, T., Tuia, D., de Morsier, F. & Golay, F. (2014) Do geographic features impact pictures location shared on the Web? Modeling photographic suitability in the Swiss Alps. *Environnmental Multimedia Retrieval co-located with ACM International Conference on Multimedia Retrieval*, 1222, 22-29.

Simonyan, K. & Zisserman, A. (2014) Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv preprint arXiv:1409.1556.*

Szegedy, C., Vanhoucke, V., Ioffe, S. & Shlens, J. (2016) Rethinking the Inception Architecture for Computer Vision. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA*

Tracewski, L., Bastin, L. & Fonte, C.-C. (2017) Repurposing a deep learning network to filter and classify volunteered photographs for land cover and land use characterization. *Journal, Geo-spatial Information Science*, 20(3), 252-268.

Wegner,J.-D., Branson, S., Hall, D., Schindler, K. & Perona, P. (2016) Cataloging public objects using aerial and street-level images-urban trees. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6014-6023.

Workman, S., Zhai, M., Crandall, D.-J. & Jacobs, N. (2017) A Unified Model for Near and Remote Sensing. *IEEE International Conference on Computer Vision. Venice, Italy.*

Xu, G., Zhu, X., Fu, D., Dong, J. & Xiao, X. (2017) Automatic land cover classification of geo-tagged field photos by deep learning. *Environmental Modelling & Software*, 91, 127-134.

Yokoya , N., Ghamisi , P., Xia , J., Sukhanov, S., Heremans, R., Debes , C., Bechtel , B., Le Saux , B., Moser, G., & Tuia, D. Open data for global multimodal land use classification: Outcome of the 2017 IEEE GRSS Data Fusion Contest. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, in press.

Zhou, B., Lapedriza, A., Khosla, A., Oliva, A. & Torralba, A. (2017) Places: A 10 million Image Database for Scene Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, in press.