

# Building Instance Classification using Social Media Images

Eike Jens Hoffmann<sup>1</sup>, Martin Werner<sup>2</sup>, Xiao Xiang Zhu<sup>1,2</sup>

<sup>1</sup>Signal Processing in Earth Observation (SiPEO), Technical University of Munich (TUM)

<sup>2</sup>Remote Sensing Technology Institute (IMF), German Aerospace Center (DLR), Wessling, Germany  
eike.jens.hoffmann@tum.de, martin.werner@dlr.de, xiaoxiang.zhu@dlr.de

**Abstract**—Understanding urbanization and planning for the upcoming changes require detailed knowledge about the places where people live and work. Thus, knowing the usage of buildings is inevitable to distinguish between residential and commercial places. Assessing the usage of buildings from an aerial perspective alone is challenging and results in unresolvable ambiguities. As complementary data sources, social media images taken from ground level allow access to the building façades, as well as ongoing social activities around the buildings, which are very valuable information while coming to accessing the building usages. Towards the fusion of social media images and remote sensing data for this purpose, in this work we present a method to assess building usages from social media images taken in their neighborhood. Using a straight forward next neighbor classifier for mapping images to buildings and pre-trained networks for dimensionality reduction we trained a logistic regression classifier to distinguish between five different building usage classes. Applied to a study area of Los Angeles metropolitan area, USA, we obtain an average precision of 0.67. Hence, we show that social media images can be a valuable additional source to remote sensing data.

**Index Terms**—Building Classification, Social Media, Building Usage, Social Media Image, Complementary Data Source

## I. INTRODUCTION

The United Nations estimate that 68 % of the global human population will live in cities by 2050 [1]. Thus, metropolitan areas will grow in general and especially informal settlements will expand. To manage this growth and to develop cities accordingly, detailed insights into urban dynamics are necessary. Sustainable urban planning requires detailed morphological and cartographic information as well as insights about the population dynamics including population densities and hot spots.

For the creation of knowledge about the places where people live, work, and buy everyday necessities remote sensing is a crucial source of information. However, the aerial view cannot reveal every detail of urban areas: unresolvable ambiguities remain, which can only be determined from ground view.

Land-use classification in urban areas on building instance level is still a challenging task [2]. Buildings of different functions stand side by side and therefore, estimating their

This work is jointly supported by the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation programme (grant agreement No. [ERC-2016-StG-714087], Acronym: *So2Sat*), Helmholtz Association under the framework of the Young Investigators Group “SiPEO” (VH-NG-1018, [www.sipeo.bgu.tum.de](http://www.sipeo.bgu.tum.de)), and the Bavarian Academy of Sciences and Humanities in the framework of Junges Kolleg.

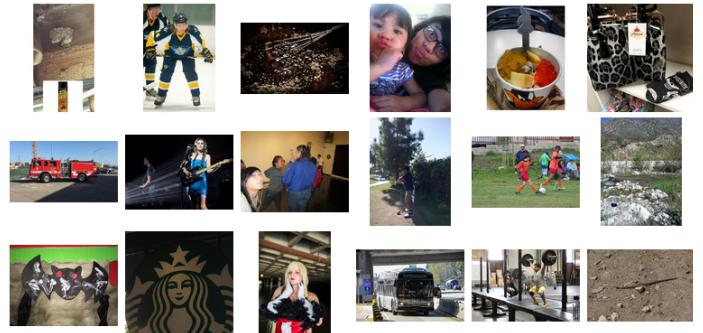


Figure 1: Examples of geo-tagged Flickr images

functions using optical satellite images is virtually impossible. Hence, Workman et al. proposed a unified model for near and remote sensing by fusing panoramic images from Google Street View and Bing Maps aerial [3]. For a fine-grained building instance classification Kang et al. used Google Street View images with a dedicated heading and sophisticated filtering [4].

Beyond Google Street View there are several social media platforms to which their users upload millions of photos every day. This is an additional and rich source of ground view imagery with various types of motifs included (cf. Figure 1). These images contain information about locations and activities of their users and hence bringing spatial and semantic knowledge together in one context. In this paper, we propose a framework to estimate building instance functions using plain geo-tagged social media images not containing any heading information. We show the effectiveness of our method on a study size of 8,270 km<sup>2</sup> including 2.6 million buildings around Los Angeles.

## II. RELATED WORK

Land-use classification using remote sensing images has been an active field of research adapting progress from computer vision and pattern recognition rapidly. Traditionally, it was mainly adopting feature extraction methods like SIFT [5] with bag-of-words approaches [6].

With the evolution of deep learning methods, new approaches for dealing with remote sensing images are being presented [7]. Chen et al. showed how stacked autoencoders for hierarchical feature extraction can be used for land use

classification in hyperspectral images [8]. Since labeled data in remote sensing on deep learning is very scarce, Marmanis et al. [9] used features extracted from pre-trained networks on ImageNet [10] to classify aerial images. Beside CNNs, recurrent neural networks and generative adversarial networks are also employed for different remote sensing tasks, for example in [11]–[13].

Parallel to developments on aerial imagery there has been a lot of research on geo-tagged images from Google Street View and social networks like Flickr. Leung and Newsam used a combination of Flickr and the Geograph British Isles project<sup>1</sup> for a binary land coverage classifier. This method called *Proximate Sensing* achieved an accuracy of 75 % on their study area covering the London metropolitan area and its surroundings.

Gebru et al. used Google Street View images with a deep learning method to classify car models [14]. These classification results were used as a proxy for estimating income, race, education, and voting patterns in the United States. Furthermore, Kang et al. used Google Street View images to classify building usage on single buildings using façade images. They filtered out all images with occlusions or indoor scenes and fine-tuned four pre-trained networks to distinguish between eight usage classes obtained from OSM.

Moreover, hybrid datasets like the Cross-View USA (CVUSA) including geo-tagged images from Google Street View and Flickr [15] provide a valuable source of information for land cover estimation. Using a pre-trained object detection algorithm, Greenwell et al. presented spatial correlations between detected objects and land coverage [16].

In contrast to the given related work, we use unstructured social media imagery from Flickr and do not apply any filtering of images in order to predict building instance classes. More concretely, we address the problem of building instance classification into five classes given by OSM. We solely rely on unfiltered geo-tagged social media images from Flickr. We exploit the spatial proximity of images and buildings indicating a possible semantic relation knowing and expecting that many images do not directly relate to the given classification task. Still, we can show that the overall spatial distribution of images is tightly linked to the building functions.

### III. METHODOLOGY

In general, we are going to create a classification system which can be given a set of images assigning one of the classes. Therefore, we need several steps: First, we need to associate the geo-tagged imagery with individual building instances. Our approach for this aspect is outlined in Section III-A.

Then, for all images that are sufficiently near to a building, we extract a visual feature vector from applying a pre-trained vision network. Concretely, we employ VGG-16 [17] for its good balance of visual performance and prediction speed. The extracted feature vectors are then used in a logistic regression

setup trying to predict the class of the building a set of images was associated to. Multiple results stemming from multiple images are combined by using a voting scheme. This prediction step is subsumed in Section III-B.

#### A. Assigning images to buildings

The assignment of an image to a building is based on a spatial nearest neighbor join. Intuitively, we assume that geo-tagged images next to the same building share some context and, therefore, we map each image to the nearest building. An image can be assigned to only one building, but one building can have multiple images assigned to it.

More formally: let  $I$  be a collection of pairs  $(i, p)$ , where  $i$  is an image and  $p$  is a spatial point representing the location of the image. Let further  $B$  be a collection of pairs  $(b, c)$  of a building polygon  $b$  and a class  $c$  representing its usage according to OSM. Then, we perform a spatial nearest neighbor left join on  $I \times B$  resulting in a set of tuples  $(i, c, d)$  assigning to each image  $i$  the class  $c$  of the nearest building to the image location as well as the distance  $d$  between the building polygon and the image location.

The closer the distance between an image point and a building polygon the more likely is a semantic relationship. Therefore, we define a distance threshold  $d_{thres}$  discarding all images, which are assigned to a building that is more than  $d_{thres}$  meters away.

#### B. Predicting building functions

For all images that have been selected in the previous step we compute feature vectors,  $f(i)$ , from a pre-trained network. For this, we use the VGG-16 architecture [17] trained on ImageNet [10] and extract the fully connected layers  $fc1$  and  $fc2$  in front of the prediction layer. The feature vectors obtained from these layers have 4096 dimensions.

This instantiates a classification problem of mapping one of these feature vectors to the building class. We decide to apply a logistic regression classifier trained using the SAGA optimizer, which is known to work well if the feature space is high-dimensional, the training data is very small, and the dataset is large [18]. It applies a certain form of stochastic gradient decent and it is known to converge very fast and even for objective functions that are not strictly convex.

The classifier

$$cl_{i \rightarrow c} : i \xrightarrow{VGG-16} fc1 \xrightarrow{\text{Log. Reg. w. SAGA}} c$$

is trained on pairs of images and classes to distinguish between the different types of usage.

For our final goal, however, we need to assign a class to each building and we do so by combining the results for all images assigned to a certain building through majority voting:

$$cl_{b \rightarrow c} : b \xrightarrow{\text{lookup images}} (i_k) \xrightarrow{cl_{i \rightarrow c}} (c_k) \xrightarrow{\text{maj. vote}} c$$

In this classifier, the first operation extracts all images that have selected the given building as its nearest neighbor, applies the image to class classifier, and combines the results selecting the most frequent class among the results of the image

<sup>1</sup><http://www.geograph.org.uk>




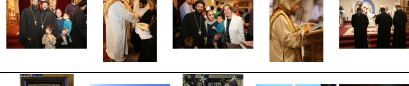
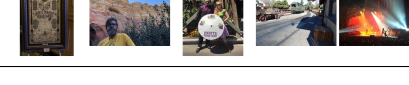
Class	Sample Images
a) accommodation	
b) civic	
c) commercial	
d) religious	
e) other	

Figure 2: Example images from five different usage clusters

to class classifier. To estimate the classifier’s performance thoroughly, we apply a 10-fold cross-validation to train our logistic regression classifier.

#### IV. EVALUATION

First, we give a brief introduction to the Los Angeles dataset used for training and testing. Then, we show the results of our methodology starting with assigning the images to buildings and discussing our classification results with different parameters in the end.

##### A. Dataset

Our study area focuses on the Los Angeles metropolitan area because it has a very dense and convincing building function assignment in OSM. 2,619,306 building polygons in this area of 8,270 km<sup>2</sup> have a building tag corresponding to the proposed labeling schema by OSM<sup>2</sup>. This labeling schema is clustered into five high level classes: “accommodation”, “civic”, “commercial”, “religious”, and “other”. Consequently, these building tags can be mapped to one of the high level classes. Figure 2 depicts example images for all classes. For our study area we collected 343,711 public, geo-tagged images from Flickr.

##### B. Assigning images to buildings

Since the assignment algorithm maps each image to a building no matter how far away it is, the distribution of assignment distances follows a log-normal distribution. There are many images with a small distance to the next building and few buildings with a larger distance. Figure 3 shows the distribution of assignment distances on a log-scale. 56.6 % of all images are less than 100 m away from the next building, 89.3 % are within a 1,000 m distance threshold. The median distance is 75.6 m.

Figure 4 depicts the spatial join assignments as lines on a map of Los Angeles area. As expected, the spatial density of

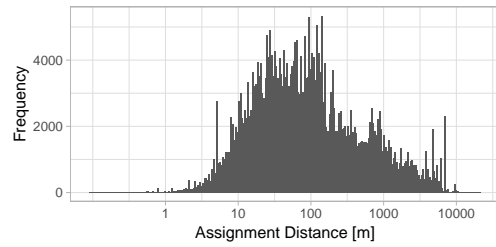


Figure 3: Distribution of Spatial Join Distances

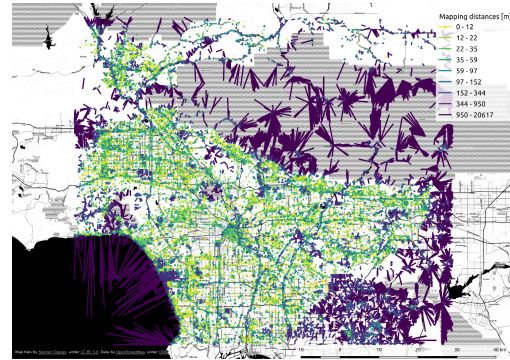


Figure 4: Image assignment distances (color = distance)

geo-tagged images from Flickr is higher around the city center (about 97 m) as opposed to the rural areas (up to 20,617 m). These long lines stem from assigning images on the ocean or from hiking in the mountains to the nearest classified building.

In summary, 343,600 images are assigned to 34,318 buildings with proper class labels. Figure 5 shows the distribution of the number of images assigned to a building and its frequency. 17,029 buildings have a single image assigned to it, 17,298 have more than one image for classification. A fraction of 4,341 buildings has 10 or more images assigned to it.

##### C. Predicting building functions

First, we present experiments regarding the choice of the distance threshold  $d_{thresh}$ . Figure 6 shows the classification performance of the framework for choosing both fc1 and fc2, respectively. The classification performance is measured with precision, recall, and accuracy as a function of the distance threshold  $d_{thresh}$ . Additionally, they depict the relative number of classified buildings compared to the number of buildings that have at least one image assigned to it.

Concretely, the fraction of images that are actually being labeled from the OSM data varies significantly and starts to

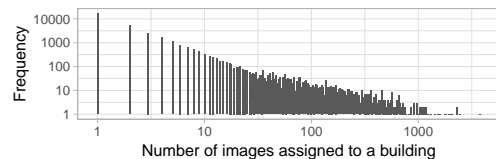


Figure 5: Number of Images per Building (log-log-scale)

<sup>2</sup><https://wiki.openstreetmap.org/wiki/Key:building>

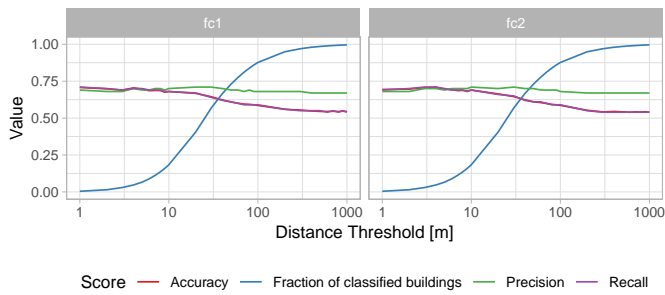


Figure 6: Classification Performance for Features using Layers fc1 and fc2

True Label \ Predicted Label	Accommodation	Civic	Commercial	Other	Religious
Accommodation	10123	1062	4220	2014	25
Civic	260	192	248	114	1
Commercial	2442	538	6991	1192	5
Other	166	19	139	323	1
Religious	7	2	4	4	3

Figure 7: Confusion Matrix. Background color visualizes column-normalized precision.

stabilize around 100 m with 87.7 % (18.3 % at 10 m, 66.7 % at 40 m). In addition, increasing the distance does reduce precision and recall, but not significantly. Note that recall is measured in relation to the dataset of assigned images, not in relation to the number of images that would be available. In addition, we observe that the performance of using fc1 or fc2 is comparable.

In summary, the performance fluctuates around 75 % for precision, recall, and accuracy and starts to reduce with increasing distance threshold at about 20 m. However, increasing the distance threshold further increases coverage in the sense that more buildings get images. Therefore, we propose to use a threshold of 100 m for this tradeoff. For the final setting of  $d_{thres}=100$  m, Figure 7 depicts a confusion matrix. In summary, the most frequent classes accommodation and commercial show convincing performance of 0.58 and 0.63. This is a very encouraging performance. In the future, we envision to use this information for augmenting classification based on satellite imagery.

## V. CONCLUSION AND OUTLOOK

In this study we presented a classifier for predicting building usage based on spatial proximate social media images. In the Los Angeles metropolitan area our method is able to classify 87.6 % of all buildings that have at least one image in a proximity of 100 m with an average precision of 0.68.

For future work, we envision three important directions: first, use this in a fusion manner together with spatially

accurate yet semantically ambiguous remote sensing imagery. Second, replace the simple kNN associations with a higher degree probabilistic spatial association rule such that images contributes to more than one building. Finally, we envision to use filtering and motif detection in order to reduce the amount of misleading images from the classification system.

## REFERENCES

- [1] U. Nations, *World urbanization prospects*. 2014.
- [2] I. Demir, K. Koperski, D. Lindenbaum, G. Pang, J. Huang, S. Basu, F. Hughes, D. Tuia, and R. Raskar, “Deepglobe 2018: A challenge to parse the earth through satellite images,” *ArXiv e-prints*, 2018.
- [3] S. Workman, M. Zhai, D. J. Crandall, and N. Jacobs, “A unified model for near and remote sensing,” in *The IEEE International Conference on Computer Vision (ICCV)*, pp. 2707–2716, 2017.
- [4] J. Kang, M. Krner, Y. Wang, H. Taubenbck, and X. X. Zhu, “Building instance classification using street view images,” *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 145, pp. 44 – 59, 2018.
- [5] D. G. Lowe, “Object recognition from local scale-invariant features,” *Proceedings of the Seventh IEEE International Conference on Computer Vision*, vol. 2, no. 8, pp. 1150–1157, 1999.
- [6] Q. Zhu, Y. Zhong, B. Zhao, G.-S. Xia, and L. Zhang, “Bag-of-Visual-Words Scene Classifier With Local and Global Features for High Spatial Resolution Remote Sensing Imagery,” *IEEE Geoscience and Remote Sensing Letters*, vol. 13, no. 6, pp. 747–751, 2016.
- [7] X. X. Zhu, D. Tuia, L. Mou, G.-S. Xia, L. Zhang, F. Xu, and F. Fraundorfer, “Deep learning in remote sensing: a comprehensive review and list of resources,” *IEEE Geoscience and Remote Sensing Magazine*, vol. 5, no. 4, pp. 8–36, 2017.
- [8] Y. Chen, Z. Lin, X. Zhao, G. Wang, and Y. Gu, “Deep Learning-Based Classification of Hyperspectral Data,” *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 7, no. 6, pp. 2094–2107, 2014.
- [9] D. Marmanis, M. Datcu, T. Esch, and U. Stilla, “Deep Learning Earth Observation Classification Using ImageNet Pretrained Networks,” *IEEE Geoscience and Remote Sensing Letters*, vol. 13, no. 1, pp. 105–109, 2016.
- [10] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, “ImageNet Large Scale Visual Recognition Challenge,” *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.
- [11] H. Lyu, H. Lu, L. Mou, W. Li, J. Wright, X. Li, X. Li, X. X. Zhu, J. Wang, L. Yu, and P. Gong, “Long-term annual mapping of four cities on different continents by applying a deep information learning method to landsat data,” *Remote Sensing*, vol. 10, no. 3, pp. 1–23, 2018.
- [12] N. Merkle, S. Auer, R. Miller, and P. Reinartz, “Exploring the potential of conditional adversarial networks for optical and sar image matching,” *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 11, no. 6, pp. 1811–1820, 2018.
- [13] L. H. Hughes, M. Schmitt, and X. X. Zhu, “Mining hard negative samples for sar-optical image matching using generative adversarial networks,” *Remote Sensing*, vol. 10, no. 10, pp. 1–17, 2018.
- [14] T. Gebru, J. Krause, Y. Wang, D. Chen, J. Deng, E. L. Aiden, and L. Fei-Fei, “Using deep learning and Google Street View to estimate the demographic makeup of neighborhoods across the United States,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 114, no. 50, pp. 13108–13113, 2017.
- [15] S. Workman, R. Souvenir, and N. Jacobs, “Wide-Area Image Geolocalization with Aerial Reference Imagery,” in *2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 3961–3969, 2015.
- [16] C. Greenwell, S. Workman, and N. Jacobs, “What goes where: Predicting object distributions from above,” *arXiv preprint arXiv:1808.00995*, 2018.
- [17] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *arXiv preprint arXiv:1409.1556*, 2014.
- [18] A. Defazio, F. Bach, and S. Lacoste-Julien, “Saga: A fast incremental gradient method with support for non-strongly convex composite objectives,” in *Advances in neural information processing systems*, pp. 1646–1654, 2014.