Mutual Information Analysis of Social Media Images and Building Functions

Eike Jens Hoffmann[1], Martin Werner[2], Xiao Xiang Zhu[1,2]
[1]*Signal Processing in Earth Observation (SiPEO), Technical University of Munich (TUM)*
[2]*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 urban dynamics requires detailed insights into urban land use. On the most fine-grained level this classification is done on single building instance levels. This level of detail can hardly be solved using remote sensing only, but requires complementary data. Social media images are a promising additional image data source since they are captured on a global scale in vast volumes.**

In this study we investigate the relation between objects showing up in geotagged social media images and functions of buildings proximate to the image location. We propose a rasterization approach to embed features from images and labels from a target domain to calculate mutual information both domains share. In our study area of Los Angeles, USA, we show that using object detection is a valuable way of extracting features from social media images to predict building functions. Furthermore, we present the most significant object types for five types of buildings.

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

I. Introduction

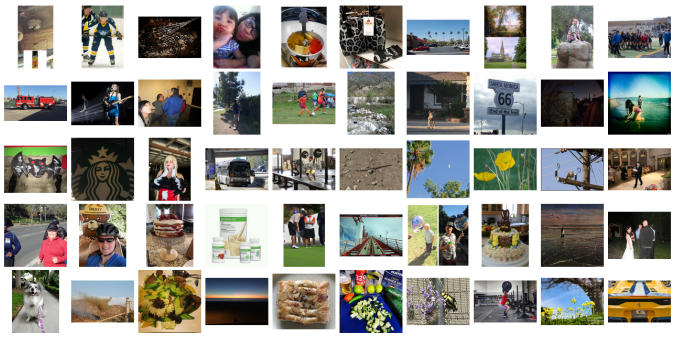

Figure 1: Examples of geo-tagged Flickr images

Remote sensing data is a comprehensive source of data in spatial as well as in temporal domain. As a constant stream of imagery acquired from space it allows land cover and land use classification as well as detecting changes over time. However, capturing land use in urban areas from aerial perspective is challenging due to ambiguities like distinguishing office buildings and apartments from above.

Resolving these ambiguities requires additional data sources like ground view data. This kind of imagery can be either obtained from proprietary map data providers like Google Street View or from public social media platforms like Flickr. Although commercial providers acquire their data in a structured manner from a specific perspective there might still be occlusions like buses and trees hiding the building itself. Additionally, the update frequency of commercial ground view data lies with the provider. And last but not least, capturing data with high quality requires a huge effort, which has to be justified.

In these cases, photos from different perspectives on social media platforms can help out. They provide publicly available photos taken by their users in different perspectives and showing various motifs (Figure 1). Tourists take images at landmarks and other touristic hotspots, whereas residents of areas take photos of their activities. Both kinds of images can be an additional source of information on the way towards building a semantic map. While the first gives insights about the things present at a location the latter shows what people usually do in these places.

In this work we investigate the relation between the number of different objects in an image and the building functions at the same location. First, we present a method to co-register image features and building functions as well as an approach to compare them. For our method, we apply a state-of-the-art object detection algorithm to a set of geo-tagged images from Flickr and rasterize the object class counts at different spatial resolutions. Second, we show our results in a study area covering Los Angeles metropolitan area.

II. Related Work

With the exponential growth of social media platforms there has also been an increased scientific interest in exploiting the data in these platforms in order to understand social, economic, and ecological patterns.

Beyond land cover and land use Lee et al. studied the value of Flickr images to predict 15 socio-economic binary labels as a multilabel binary classification problem [1]. By applying a CNN on their database of 40 million images they achieved accuracies 8% to 24% improvement compared to a random baseline on a global scale.

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), and the Bavarian Academy of Sciences and Humanities in the framework of Junges Kolleg.

Antoniou et al. investigated the feasibility of geotagged photos for land cover classification by manual labeling of images regarding their usefulness for this classification task [2]. Labelling was done on a dataset of images from London metropolitan area, UK, and revealed that 40% of all Flickr images were considered unusable for land cover classification. Interestingly, the most useful Flickr images are indoor photos revealing activities happening inside of buildings.

One of the most fine-grained hierarchical classification scheme with five top classes, 16 middle classes, and 45 fine-grained classes was proposed by Zhu et al. [3]. To deal with the noisiness in their Flickr dataset they applied search-based training set augmentation and online adaptive training. In combination with their two-stream CNN, one for recognizing objects and one for recognizing scenes, they were able to achieve over 29% recall at land parcel level on 45 classes.

Fang et al. presented a hierarchical parcel scheme derived from the OpenStreetMap road network and extrapolate the land use shown in images to street level blocks [4]. In their study area of London, they used geotagged images from the Geograph Britain and Ireland project and predicted a five-class land use schema based on a specialized Object Bank [8].

Our work focuses on using high-level, human understandable features for building function classification. We investigate the relation between the number of different objects in an image and the building functions around the places, where the image was taken. This study gives an outlook to the maximum performance that classifiers could reach when trained on this data. In addition, this approach adds to the explainability of the results as the basis for the decision remains a human-understandable statistic of the number of things seen in an image.

III. Methodology

We first show how we obtain data for the feature as well as for the target domain before we present a way to embed both modalities in a common raster. Finally, we explain how to relate feature and target domain using mutual information.

A. Feature Extraction from Social Media Images

Let $I = \{(i_1, l_1), \ldots, (i_n, l_n)\}$ be a set of geotagged images with i_x being an image and l_x its associated location. We apply an image detection algorithm on each i_x and yield counts $c_x \in \mathbb{N}^{80}$. We use the SSD object detection algorithm [5] trained on the COCO dataset [6] with a ResNet50 architecture [7] for feature extraction, which is able to assign labels for 80 object classes. This combination has been shown to have a good tradeoff between accuracy and speed (mAP 35 and 76 ms for inference on single image) [8]. Therefore, feature c_x has 80 dimensions with each dimension counting the number of detected objects in one image. In our case, we append c_x to each image (i_x, l_x, c_x).

B. Building Functions from OpenStreetMap

To obtain clear building functions we use the official labelling scheme from OpenStreetMap (OSM) for the key "building" according to their wiki[1]. All buildings that have a label other than this are neglected.

Next, all buildings having one of these tags are mapped to their cluster class according to Table I. In summary we have five cluster classes "accommodation", "commercial", "religious", "civic", and "other". Thus, for each building there is a building tag, a cluster class and its geometry as a multipolygon.

Cluster	OSM building tag
accommodation	apartments, farm, hotel, house, detached, residential, dormitory, terrace, houseboat, bungalow, static_caravan, cabin
commercial	commercial, office, industrial, retail, warehouse, kiosk
Religious	religious, cathedral, chapel, church, mosque, temple, synagogue, shrine
civic	bakehouse, kindergarten, civic, hospital, school, stadium, train_station, transportation, university, grandstand, public
other	barn, bridge, bunker, carport, conservatory, construction, cowshed, digester, farm_auxiliary, garage, garages, garbage_shed, greenhouse, hangar, hut, pavilion, parking, riding_hall, roof, shed, sports_hall, stable, sty, transformer_tower, service, ruins, water_tower

Table I: Mapping of OpenStreetMap tags to target classes

C. Rasterization

Next, we rasterize geometries g by summing up the counts for each class. For aerial images, we sum up feature counts c for all images taken inside the area of a grid cell. In case of buildings, we count the number of building functions in each grid cell if a multipolygon overlays it. More formally: for rasterizing the object counts c for each grid cell $g = (c_g, e_g)$ in the raster with e_g as the spatial extent of the grid cell and c_g the band values of the grid cell.

$$c_g = \sum c_i \forall (c_i, l_i) \text{ s.t. } l_i \text{ in } e_g$$

By adding up all c for each raster grid cell at l, we obtain a raster brick with n bands, one band for each dimension of the feature space c.

D. Mutual Information

After rasterization, we evaluate the relation between feature domain, i.e. object counts in images, and target domain, i.e. building functions. This evaluation is based on mutual information between each band in feature and in target domain, F and T, respectively.

$$I(F,T) = \sum_{f \in F} \sum_{t \in T} p(f,t) \log \frac{p(f,t)}{p(f)p(t)}$$

The mutual information is a metric measuring the dependence of two random variables [9]. While correlation measures the linear dependence of two random variables, mutual information calculates how much information in the

[1] https://wiki.openstreetmap.org/wiki/Key:building

joint probability distribution is common in both marginal distributions.

IV. Results

First, we give a brief introduction to the data captured in Los Angeles metropolitan area. Then, we show the results of our methodology starting with detection of objects in images and spatial statistics of the rasterization process. Finally, we present the mutual information correlation at different resolutions as well as a selected example of mutual information.

A. Study Area and Dataset

We evaluate our method in the Los Angeles metropolitan area using geo-tagged images from the social platform Flickr. In this area of 8,270 km^2 we collected 343,714 images using the Flickr API. Figure 2 shows the spatial distribution of the image dataset in our study area. At the beach, in the city center, and in Disney World, hotspots are visible, whereas the residential areas are sparsely covered.

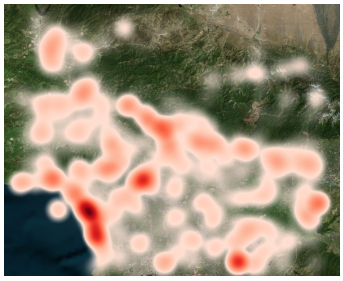

Figure 2: Spatial density of Flickr images in Los Angeles metropolitan area, background image ©Bing Aerial Maps

We get our building functions from OSM building labels. By applying the aforementioned scheme, we have in summary 2,213,834 labeled buildings. Table II shows that the majority of the buildings, namely 93.8%, are fall in the accommodation class, with commercial use as the second most frequent usage.

Class	Number of Buildings
Accommodation	2,077,275
Civic	8,733
Commercial	122,099
Other	5,631
Religious	78

Table II: Number of buildings for each target class

B. Detecting Objects in Social Media Images

In summary, the SSD algorithm detected 2,215,680 objects in 109,741 images with 1,083,000 of these detections being persons. The second most common class is car with 214,842 instances. Note that we do not set an uncertainty threshold for the object detection algorithm, hence there can be false or misclassified objects in our dataset. This is likely the reason why 414 zebras and 518 giraffes are in the dataset. Still, we believe that these noisy observations are helpful and that they do not have a very negative effect on the classification as we are usually aggregating many detections from multiple instances for a decision.

C. Rasterization

To predict land use classes from the object counts, we summarize the object counts for each raster grid cell and each class. Since the spatial density of social media images is highly volatile, we raster the counts in ten different spatial resolutions, starting with 10 m to 100 m ground sampling distance. Table III shows the number of grid cells that have a least one object count in any class compared to the number of all grid cells. At 10 m resolution only 0.05% of all grid cells have at least one detected object. Even at 100 m resolution only 2.3% of all grid cells are occupied with detections.

Extracting land use classes from OSM building tags results in similar spatial densities. At 10 m resolution 2.7% of all grid cells have a clear OSM land use class and at 100 m resolution the fraction increases to 13.1%.

Spatial Resolution [m]	#Grid cells	#Filled Feature Cells	#Filled Target Cells
10	131,077,000	77,599	3,575,067
20	32,769,250	65,542	2,022,459
30	14,569,296	58,029	1,256,380
40	8,194,050	52,717	841,979
50	5,243,080	48,583	582,346
60	3,642,324	45,144	424,090
70	2,677,128	42,405	323,941
80	2,049,102	40,122	256,384
90	1,619,160	38,020	208,299
100	1,310,770	36,037	172,996

Table III: Number of filled grid cells after rasterization of features, i.e. object detections, and targets, i.e. OSM building functions

D. Mutual Information

We investigate the relationship between object counts and building functions in order to have an estimate on how hard this problem will be for a machine learning algorithm. The better the target classes can be separated by different features values, the easier a classifier could fit the training data. Thus, we want low correlation between the feature vectors, i.e. the mutual information between different OSM classes should be less than one. After calculating the mutual information between all features and target classes, we calculate the correlation of mutual information between target classes. Figure 3 shows the distribution of correlations between OSM classes at different spatial resolutions. At 30 m and 100 m spatial resolution the median correlation is 0.485, which is the lowest value. However, the standard deviation at 30 m is lower than at 100 m, 0.12 vs. 0.16, respectively. From this point of view 30 m resolution might be a good candidate for training a classifier. This is true as well from an application point of view: 30 m resolution covers most buildings and still has room for including its outdoor environments and spatial context, where most of the images are taken.

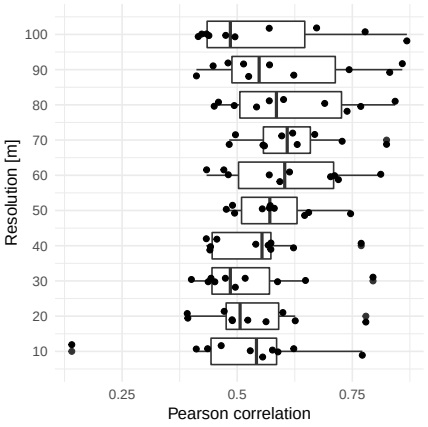

Figure 3: Correlation between target classes in terms of mutual information vectors at spatial resolutions from 10 m to 100 m

Taking a detailed look at the mutual information at 30 m resolution reveals that on the one hand the top 3 labels with the highest mutual information are dining table, truck, and cup, all together with commercial (Figure 4). On the other hand, the top 3 with the lowest mutual information are toaster, hair drier, and zebra in combination with religious. Table IV summarizes the top 5 labels with the highest mutual information between OSM classes. Although all classes share some labels, there is a clear difference between them.

accommodation	civic	commercial	other	religious
couch	person	dining table	train	potted plant
potted plant	book	truck	truck	chair
bed	handbag	cup	bus	person
bottle	chair	car	handbag	book
book	tv	bowl	bench	handbag

Table IV: Top 5 object classes by mutual information for each target class

V. Conclusion and Outlook

In this paper we investigated the relation between objects in social media images at locations and the functions of buildings nearby. By applying a state-of-the-art object detection algorithm, we extracted the frequency of 80 objects from Flickr images and summarized them using rasterization. To relate this with building functions, we gathered OpenStreetMap buildings that have a label according to the OSM labelling scheme and rasterized their respective building polygon footprints as well. We calculated the mutual information between object frequencies and buildings functions and found strong patterns indicating that a classifier could fit this problem with high accuracy. In the future we want to apply this to state-of-the-art classification algorithms to see if this holds true. Additionally, we could think of integrating further features like the relative size of the objects in the picture as well as the uncertainty the object detection classifier gives us.

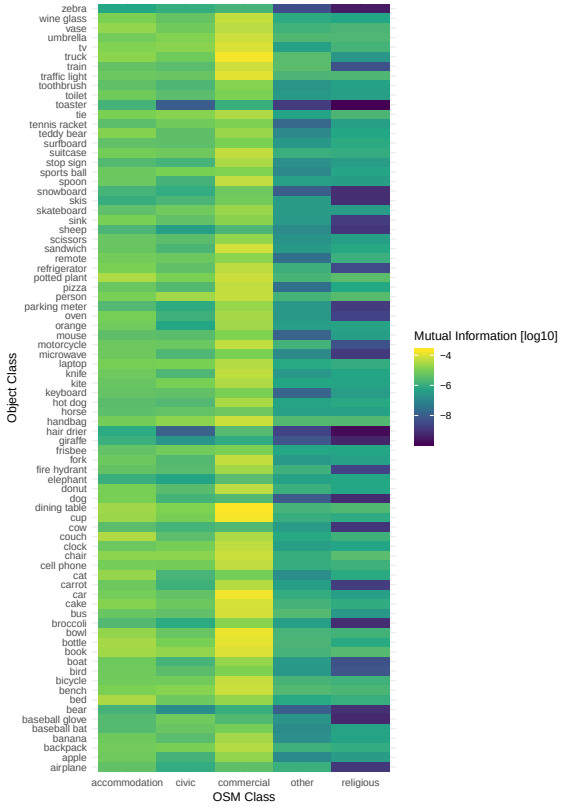

Figure 4: Mutual information between counts of detected objects and OpenStreetMap classes at log-scale and at 30 m resolution

References

[1] S. Lee, H. Zhang, and D. J. Crandall, "Predicting Geo-informative Attributes in Large-Scale Image Collections Using Convolutional Neural Networks," in *2015 IEEE Winter Conference on Applications of Computer Vision*, pp. 550–557, IEEE, jan 2015.

[2] V. Antoniou, C. Fonte, L. See, J. Estima, J. Arsanjani, F. Lupia, M. Minghini, G. Foody, S. Fritz, V. Antoniou, C. C. Fonte, L. See, J. Estima, J. J. Arsanjani, F. Lupia, M. Minghini, G. Foody, and S. Fritz, "Investigating the Feasibility of Geo-Tagged Photographs as Sources of Land Cover Input Data," *ISPRS International Journal of Geo-Information*, vol. 5, p. 64, may 2016.

[3] Y. Zhu, X. Deng, and S. Newsam, "Fine-grained land use classification at the city scale using ground-level images," *arXiv preprint arXiv:1802.02668*, 2018.

[4] F. Fang, X. Yuan, L. Wang, Y. Liu, and Z. Luo, "Urban Land-Use Classification From Photographs," *IEEE Geoscience and Remote Sensing Letters*, vol. 15, pp. 1927–1931, dec 2018.

[5] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *European conference on computer vision*, pp. 21–37, Springer, 2016.

[6] T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, and P. Dollár, "Microsoft COCO: Common Objects in Context," may 2014.

[7] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, IEEE, jun 2016.

[8] J. Huang, V. Rathod, C. Sun, M. Zhu, A. Korattikara, A. Fathi, I. Fischer, Z. Wojna, Y. Song, S. Guadarrama, and Others, "Speed/accuracy trade-offs for modern convolutional object detectors," in *IEEE CVPR*, vol. 4, 2017.

[9] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York, NY, USA: Wiley-Interscience, 1991.