Building Facade-based City Classification from Aerial View Images

Liu, Shikun Liu, Yanxi

May 26, 2016

Abstract

Building facades are key visual elements in urban scene analysis. State of the art building-facade extraction algorithms are challenged by various uncertainties and noise in the image data, such issues are particularly severe from aerial-view images of a city. We focus on using a machine learning algorithm to screen automatically detected facades. Furthermore, we use building facades as basic elements to determine which city this facade comes from. Our experimental validations on more than 4,000 building-facades provide promising results on both facade (non-facade) classification and city (NYC, Rome, SF) classification.

Keywords: City classification, Facades extraction, Urban scene analysis, Bag-of-Visual words, Lattice Network

Contents

1 Introduction **3**

2 Our Approach **3**

2.1 Facade Extraction 4

2.2 Feature Computation 5

2.3 Facade Classification 6

2.4 City Classification 7

3 Experimental Results **7**

3.1 Facade Classification 7

3.2 City Classification 9

3.3 Discussion 11

4 Conclusion **11**

References **12**

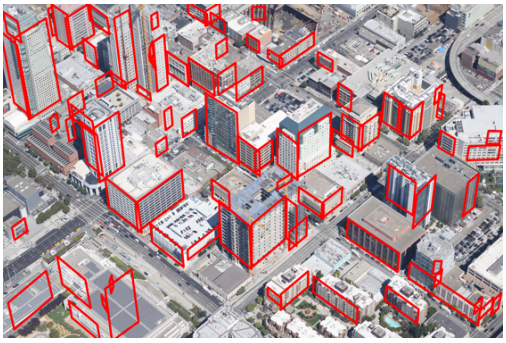
1 Introduction

Building facades are key visual elements in urban scene analysis. Previous works have used the building facades for building style classification [1, 2, 3], orientation estimation [5], geo-tagging [6] and 3D urban scene reconstruction[7, 8, 9]. Most algorithms for automatic facade extraction from urban scenes are based on finding regularities or symmetries in the image [5, 10, 11], which is either dominated by the building facade and/or is relatively clean with a non-oblique view. Robust algorithms for facade detection from high resolution, city-scale aerial view images are rare. In 2014, Liu and Liu [12] developed one of the state of the art building facade extraction algorithms that use local edge-based regularity measured by GINI-index scores [13] and a greedy adaptive region expansion procedure to search for building facades from city-scale aerial view images (NYC, SF, Rome). They demonstrated that from a single aerial view, 200-300 facades can be extracted. Though most captured facades in [12] are true building facades, there is also a considerate amount of false positives (Figure 1).

To improve the quality and the throughput of facade extraction from high resolution aerial images, one of our goals is to use machine learning method to screen the output of [12] into facades and non-facades. Second, we explore a novel topic of city-classification from building facades, in hope of finding the most representative facades from each city automatically. Different from previous works on photo-geolocation with restricted conditions, for example only when landmark buildings [14] or street-view images [15] are available, we focus on building facades extracted from aerial-view images. The latest related work from Google's PlaNet project [16] achieved a super-human level of accuracy combined with a long short-term memory (LSTM) architecture, and it trained using 91 millions of geo-tagged images, while our work only utilizes less than 5,000 labeled facades from three cities (NYC, SF, Rome).

2 Our Approach

We start with a massive extraction of building facade extraction using the method described in [12] from high-resolution aerial images of New York City, San Francisco and Rome. Due to viewing angles (some straightly top down views) and regions with no buildings, some of these images contain very few facades which are excluded from our experiment. Given such an initial dataset, we take a 2-step approach. First, build a binary classifier for facade/non-facade distinction; and second, to construct a 3-city classification using facade images only, both of these are utilizing supervised learning methods.



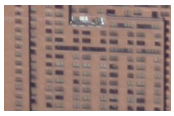
(a) San Francisco

(b) Rome

Figure 1: Using Liu and Liu's algorithm [12] on regularity-driven facade detection, we show the automatically detected building-facades from aerial views (a partial view) of San Francisco and Rome respectively. Though most captured facades (bounded by a red rectangle) are correct, one can observe several false positives, i.e. the region captured by a red rectangle is not a true building-facade. One of our tasks is to train a classifier to screen the output of [12] into facades and non-facades.

2.1 Facade Extraction

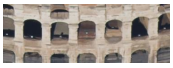
We choose 20 aerial images (each image is 3744 × 5616 in size) per city (NYC, SF, Rome) and run the algorithm from [12] on them to obtain roughly 16,000 automatically extracted potential building-facades without any post-processing. These facades are then hand-labeled into facades and non-facades. The labeling process turns out to be non-trivial since some of these extracted facades are ambiguous. See Figure 2 for sample facades and non-facades extracted automatically and their corresponding class labels.



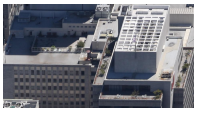
(a) True Facades Sample 1

(b) True Facades Sample 2

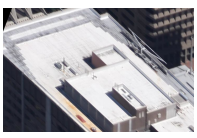
(c) True Facades Sample 3



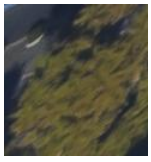
(d) True Facades Sample 4



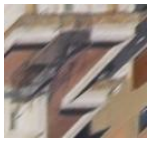
(e) False Facades Sample 1



(f) False Facades Sample 2

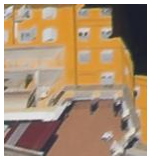


(g) False Facades Sample 3

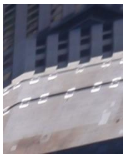


(h) False Facades Sample 4

(i) Ambiguous Facades Sample 1



(j) Ambiguous Facades Sample 2



(k) Ambiguous Facades Sample 3

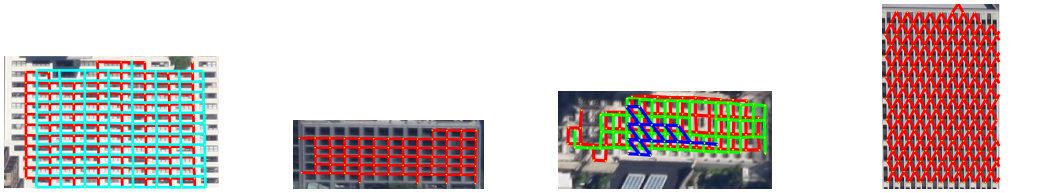
(l) Ambiguous Facades Sample 4

Figure 2: Top row: true facades extracted from NYC, SF, and Rome. Middle row: non-facades extracted. Bottom row: ambiguous cases which could be labeled either as facade or non-facade.

2.2 Feature Computation

For discriminating facade from non-facade (Figure 2), we aim to extract a set of features that may capture the regularity of the facade while robust to image noise. Since building facades are in the general category of near-regular textures [17, 18] and their underlying lattice is shown to be an effect descriptor for regularity [5, 6], we apply a lattice extraction algorithm from [19] on each facade. Figure 3 shows examples of detected lattices on true building facades versus non-facades. Once a lattice is detected from a facade, it divides the facade into multiple, relatively equal-sized tiles. Given the fact that randomness and regularity co-exist in a near-regular texture like building facade, we define and compute the following features for each facade:

1. **Number of tiles**: the maximum number of tiles in all the detected lattice on one extracted facade.
2. **A-score**: this is a measure of the appearance similarity of tiles of an extracted lattice as proposed in [17] and defined as
$$\text{A-score} = \frac{\sum_i std(T_i(1), T_i(2), \cdots, T_i(n))}{m}$$
where m is the total number of pixels in each aligned texel (tile) T_i.
3. **Occupancy:** the area of the extracted lattice with max number of tiles divided by the rectified facade area (image patch).
4. **Entropy:** $\sum -p \log_2 p$ where p contains the histogram of counts for each gray-scale value in the smoothed image using a $[8 \times 8]$ Gaussian kernel.
5. **Facade area**: the area of red bounding box on the input image from the output of.
6. **Facade probability:** a Gini-index [13] used to measure the probability of regularities in the local region of interest.

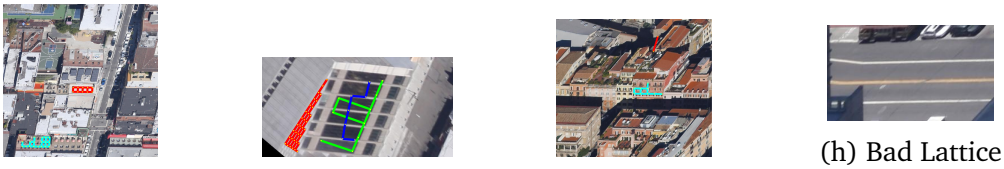


(a) Good Lattice Sample 1

(b) Good Lattice Sample 2

(c) Good Lattice Sample 3

(d) Good Lattice Sample 4



(e) Bad Lattice Sample 1

(f) Bad Lattice Sample 2

(g) Bad Lattice Sample 3

(h) Bad Lattice Sample 4 (No lattice found)

Figure 3: Lattice Detection [5] output: Top row contains building facades properly captured by the detected lattice. Bottom row shows cases where either no lattice detected or a very small lattice is detected on non-facades or noisy-facade regions.

2.3 Facade Classification

For facade classification, we use a soft-margin support vector machine method by directly applying a sub-gradient descent algorithm [22] to the expression

$$f(\mathbf{w},\mathbf{b}) = \left[\frac{1}{n}\sum_{i=1}^{n}\max(0, 1 - y_i(\mathbf{w}\cdot x_i + \mathbf{b}))\right] + \lambda\|\mathbf{w}\|^2,$$

in which x_i is our input data with p dimension, y_i represents its binary class, and $\mathbf{w}$ is the normal vector to the hyper-plane. Note that small value of λ can yields a hard margin classifier. Since f has proved to be convex, [23] the sub-gradient algorithm will give an efficient solution without depending on its number of data points. We define an augmented variance ratio (AVR) as

$$\mathrm{AVR}(F) = \frac{\mathrm{Var}(S_F)}{\frac{1}{C}\sum_{i=1\cdots C}\frac{\mathrm{Var}_i(S_F)}{\min_{i\neq j}(|\mathrm{mean}_i(S_F) - \mathrm{mean}_j(S_F)|)}}$$

where $\mathrm{Var}(S_F)$ is the cross-class variance of feature F, $\mathrm{Var}_i(S_F)$ and $\mathrm{mean}_i(S_F)$ are the within-class variance and mean of feature F for class i out of C distinct classes. Similar to Fisher criteria, [20] AVR is the ratio of cross-class variance of the feature over within-class variance, with an added penalty to features that have close inter-class means. AVR ranked features provide us with a quantitative basis to screen out non-discriminative features as a form of feature subset selection [21].

Given the six features defined in Section 2.2, we use AVR values as a measure for the discriminativeness of each feature, to understand those features that play an important role in facade classification.

2.4 City Classification

As a novel exploration using building facades extracted algorithmically from aerial images, we propose to seek whether building facades alone from different cities are distinguishable. An affirmative answer to this question may lead to the discovery of unique characteristics from each city's architectural structures reflected by their building facades. For this application, we propose to use a bag-of-features model [25] since all facades have a shared regularity and the difference may reside in lower-level features. We choose the top 80% strongest SURF features in each city [26] using the grid selection method. Since different city has different shooting angel for each image, SURF detector may provide scale invariant descriptors. We then use K-Means clustering [27] to create a 500 visual vocabulary for each class, and train a SVM classifier using error-correcting output codes (ECOC) framework [28]. The classifier uses the bag-of-features to encode the training images and use nearest neighbor method to construct the feature histogram of each image.

3 Experimental Results

3.1 Facade Classification

Since the automatically extracted facades [12] vary in shape and size, we would like to assess the relationship between the detected facade size and the discriminative power of the proposed features. We rank the facades data by the area feature. We pick the top N true facades and top N false facades to make sure they have same number of input data for each binary classification. We carry out 10 random splits of training and testing data into 8/2 ratio. We report precision and recall rate for the mean value of overall (Table 1). We denote our features by: *P:Probability, A:Area, E:Entropy, S:A-Score, T:Tiles, O:Occupancy*. AVR score and the mean AVR value for each feature dimension are shown in Table 2 and 3.

	$N = 100$	$N = 200$	$N = 500$	$N = 1000$	$N = 2000$	$N = 4000$
Precision	0.98	0.95	0.91	0.87	0.84	0.77
Recall	0.94	0.91	0.86	0.80	0.72	0.68
Precision	0.98	0.95	0.91	0.87	0.84	0.77
Recall	0.93	0.91	0.85	0.80	0.71	0.67

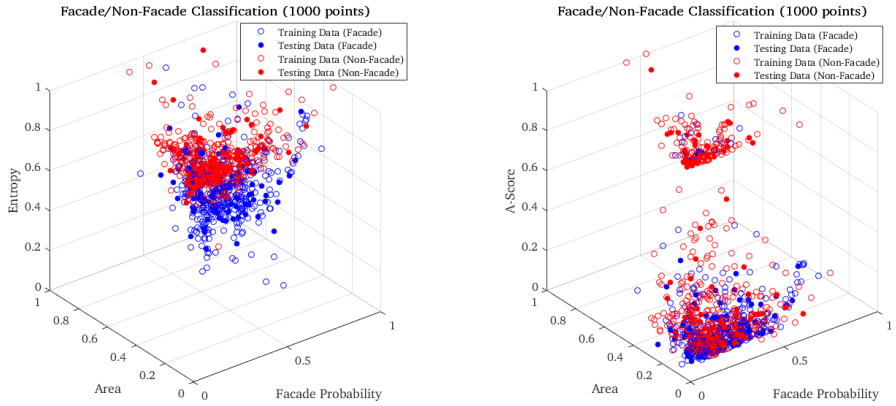
Table 1: Facade Classification for $2N$ size of true and false facades data points. Top: training. Bottom:testing.

	$N = 100$	$N = 200$	$N = 500$	$N = 1000$	$N = 2000$	$N = 4000$
	E:0.8493	E:0.7166	E:0.3791	S:0.3722	S:0.3557	S:0.2367
Highest	S:0.3339	S:0.4103	S:0.3782	E:0.3017	E:0.2184	E:0.1690
↓	P:0.1649	P:0.1306	T:0.1178	T:0.0851	T:0.0594	T:0.0297
Lowest	T:0.1508	T:0.1150	P:0.0525	A:0.0164	A:0.0120	P:0.0236
	O:0.0676	O:0.0287	O:0.0338	P:0.0154	P:0.0090	A:0.0045
	A:0.0233	A:0.0175	A:0.0215	O:0.0088	O:0.0003	O:0.0022

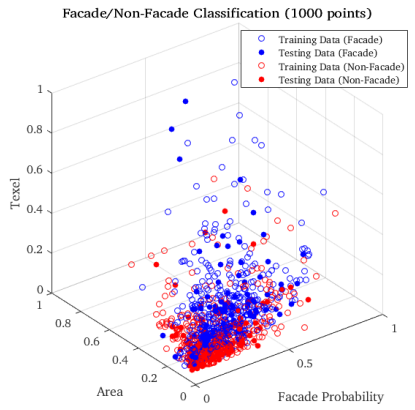
Table 2: AVR score for each feature when varying the input data size. Here N means that the data set contains N facades and N non-facades, thus the total number of the data points is $2N$.

	Probability	Area	Entropy	A-Score	Tiles	Occupancy
Mean	0.0660	0.0159	0.4390	0.3478	0.0930	0.0236

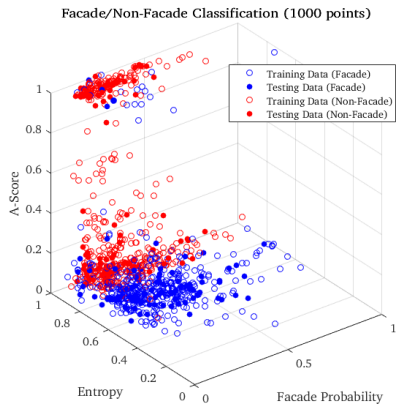
Table 3: Mean AVR score for each feature dimension. The ranking by the mean AVR (from most discriminative to least): Entropy → A-Score → Tiles → Probability → Occupancy → Area. We note that the area feature usually works well when combined with other features.



(a) Entropy-Area-Facade Probability (b) A-Score-Area- Facade Probability



(c) Texel-Area-Facade Probability



(d) A-Score-Entropy-Facade Probability

Figure 4: Data distribution visualization in the most discriminative feature subspace for facade/non-facade classification. Blue: facade points, red: non-facade points, solid: test data, empty: training data. One can observe clearly that the entropy feature is one of the most discriminative features.

3.2 City Classification

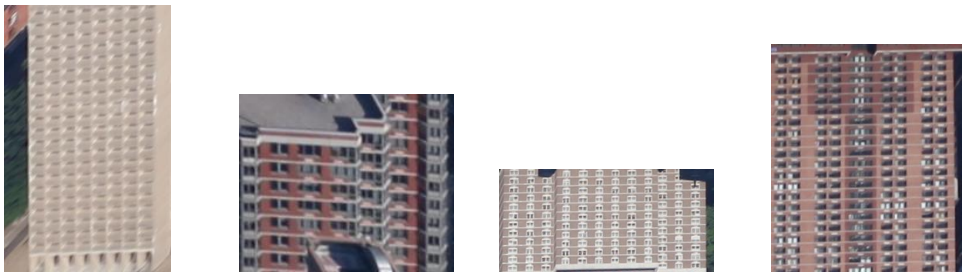
Using MATLAB computer vision toolbox, we extract SURF features in grid $[8 \times 8]$ grid step, and the minimum number features in three cities, and the strongest 80 SURF features to create a 500-bag-of-words model. Since facades quality and SURF feature quality largely depend on facade area, we use the area (A) as a screening measure for the data set used. To evaluate the relation between facade patch area and classification rate, we document the classification results with a 80/20% training/testing ratio in Table 4, where *A: Area, #F: number of input facades with each area threshold.*

Figure 5 below provides the most discriminative sample facades in each city (Results from bag-of-words).

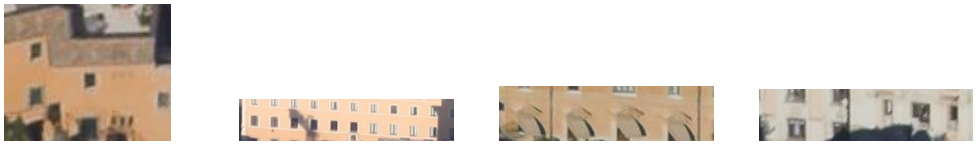
Our experiments are carried out on an Intel i7-4770 with 3.40 GHz, Ram:16Gb, with Windows 10 Pro. It spent approximately 5min for binary classification and about 300 min for all the city-classification training and testing steps.

NYC	A ≥ 1000 # F: 196	A ≥ 500 # F: 565	A≥ 200 # F: 1295	A ≥ 100 # F: 1845	A ≥ 50 # F: 2292	A ≥ 20 # F: 2506	A ≥10 # F: 2506	A ≥ 0 # F: 2506
Precision	0.87	0.87	0.91	0.92	0.92	0.90	0.87	0.91
Recall	0.94	0.93	0.83	0.80	0.77	0.74	0.73	0.75
Precision	0.82	0.82	0.78	0.89	0.87	0.89	0.91	0.85
Recall	0.92	0.90	0.97	0.75	0.73	0.67	0.76	0.71
ROME	A ≥ 1000 # F: 40	A ≥ 500 # F: 130	A ≥ 200 # F: 364	A ≥ 100 # F: 598	A ≥ 50 # F: 846	A ≥ 20 # F: 981	A ≥10 # F: 981	A ≥ 0 # F: 981
Precision	0.89	0.86	0.72	0.72	0.69	0.70	0.70	0.69
Recall	1.00	0.99	0.93	0.94	0.91	0.89	0.89	0.90
Precision	0.87	0.72	0.82	0.78	0.63	0.64	0.65	0.60
Recall	0.87	0.88	0.88	0.82	0.86	0.85	0.89	0.81
SF	A ≥ 1000 # F: 417	A ≥ 500 # F: 615	A ≥ 200 # F: 817	A ≥ 100 # F: 921	A ≥ 50 # F: 997	A ≥ 20 # F: 1047	A ≥10 # F: 1047	A ≥ 0 # F: 1047
Precision	0.97	0.96	0.88	0.82	0.80	0.74	0.75	0.76
Recall	0.86	0.86	0.88	0.86	0.87	0.86	0.86	0.87
Precision	0.97	0.96	0.88	0.87	0.74	0.67	0.70	0.73
Recall	0.92	0.82	0.87	0.82	0.78	0.86	0.80	0.78

Table 4: The precision and recall rates from training (top)/testing(bottom) of 3-city classification using facades from NYC, Rome and SF respectively.



(a) NYC Facade 1 (b) NYC Facade 2 (c) NYC Facade 3 (d) NYC Facade 4



(e) Rome Facade 1 (f) Rome Facade 2 (g) Rome Facade 3 (h) Rome Facade 4

(i) SF Facade 1 (j) SF Facade 2 (k) SF Facade 3 (l) SF Facade 4

Figure 5: Some of the most distinguishable facades from NYC (top), Rome (middle), SF (bottom) respectively.

3.3 Discussion

Our facade classification and facade-based city classification demonstrate promising results using automatically extracted facades of varying shape/size from high resolution aerial images. We have learned that indeed features capturing the near-regularity of facades play an important role in facade classification (high AVR values). Entropy features computed from building facades prove to be the most discriminative features of all, establishing a direct link between entropy and regularity. It is somewhat surprising to achieve from 70+% to 90+% precision and recall rates on city classification using building facades alone, without any special attention paid to landmark buildings. The rates however are better with the increasing of the size of the facades, which contain more information and more representative of the location. Further study is required to analyze the lower level features learned from the bag-of-words and verify their discriminativeness on more cities.

4 Conclusion

We have made a unique contribution in this paper to investigate the validation and the use of building facades from city-scale building facades. We have improved previous work [12] by increasing the true positives of the detected facades, while removing false positives. We have further contributed to the geo-tagging literature by using building facades for city classification with surprisingly positive results (Table 4). Our results show that indeed building facades are powerful and descriptive elements for urban scene analysis. We plan to look in depth the features learned through supervised learning on city classification and expand our experiments on more cities worldwide.

References

[1] Shalunts, G., Haxhimusa, Y., Sablatnig, R.: Architectural style classification of building facade windows. In: Advances in Visual Computing. Springer (2011) 280-289

[2] Mathias, M., Martinovic, A., Weissenberg, J., Haegler, S., Van Gool, L.: Automatic architectural style recognition. ISPRS-International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences 3816 (2011) 171-176

[3] Martinovic, A., Mathias, M., Weissenberg, J., Van Gool, L.: A three-layered approach to facade parsing. In: Computer Vision-ECCV 2012. Springer (2012) 416-429

[4] Martinovic, A., Mathias, M., Weissenberg, J., Gool, L.: A three-layered approach to facade parsing. In: Proc. ECCV. (2012)

[5] Park, M., Brocklehurst, K., Collins, R., Liu, Y.: Translation-symmetry-based perceptual grouping with applications to urban scenes. In: Proc. ACCV. (2010) 1-14

[6] Schindler, G., Krishnamurthy, P., Lublinerman, R., Liu, Y., Dellaert, F.: Detecting and matching repeated patterns for automatic geo-tagging in urban environments. In: Proc. CVPR. (2008)

[7] Gallup, D., Frahm, J.M., Pollefeys, M.: Piecewise planar and non-planar stereo for urban scene reconstruction. In: Proc. CVPR. (2010) 1-8

[8] Mobahi, H., Zhou, Z., Yang, A., Ma, Y.: Holistic 3d reconstruction of urban structures from low-rank textures. In: Proc. ICCV workshop. (2011) 1-8

[9] Vanegas, C., Aliaga, D., Benes, B.: Building reconstruction using manhattan world grammars. In: Proc. CVPR. (2010) 1-8

[10] Ceylan, D., Mitra, N.J., Li, H., Weise, T., Pauly, M.: Factored facade acquisition using symmetric line arrangements. In: Proc. EUROGRAPHICS. (2012)

[11] Xiao, J., Fang, T., Tan, P., Zhao, P., Ofek, E., Quan, L.: Image-based facade modeling. In: Proc. ACM SIGGRAPH Asia. (2009)

[12] Liu, J., Liu, Y.: Local regularity-driven city-scale facade detection from aerial images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2014) 3778-3785

[13] Hurley, N., Rickard, S.: Comparing measures of sparsity. In: Proc. MLSP. (2008)

[14] Avrithis, Y., Kalantidis, Y., Tolias, G., Spyrou, E.: Retrieving landmark and non-landmark images from community photo collections. In: Proceedings of the 18th ACM international conference on Multimedia, ACM (2010) 153-162

[15] Chen, D.M., Baatz, G., Koser, K., Tsai, S.S., Vedantham, R., Pylya, T., Roimela, K., Chen, X., Bach, J., Pollefeys, M., et al.: City-scale landmark identification on mobile devices. In: Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on, IEEE (2011) 737-744

[16] Weyand, T., Kostrikov, I., Philbin, J.: Planet-photo geolocation with convolutional neural networks. arXiv preprint arXiv:1602.05314 (2016)

[17] Liu, Y., Lin, W.C., Hays, J.: Near-regular texture analysis and manipulation. In:ACM Transactions on Graphics (TOG). Volume 23., ACM (2004) 368-376

[18] Hays, J., Leordeanu, M., Efros, A.A., Liu, Y.: Discovering texture regularity as a higher-order correspondence problem. In: Computer Vision-ECCV 2006. Springer (2006) 522-535

[19] Park, M., Brocklehurst, K., Collins, R.T., Liu, Y.: Translation-symmetry-based perceptual grouping with applications to urban scenes. In: Computer Vision ACCV 2010. Springer (2010) 329-342

[20] Duda, R., Hart, P., Stork, D.: Pattern Classification. John Wiley & Sons, New York (2001)

[21] Bishop, C.M.: Neural Networks for Pattern Recognition. Clarendon Press (1995) ISBN:0198538499.

[22] Shalev-Shwartz, S., Singer, Y., Srebro, N., Cotter, A.: Pegasos: Primal estimated subgradient solver for svm. Mathematical programming 127 (2011) 3-30

[23] Suykens, J.A., Vandewalle, J.: Least squares support vector machine classifiers. Neural processing letters 9 (1999) 293-300

[24] Breiman, L., Friedman, J., Stone, C.J., Olshen, R.A.: Classification and regression trees. CRC press (1984)

[25] Nowak, E., Jurie, F., Triggs, B.: Sampling strategies for bag-of-features image classification. In: Computer Vision-ECCV 2006. Springer (2006) 490-503

[26] .Bay, H., Ess, A., Tuytelaars, T., Van Gool, L.: Speeded-up robust features (surf). Computer vision and image understanding 110 (2008) 346-359

[27] Tang, R., Fong, S., Yang, X.S., Deb, S.: Integrating nature-inspired optimization algorithms to k-means clustering. In: Digital Information Management (ICDIM), 2012 Seventh International Conference on, IEEE (2012) 116-123

[28] Frome, A.L.: Learning distance functions for exemplar-based object recognition. ProQuest (2007)