Beyond the Google StreetView: learning SVMs for urban analysis

Anonymous ECCV submission

Paper ID 13

Abstract. Given a large database of geotagged imagery of a whole city the goal is to evaluate a distribution of different architecture styles across the city, to detect areas with open or close views, areas with dense or loose vegetation or places with graffiti on the walls. We first download 180,000 panoramas of the city of Madrid from the Google street view, then we generate a random set of 7,000 perspective images and label them. We use the labeled images to train a set of linear SVM predictors for each class. Finally, we uniformly sample 120,000 random images across the city of Madrid covering roughly area of $30\times20km$ and generate a set of heatmaps showing response of the learned predictors for each of the classes. The contribution of the paper is three fold: (i) We propose a simple method for detection of architecture style, graffiti, vegetation and type of view. (ii) We have created a labeled set of images that is publicly available. (iii) We visualize the results by set of heatmpas and analyze the results with an urbanist expert. The analysis shows coherence between the results and historical context of the city and brings interesting insights to the reading of the territory in a quantified way.

Keywords: SVM, architecture style recognition, urbanism analysis

1 Introduction

The goal of this work is to learn, visualize and analyze a set of given urban attributes of the city. The attributes of interest are architecture style, amount of vegetation, amount of open view or presence of graffiti. Given a large database of geotagged imagery such as Google Street View the goal is to evaluate a distribution of different attributes across the city. Then we then visualize the results on maps and analyze it with an urbanist expert. The analysis shows coherence between the results and historical context of the city and brings interesting insights to the reading of the territory in a quantified way. This work shows that state of the art computer vision and machine learning techniques along with sufficient amount of data can serve as useful tool for urban analysis of the city.

2 Related work

Since Google released its first StreetView API in 2007 there were no doubts that researchers and computer vision community will aim on using the Street View

as powerful source of data. Some of the early works [1], [2] used Google street view data for place recognition, 3D reconstruction [3], [4] or for identification of commercial entities in street view imagery [5]. More recent works utilize support-vector-machines (SVM) to learn a set of patches that are specific for given city [6] or to learn place-specific classifiers for place recognition [7].

Our work is mostly related to two decent papers. The former one [8] explores the problem of facade recognition. This early work has several limitations that we are bridging in this paper. The dataset used is small and takes into account only three styles. Secondly, the approach for segmenting facades is cumbersome and finally, labels of adjacent facades are exploited in the learning model. We want to scale our analysis to hundred thousands of images, simplify the annotation protocol and generalize styles to process multiple cities in a country and across countries.

The latter is nice paper of Doersch et al. [6]. They aim on recognizing visual elements that are both frequent and discriminative for particular city. They collect a street-view imagery for several cities and generate millions of random patches, then they train linear SVMs in iterative manner while adding a weak geographic constraint and analyze learned wight vectors in order to find discriminative patches.

Unlike [6] our goal is not to learn discriminative patches. Instead we aim on learning our linear SVM on bag-of-words-like representation since we want to classify the whole appearance of the image on order to detect different architecture styles and visual attributes. Moreover, our goal is to learn a cheap and scalable pipeline that eventually allows to process millions of images.

3 Approach overview

We opt for a classic approach [9], [10] to learn discriminative classifiers separating images of buildings based on architectural style and other attributes such as level of vegetation, open or closed view. In this section we first describe the strategy we used to collect and annotate google street view images as well as the image analysis pipeline to efficiently process such images.

3.1 Database

Similar to other works [7], [6] we follow [3] and build our database from a collection of Google StreetView images. We have downloaded about 180,000 panoramic image of the city of Madrid. For a subset of panoramas we have generated four perspective images that capture both building facades and street level scene (details are given next in section 4). In the same manner we have generated a disjunctive set of images that are be subsequently annotated by humans.

In order to generate a coherent set of candidate facade labels, we took advantage of the help of an urbanist expert on spanish urbanism. This interaction resulted into the definition of four prevailing styles relevant for the city of Madrid and generalisable to the most part of spanish cities. We briefly summarize these four architectural styles in the tables 1 and 2.

	Monumental	Classical residential
History	not an age-based homogeneous style built before 1920	not an age-based homogeneous style built before 1920
Original use	religious or civil architecture	92.50
Current use	cult, monumental, museums, prestigious commercial activities (banks, etc.)	commercial premises on ground floor, and mainly apartments in the upper floors
Features	volume non just parallelepiped, towers, ornaments, columns, steep slope roofs, domes, etc.	volume is parallelepiped, eclectic and neoclassic ornaments, iron balconies, straight windows (high>width)

Table 1. Monumental vs. Classical residential. Definition of prevailing styles relevant for the city of Madrid

	Contemporary residential	Contemporary non-residential
History	built after 1920	built after 1920
Original use	residential	tertiary use, sport stadium, educational, offices, hotels, both private and public buildings
Current use	residential	same as original use
Features	brick or concrete faades, wide windows (width>=high), parallelepiped volumes, flat roofs	steel and glass architecture, concrete

Table 2. Contemporary residential vs. Contemporary non-residential. Definition of prevailing styles relevant for the city of Madrid

We have designed a simple web-based annotation interface in order to collect image labels for a random set of Google street view images described above. We have asked nine annotaters to label a given subset of images. Rather then classifying each of many Madrid architecture styles (Castilian baroque, Gothic, Romanesque, Neoclassical, Baroque, Bourbon rococo etc.) we asked annotaters to classify the architecture appearance in one of four classes: *Classical residential*, *Contemporary-residential* (modern buildings, second half of 20th century), *Non-residential* (office buildings, factories, shopping centers) and *Monumental*. They were given a reference set of samples for each class as show in figures 1 and 2.

In addition we have asked them to also label amount of vegetation appearing in the image (*dense/loose* vegetation), a type of view (*close/open/partially-open* view) and presence of *graffiti*. A number of collected labels for each category is summarized in table 3. It is worth noting that non of the annotaters was an urbanist or an expert in architecture. Hence, our database is expected to contain some amount of human-induced noise which makes our task even more interesting.

3.2 Learning SVMs

Each image j is represented by its feature vector x_j. In this work we represent images by a pyramid-of-bag-of-words (PHOW) features (details are given next in section 4). For each class k we train a linear SVM by minimizing objective

$$||w_k||^2 + C_1 \sum_{x_j \in \mathcal{P}_k} h\left(w_k^T x_j + b_k\right) + C_2 \sum_{x_j \in \mathcal{N}_k} h\left(-w_k^T x_j - b_k\right), \qquad (1)$$

where $\mathcal{P}_k$ and $\mathcal{N}_k$ are positive and negative training sets for class k, h is the squared hinge loss and C_1, C_2 are penalty weights. While positive set $\mathcal{P}_k$ consist of all labeled images in class k the negative set N_k contains rest of the images

Fig. 1. *(left)* An example of monumental architectural style manually selected for the benefit of the annotators, *(right)* classical residential buildings.

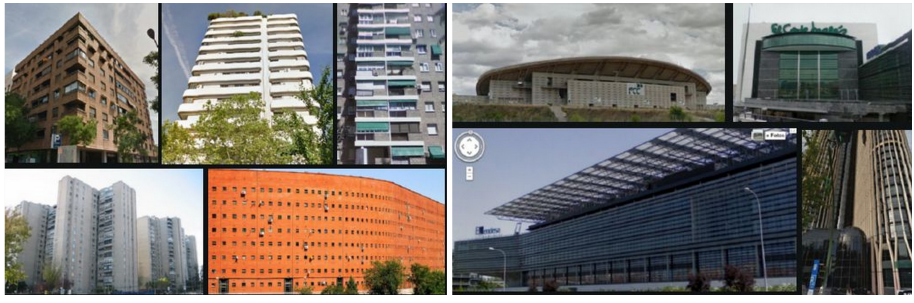


Fig. 2. *(left)* An example of contemporary residential architecture selected for the benefit of the annotators, *(right)* contemporary non-residential buildings.

within the same category. For instance, for class *Classical residential* form table 3 the negative set $\mathcal{N}_k$ contain images labeled as *Contemporary residential*, *Contemporary non-residential* and *Monumental*. Hence, we are training a set of one-versus-rest (OVR) predictors.

3.3 Evaluation

Rather then quantitative evaluation our goal is a qualitative analysis. Particularly, we aim on generating a set of heatmaps, one for each predictor, showing its response across the city on the map. These heatmaps can be interpreted as a spatial density of each class. What we expect to see is a semantics and complementarity of the maps. For instance, is the classical residential style located in the center, are there open views in the center, is moder style and office buildings located out of the city center, are there office buildings and at the same place as old castilian baroque houses, is there a lot of vegetation in the city center?

Indeed, we know the answers to the questions above because of a prior knowledge: The city center is typically the oldest part of the city, hence, we expect there is a majority of the classical residential style buildings. As the city expands a more modern buildings are being constructed at the peripheries as well as the

Category	# of labeled images	E[AP] [%]
Architecture style:		
Classical residential	2702	65.0
Contemporary residential	1431	69.0
Contemporary non-residential	1032	39.0
Monumental	236	28.2
Vegetation:		
Dense	1219	81.1
Loose	2567	94.9
View:		
Open view	911	71.5
Close view	3767	94.4
Partially open	2509	71.7
Other:		
Graffiti	605	50.6
Not-graffiti	2375	-

Table 3. A number of labeled images in each category. The right column shows an expected average precision (AP) for each class within given category. The expected AP have been estimated by 6-fold cross-validation during training. The values indicate which class has a potential to be detected by learned predictor.

office buildings and factories. There is probably not much open views in the center since it consists of lot of narrow streets and close spaces but in the suburbs we may expect a lot of open views since a density of buildings is significantly lower.

However, answering some questions may not be that straight forward. For instance, where is the majority of the graffiti in Madrid? Is it gonna be in the center, in suburbs, somewhere else, or is it uniformly distributed across the city? Beside of the class density heatmaps we are also interested in visual appearance of the top ranked images in each class and see whether or not these are coherent with the classes. In section 5 we present a set of heatmaps along with several image examples for each class and we briefly analyze the qualitative results and its semantics.

3.4 Visualization

The idea is to visualize a response of each class on a map. Since each image in the database has it's associated GPS location we aim on utilizing the Google StreetVoew API to plot a density of each class over a map of the city. To do so, for each image j in turn we compute its thresholded SVM score

$$s_{jk} = \max\left(0,\ w_k^T \cdot x_j + b_k - t_k\right) \tag{2}$$

where t_k is a threshold for given class k, and than we use this score as a voting weight in a accumulation space. Hence, the score is either zero or a positive

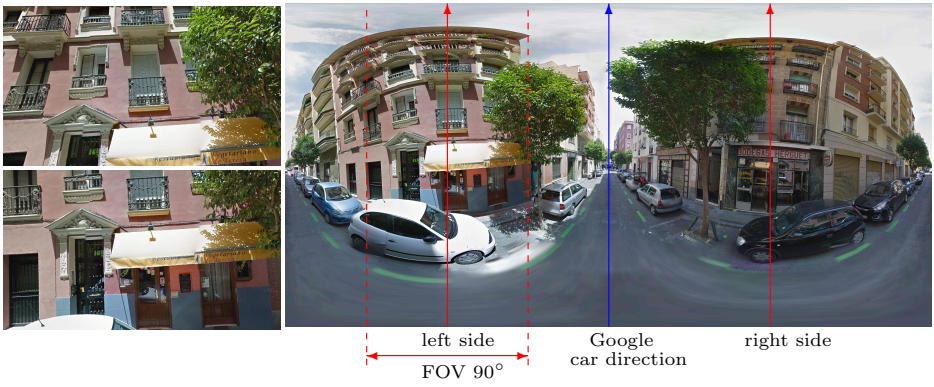


Fig. 3. A center of each panorama *(right)* corresponds to the Google car motion. For each side of the panorama we generate two perspective images *(left)* with horizontal field of view 90° in two different elevation pitches in order to capture both street view level and building facades.

distance from the threshold t_k. The accumulation space for plotting a heatmap is implemented in the Google Map Java Script API in `google.maps.visualization.HeatmapLayer` object. Details will be given later in the section 4.

4 Experimental details

Database: We have downloaded set of 180,000 panoramic images [3] roughly covering the area of $30km \times 20km$. We then split this area by a regular grid with spacing of $1km$, hence we have spit the area by 600 uniform cells. Then we have randomly selected a set of 30,000 panoramas in such a way that for each in turn we (i) have randomly picked a grid cell and (ii) from this cell we have randomly picked the panorama. In this manner we have achieved almost uniform sampling across the whole area.

Perspective images: For each panorama in turn we generate four perspective images. As shown in figure 3 the center of each panorama corresponds to the Google car direction of motion. Since our goal is to capture building facades we generate views to the right and left w.r.t. the car motion. In order to capture both street-level view and building facades, for each side we generate perspective images with two different elevation pitches, namely 4° and 28°. All the images have resolution of 960×768 pixels and horizontal field of view 90°. Finally, our database consists of 120,000 perspective images.

Features: We first extract dense SIFT features with a step of 16 pixels. Then we represent each image by a pyramid-of-histogram-of-bag-of-words (PHOW) [11] quantized into a $20k$ visual word dictionary. The dictionary has been learned by k-means from SIFT descriptors of $5,000$ randomly elected images. We have found

that PHOW representation performs better then BOW. This can be explained by the fact that rather then representing a place-specific features of the particular image we aim on representing it's global appearance.

Learning SVMs: For each category we learn one-versus-rest linear SVM using [12]. To select optimal parameters C_1, C_2 we perform a grid search with $6-fold$ cross-validation and for each coupe of parameters we compute a expected average precision (see table 3). Then we pick a set of parameters that maximizes the expected AP and re-train the SVM with all available data. Table 3 shows a maximal expected AP reached during cross-validation. The values can indicate how reliable the classifier will be. For instance, the $E[AP]$ of the class *Monumental* is only 28%, considering that *Architecture style* category has only four classes, for a random performance would be 25%.Thus a predictor of this class will be very noisy.

5 Results

Historical context

Madrid was far from being the biggest city of Spain when it was set as its capital city in 16th century, nor was it the head of a great archbishopric, so within its barely 128 hectares there were not great buildings comparable to those of other cities such as Toledo, Seville, or even Lisabon. The footprint of this small village that became the capital of Philippe II Empire is still visible on its messy mediaeval urban fabric in the city centre, known as Madrid of the Austrias. In the following two centuries the city grew in the same chaotic way, gradually covering the 700 hectares surface within the wall built by Philippe IV, with sober civil buildings, together with an increasing number of religious facilities and royal monumental buildings, leaving just the minimum necessary space for narrow and winding streets. This situation lasts until 1860, when, following the hygienists stream, a grid-shaped urban expansion plan was approved, adding 1,500 extra hectares to the city, however, its execution took over 50 years and hence it reached the 20th century.

Qualitative analysis

In this analysis *classical residential* style (Figure 5a) encloses those civil buildings built before 1920, indeed a broad and diverse period, but in which buildings rose up along 19th century in an eclectic style are the majority above those built before 1800 (currently these sums up only 1,349 buildings in a city with more than 122,000 buildings, 11,138 of them erected before 1920).

This expansion area (figure X or Y) has another characteristic that is shown by our analysis: in the new wide and straight streets there is place for trees, forming boulevards, with a reflection on *loose vegetation* label (Figure 7b), and a high density of commercial premises on ground floor. It is interesting to verify

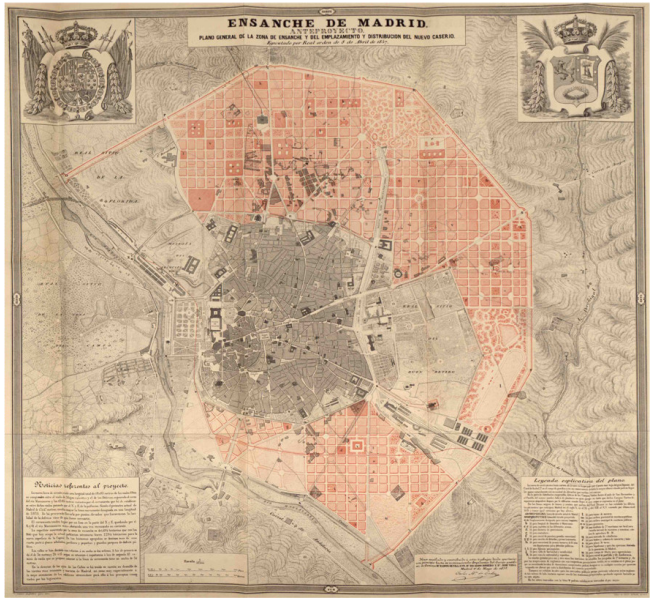


Fig. 4. A grid-shaped urban expansion plan that was approved in 1860. The city have grown according this plan till the first half of the 20th century.

how *classical residential* heatmap (Figure 5a) shows informal settlements out of the ordered area that afterwards were to be incorporated to the city such as Tetun and Vallecas.

After 1930's, when the city reached its first million inhabitants, rationalism architectural current derives into a *contemporary style* (Figure **??**b) of cheap and quick execution (concrete structures and brick faades). This factor allows a rapid growth that was steadily accelerated until 1970's, when the city reaches its current population (though up to our days new buildings have kept on filling the remaining empty land plots or replacing previous buildings). Nowadays the city is surrounded by heavy traffic highways (M30 and M40) built between 1970's and 1990's, and where many tertiary buildings are located (offices and commercial centres), this is shown in our data through the *contemporary non-residential* (steel and glass architecture), Figure 6b, and *open view* labels, Figure **??**a. The major transformation experienced by the city in the 2000's, the M30 south-west arch burying and the huge lineal green area so generated along Manzanares river, is not shown in our data (open view and vegetation labels) because the Google car gathered images inside the resulting underground tunnel.

6 Conclusion

In global terms, the project contributes with interesting insights to the reading of the territory in a quantified way structured upon an available and rich data

source. This innovative generation of urban metrics obtained thanks to a massive and fast method constitutes a broad and transversal approach that encloses multiple potential aims. To quote just some of them, this results open the possibility of correlating urban characteristics such as the architectonical predominant style, or the street average width with commercial or touristic success or decay of an area, if crossed with other data such as economic performance information, or the vegetation density with life quality attributes, such as liveability, healthiness or even other more subjective matters such as safety.

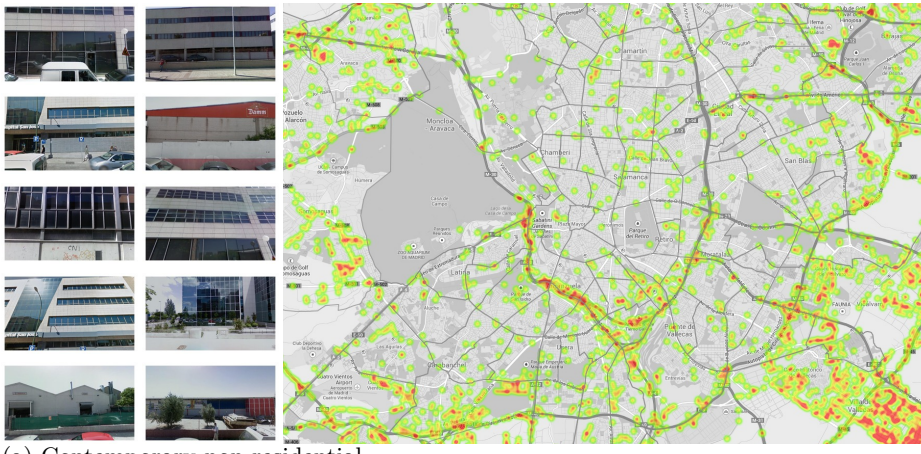
References

1. Knopp, J., Sivic, J., Pajdla, T.: Avoidng confusing features in place recognition. (2010)
2. Torii, A., Sivic, J., Pajdla, T.: Visual localization by linear combination of image descriptors. In: Proceedings of the 2nd IEEE Workshop on Mobile Vision, with ICCV 2011. (2011)
3. Gronat, P., Havlena, M., Sivic, J., Pajdla, T.: Building streetview datasets for place recognition and city reconstruction. In: Technical Report. (2011)
4. Havlena, M., Torii, A., Knopp, J., Pajdla, T.: Randomized structure from motion based on atomic 3D models from camera triplets. (2009) 2874–2881
5. Zamir, A., Darino, A., Shah, M.: Street view challenge: Identification of commercial entities in street view imagery. In: Machine Learning and Applications and Workshops (ICMLA), 2011 10th International Conference on. Volume 2., IEEE (2011) 380–383
6. Doersch, C., Singh, S., Gupta, A., Sivic, J., Efros, A.A.: What makes paris look like paris? **31**(4) (2012)
7. Gronat, P., Obozinski, G., Sivic, J., Pajdla, T.: Learning and calibrating per-location classifiers for visual place recognition. (2013)
8. 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 (2011)
9. Csurka, G., Bray, C., Dance, C., Fan, L.: Visual categorization with bags of keypoints. In: Workshop on Statistical Learning in Computer Vision, ECCV. (2004) 1–22
10. Sivic, J., Zisserman, A.: Video Google: A text retrieval approach to object matching in videos. In: Proceedings of the International Conference on Computer Vision. Volume 2. (October 2003) 1470–1477
11. Vedaldi, A., Fulkerson, B.: Vlfeat: An open and portable library of computer vision algorithms. http://www.vlfeat.org/ (2008)
12. Fan, R.E., Chang, K.W., Hsieh, C.J., Wang, X.R., Lin, C.J.: LIBLINEAR: A library for large linear classification. Journal of Machine Learning Research **9** (2008) 1871–1874

(a) Classical residential

(b) Contemporary residential

Fig. 5. Architecture style: Heatmaps *(right)* showing a density of different architecture styles across the city of Madrid. Notice that while *classical residential* style (a) is mostly concentrated in the city center the *contemporary residential* style (b) is detected away from the city center. On the *left* there are examples of several top-ranked images for given style.



(a) Contemporary non-residential

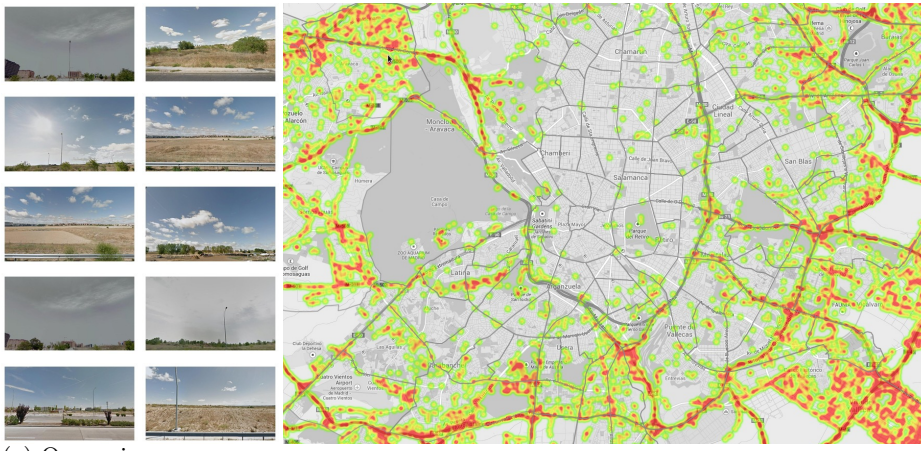
(b) Monumental

Fig. 6. Architecture style: Heatmaps *(right)* showing a density of different architecture styles across the city of Madrid. Notice that while *classical non-residential* style (a) is spread in the peripheries of the city. The *monumental* style (b) is mostly detected in the center however the detections are biased towards the training data. The *left* column shows several top-ranked images by learned predictor.

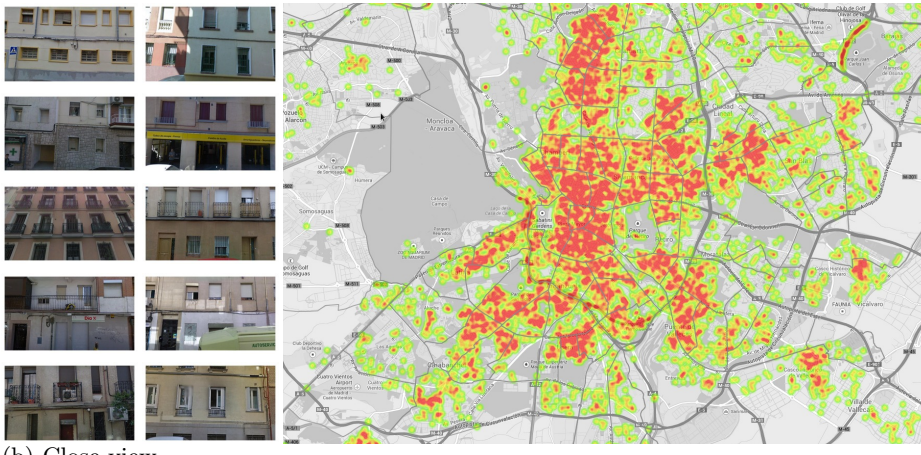
(a) Dense vegetation

(b) Loose vegetation

Fig. 7. Vegetation: Heatmaps *(right)* showing a density of vegetation across the city of Madrid. Notice a complementarity of the heatmaps. The *left* column shows several top-ranked images by learned predictor.



(a) Open view



(b) Close view

Fig. 8. View: Heatmaps *(right)* showing a type of view. The *open view* (a) is characteristic by long distance visibility without close-by obstacles. It is located in the suburbs while *close view* is spread across the city.