Facade Proposals for Urban Augmented Reality

Antoine Fond, Marie-Odile Berger, Gilles Simon

Facade Proposals for Urban Augmented Reality

Antoine Fond[1], Marie-Odile Berger[2], and Gilles Simon[1]

(a) Image from the database (b) Facade detection and recognition + warping of a 2D virtual object (c) Pose computation and projection of a 3D virtual object

Figure 1: Facade proposals for Urban AR. A facade of the Nantes Event Center building (a) is automatically detected and recognized in two views of the building (b, red polygons). From these results, any planar virtual object added to the facade (here the ISMAR logo) can be warped according to the transformation of the facade. (c) When some geometric information about the facade is available, any 3D virtual object expressed in the same reference frame as the facade (here the ISMAR building) can be added to the view.

ABSTRACT

We introduce a novel object proposals method specific to building facades. We define new image cues that measure typical facade characteristics such as semantic, symmetry and repetitions. They are combined to generate a few facade candidates in urban environments fast. We show that our method outperforms state-of-the-art object proposals techniques for this task on the 1000 images of the Zurich Building Database. We demonstrate the interest of this procedure for augmented reality through facade recognition and camera pose initialization. In a very time-efficient pipeline we classify the candidates and match them to a facade references database using CNN-based descriptors. We prove that this approach is more robust to severe changes of viewpoint and occlusions than standard object recognition methods.

1 INTRODUCTION

The purpose of Augmented Reality (AR) is to enhance the user's view of the real world with context specific information in such a way they appear to naturally belong to that world. In urban environments AR, buildings play a twofold role : they are the main object of interest for many AR applications [18] and they are semantically meaningful city-scale landmark to rely on for localization in GPS-denied areas [2, 4, 16, 17].

To overlay annotations on building facades such as anecdotes for tourism, advertisements for shopping, or addresses for city navigation aid in GPS-denied areas it is necessary to detect and recognize the building of interest in the camera image, followed by an estimation of the camera pose to correctly project virtual objects into the image.

Building detection from monocular images is a challenging task due to perspective deformations, repetitive structures and partial occlusions. Two categories of methods have been proposed in the past for the detection of building facades. Geometry-based methods attempt to identify rectangle facades in images rectified thanks to orthogonal vanishing points [9, 16]. Various geometric and photometric criteria are then used to characterize facades. However, they are generally too strict to take into account the variability of the facades encountered in urban areas.

On the other hand, with the advent of learning based techniques for classification, several methods have been designed with the aim to classify pixels or superpixels into categories. Among them, [8, 13] show examples of classification, "building" being one of the categories. Some are even tailored for classifying building subparts [11], [20]. Though promising, these methods do not allow to distinguish between adjacent facades. Recently a novel deep convolutional encoder-decoder architecture has been proposed in [3] for semantic pixel-wise labeling. By learning decoders to map the deepest layer features to full image dimensions, smoother predictions are obtained compared to [8]. The inference is also faster than [11],[20].

Object detection is traditionally formulated as a classification problem in the sliding windows paradigm in which classification is performed at each image location and scale in the image. Object detection has made great strides in recent years with the emergence of object proposals techniques. Their goal is to generate at high speed a reduced set of object bounding box proposals. A more complex classifier is then used in a second step for scoring.

The idea of defining an "objectness measure" for the pre-selection step, designed to produce a small number of regions such that top-ranked regions are likely to contain some categories, is developed in [1, 7, 29]. Though some methods integrate that objects have well defined closed boundaries [29], these methods are too general to be applied to the pre-selection of facades.

We thus propose in this paper a "facadeness" measure of image windows that can be evaluated rapidly and integrates geometric, photometric and semantic constraints. With respect to existing object proposal techniques, we especially introduce symmetry and repetitions constraints which are specific to building objects. We also propose to use semantic constraints based on the labelling of a SegNet-like network.

The aim of this work is to detect and identify facades for overlay purpose. Our framework for facade recognition does not require a city-scale 3D model but only a collection of fronto-parallel facade images and their associated virtual objects. However, as our method is partially geometry-based, we can also estimate the camera pose relatively to the detected facade as a by-product. Thus if this facade is part of a larger georeferenced 3D model our method can propose an approximate camera pose for geo-localization in the sense of [2, 6].

[1]Antoine Fond and Gilles Simon are with the Université de Lorraine, Loria, Vandœuvre-lès-Nancy, 54506, France antoine.fond@loria.fr, gilles.simon@loria.fr

[2]Marie-Odile Berger is with the Inria Nancy Grand Est, Villers-lès-Nancy, 54600 France marie-odile.berger@inria.fr

The paper is organized as follows: related work is described in section 2. Our method for facade proposals is explained in section 3. Application of our method for facade recognition is described in section 4. Results about facade proposals and facade recognition are presented in section 5.

2 State of the Art

2.1 Urban Augmented Reality

Many approaches have been proposed for augmented-reality in outdoor urban environnements. These methods can either be geometry-based or image-based. The former ones assume that a 3D model of the scene is available and seek to estimate the camera pose in order to project virtual objects into the image. In [22] the camera pose is tracked using measurements from prominent edges in the image and inertial-sensors data fused together with a Kalman filter. This method needs a fully textured 3D model. To relax this constraint, in [15] the authors propose to rely on building silhouettes rather than internal edges. Initial camera pose is refined using silhouettes matching through shape context descriptors and then tracked in a very similar way to [22]. The 3D model has to be sufficiently detailed so that building shapes can be matched. On the contrary in [2] the 3D model is very coarse using only building footprint and height. After finding vanishing points of the scene the method relies on vertical edges to generate translational hypotheses. Semantic segmentation is used to design a facade likelihood maximized relatively to the camera pose. In a similar way in [6] edges, semantic segmentation and windows detection are combined in a cost function that is minimized to find the pose. The main drawbacks of all these geometry-based approaches is that they strongly rely on GPS and inertial sensors to initialize the camera pose and they require an accurate 3D model. On the other hand, image-based methods do not require any scene geometry. Buildings are detected in the image and then matched to an image references database.

2.2 Facade detection

Several methods have been proposed in the past to detect rectangular structures in Manhattan worlds. In [16], line segments are automatically detected and intersected to generate hypotheses of rectangles in agreement with the vanishing points. For each hypothesis, the input image is orthorectified and a histogram of gradient (HOG) is computed inside the warped rectangle. Hypotheses whose HOG contains more than two dominant horizontal and vertical directions are discarded. This method is computationally expensive generating many superfluous hypotheses. To keep the problem tractable and efficient, Micusik et al. formulate the detection of the rectangles on a restricted neighborhood structure given by Delaunay triangulation [21]. The problem is then expressed as a search for the maximum aposteriori probability solution of a Markov random field. In [9], right-angle corners are detected in the orthorectified image using a Support Vector Machine. A Delaunay triangulation is performed from the right-angle corners and a min-cut-like algorithm is used to generate windows in which a high density of right corners is observed.

All these methods allow to detect rectangular structures appearing on facades, like windows or rows of windows, but not, in general, entire facades. In [19], the Gini Index is used to form an edge-based regularity metric relating regularity and distribution sparsity. The facade region detection is treated as a regional regularity/sparsity maximization problem, which is solved using greedy adaptive expansion over a down-sampled grid. Integer Quadratic Programming is then used to select a subset of facades that have maximum regularity score and facade coverage, with minimum overlap. However, the method still suffers from the use of a grid, and the regularity assumption makes it more suitable to large building facades with many regularly spaced windows than to the various kind of facades we consider in this work.

2.3 Object proposals

Landmark or building identification is of high interest in place recognition or pose computation with the goal to extract only roughly the buildings. As recognition from full images suffers from changes in viewpoints changes, the focus of research in place recognition has recently moved towards the identification of prominent landmarks in the scene through the use of object proposal methods thereby avoiding exhaustive sliding window search across images. [1] was the first to define the concept of "objectness", that is expressed as a score based on the combination of multiple cues (color contrast, edge density,...). Since then, several detection proposals have been proposed. For example, segmentation achieved at different scales is used as a selective search strategy in [27] whereas EdgeBox [29] uses object boundaries estimates for scoring. EdgeBox appears as the best compromise in speed versus quality in a comparative study conducted in [14] on then object proposal methods.

These techniques have been recently applied to place recognition [26]. Indeed comparing images on the basis of the whole image is sensitive to viewpoint changes. Comparing regions between images as proposed by EdgeBox [29] has proven to be more robust to viewpoints changes [26]. However, the cost of the method can be prohibitive if the fraction of buildings recovered by the object detector is small with respect to the number effective buildings present in the image. This leads to use a large number of proposal regions to be sure that a sufficient number of buildings can be matched between images.

We thus propose in this paper an efficient method for facade proposals which outperforms existing objectness methods. On the tested datasets, we show that only 100 candidates are required to obtain 84% of recall. This thus opens the way towards efficient methods for relocalization, place recognition and pose initialization.

3 Facade Proposals

Our algorithm for facade proposals consists in a two-stage procedure. A first set of facade candidates relying on contours is initialized. Ad hoc facade features (facadeness cues) are then evaluated on that set, and the best facade candidates are selected by combining the obtained values in a machine-learning framework. A database of 1500 images from Google Street View and ImageNet was used for learning purpose whereas testing was done on the 1000 images of Zurich Building Database (ZuBuD)[0]. Each image is orthorectified and the bounding boxes of the facades are provided manually. These bounding boxes are referred as ground truth (GT) in the following.

3.1 Geometry of the scene and rectification

Manhattan hypotheses are well suited for modeling urban environments. All the buildings of the scene are considered to be 3D boxes and shall be parallel one to another. Each of the box faces is a facade except for top and down faces. Thus the geometric shape of a facade is a rectangle and its texture is defined by different visual characteristics that will be described below.

The method of [24] is used in this paper to find the Manhattan vanishing points. The intrinsic parameters of the camera as well as the homographies that warp the image of the scene to orthorectified images are automatically computed. Doing so, all the vertical facades of the scene appear in either of these orthorectified images as in a frontal view.

3.2 Joint semantic parsing and contours detection

Semantic parsing does not solve the problem of facade detection by itself as it cannot distinguish between two nearby facades. How-

[0] http://www.vision.ee.ethz.ch/showroom/zubud/

ever we exploit this pixel-wise labeling in ours facadeness cues. We trained a modified version of SegNet to infer 7 semantic classes (background, facade, window, balcony, door, sky and road) as well as contours (see Fig. 2). Jointly solving these two problems enables semantic parsing to be more aware of edges and contours to be more located on meaningful edges (Fig. 3). The training/testing database is a label-consistent merge of CMPfacadeDB[1], eTrims[2], ECP[3], INRIA[4] and LabelMeFacade [10]. It contains a variety of classic and modern European style buildings from different cities (Paris, Prague, Berlin,...). Ground truth contours of that database are contours between different semantic areas. The network architecture is the same as the regular SegNet but the final deconvolutional layer has 9 outputs (7 for semantics and 2 for contours) that are sliced into 2 different layers. The final loss function used for training is a weighted sum of the logistic loss functions of these layers.

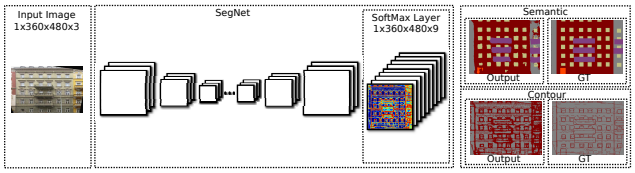

Figure 2: Architecture of our modified version of SegNet with two outputs : a semantic labeling map and a contour map.

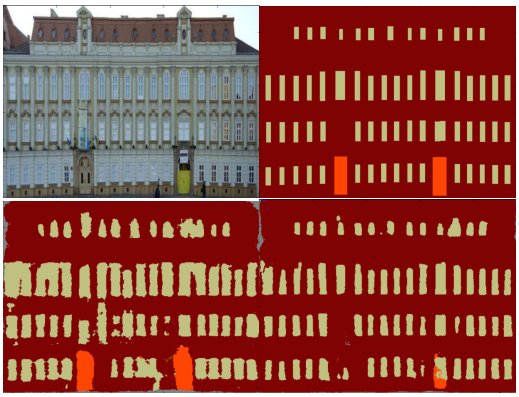

Figure 3: Result of the joint training on semantic inference map. First row shows an image of facade (left) and its ground truth semantic labeling (right). Second row shows the inference map of a standard SegNet (left) and our modified SegNet (right). The shape of the windows are much more rectangular in our version due to the joint training of semantic and contours.

3.3 Rectangular candidate sampling

The main hypothesis we made on facades is that they are rectangular-shaped. As we work with orthorectified images we are explicitly looking for rectangles. We choose to rely on the contours of the image to generate a first set of candidates. Indeed, the border of a facade should create high gradient values on the image. Edge map E is one of the two outputs of our modified version of SegNet. These edges are then accumulated in both a histogram of vertical-projected edges H_x and a histogram of horizontal-projected edges H_y. The product $H_x H_y^T$ can be seen as a corners likelihood map (see Fig. 4). The n local maxima of that map are used to generate $\frac{n(n-1)}{2}$ rectangles. Actually, as both (top-left,bottom-right) and (top-right,bottom-left) pair of corners define the same rectangle, only a set of $\frac{n(n-1)}{4}$ facade candidates are retained. For instance, for the 1500 images in our learning database, the average number of facade candidates per image is 16288 at this very first step.

[1] http://cmp.felk.cvut.cz/ tylecr1/facade/
[2] http://www.ipb.uni-bonn.de/projects/etrims_db/
[3] http://vision.mas.ecp.fr/Personnel/teboul/data.php
[4] https://github.com/raghudeep/ParisArtDecoFacadesDataset/

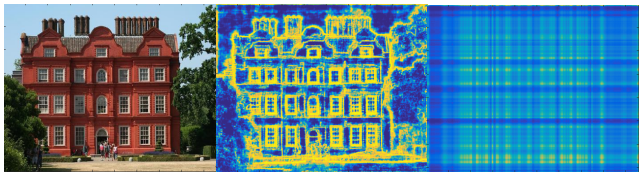

Figure 4: Example image from our database (left), with the contour map (middle) and the corners likelihood (right).

3.4 Facadeness cues

Facades share several common visual characteristics. They are usually composed of rectangular features such as floors, windows, balconies, doors. These features repeat themselves along the facade in both vertical and horizontal directions. Facades are also roughly symmetrical. Eventually facades are homogeneous in color at least compared to their background. For each of these facade candidates we evaluate 6 different had hoc features (cues). We here reuse the color contrast and shape cues defined in [1]. With respect to [1] a stronger edge cue is defined which favors vertical and horizontal segments. Three new criteria are introduced aiming at characterizing semantic contrast, symmetry and repetitive patterns on facades. Subsequently the combination of all these cues enables us to discard the candidates that do not match our facade hypotheses and keep only the best ones. For each cue presented below, Fig. 5 shows the best rectangle obtained among all candidates, in an example image of our database (right column) and the probabilities of the cue values to be obtained on a facade (in green) or on a non facade (in red).

3.4.1 Shape cue

Facades are rectangular-shaped, but all rectangles are not as likely to be observed. Indeed, architectural rules allow just a few values of the facade aspect-ratio. Extremely thin facades are almost impossible for example. We have learned the probability distribution of two rectangular parameters (height and width) on our 1500 learning images (Fig. 5, top-right image) in the same way as in [1]. We use a 24×24 discretized version H of that distribution for efficiency $s_{shape}(r) = H(h,w)$.

3.4.2 Color cue

Color itself is a poorly informative feature to describe facades as they can have many different colors. However the color homogeneity of a facade compared to its local context is a much more interesting feature as it is described in [1]. The difference of color distribution between the inside of the rectangle and the surrounding region can distinguish facades as in Fig. 5, last row:

$$s_{color}(r) = 1 - \exp\left(-d_{\chi^2}\left(H_c^{b(r,\beta)}, H_c^r\right)/\sigma_c\right) \qquad (1)$$

where H_c^r and $H_c^{b(r,\beta)}$ are respectively the color histogram of the inside of r and the color histogram of the band of thickness β surrounding r. We use LAB color space quantized into $256 = 4 \times 8 \times 8$ bins.

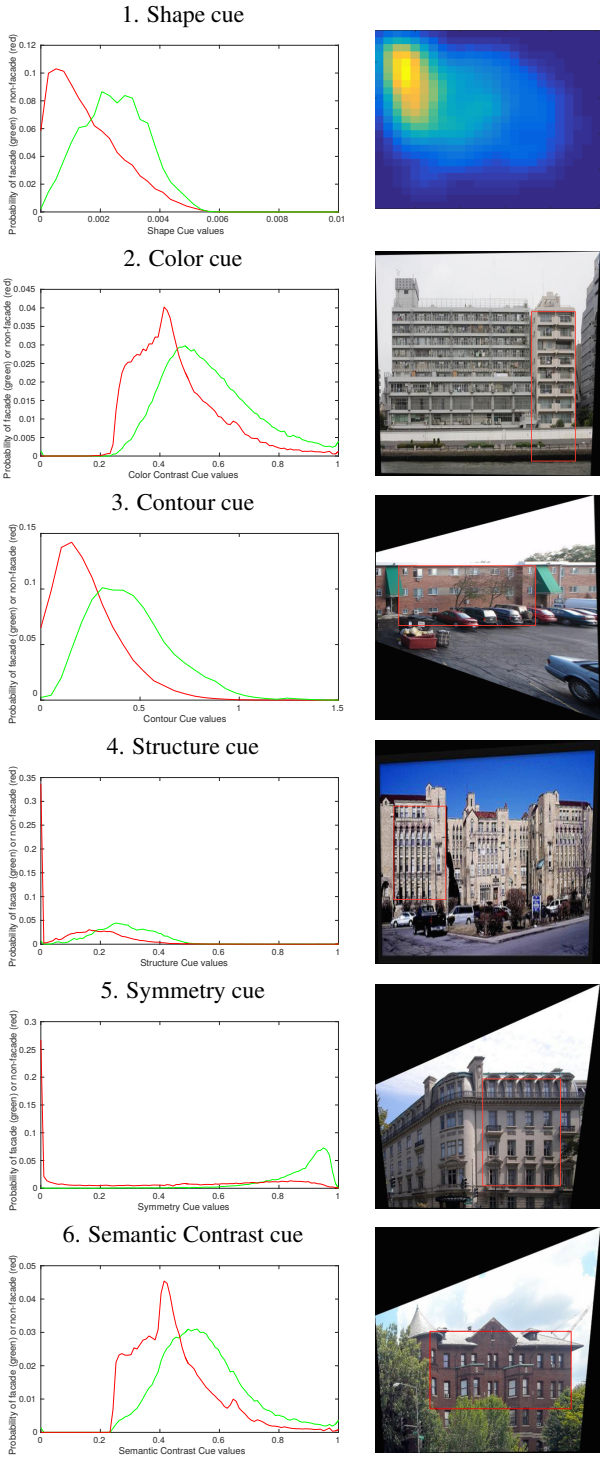

Figure 5: Left column: probabilities for each cue values to be obtained on a facade (in green) or on a non facade (in red). Right column: best rectangle obtained among all candidates for each cue, in example images of our database (for the shape cue, the whole heatmap of the histogram H is shown).

3.4.3 Oriented contour cue

As facades are rectangular-shaped, we can expect high gradient values along their border. More precisely we can expect vertical (respectively horizontal) contours along the vertical (respectively horizontal) edges of their bounding box :

$$s_{cont}(r) = \frac{1}{2\alpha(l+h)} \sum_{b_l(r,\alpha)\cup b_b(r,\alpha)} E_x + \sum_{b_l(r,\alpha)\cup b_r(r,\alpha)} E_y \quad (2)$$

where α is the thickness of the band $b_x(r,\alpha)$ positioned on the $x \in \{top, bottom, right, left\}$ of the rectangle r. E_x and E_y are the binary images of the horizontal and vertical SegNet contour. (h,w) are the height and width of r, respectively.

3.4.4 Structure cue

Windows and balconies repeat themselves along both vertical and horizontal directions on facades. The "window" or "balcony" labels projected on the horizontal axis are binned in an histogram H_x^r so are the same labels projected on the vertical axis in H_y^r. The autocorrelation of both these two signals is sparse if there are strong repetitions. We thus define the structure cue :

$$s_{struc}(r) = W(r)\left(\frac{\sum_{peaks} R(H_x^r)}{\sum R(H_x^r)} + \frac{\sum_{peaks} R(H_y^r)}{\sum R(H_y^r)}\right) \quad (3)$$

where H_x^r and H_y^r are the 32 bins-normalized-histograms defined above for the rectangle r. $R(f) = \mathscr{F}^{-1}|\mathscr{F}(f)|^2$ is the autocorrelation of f. $\mathscr{F}$ and $\mathscr{F}^{-1}$ are respectively the Fourier transform and inverse Fourier transform. Peaks are local maxima of the signal. $W(r)$ is the sum of the labels "facade", "window", "balcony", "door" inside r normalized by the area of r.

3.4.5 Symmetry cue

Facades have a non-perfect axial symmetry. What we want is a metric that evolves continuously with the symmetrical aspect of the facade. For example the cross-correlation between the left and the right half of the image would be very high for even a small asymmetry. We propose to subdivise the rectangle into 16 patches. For each of these patches we compute the HOG descriptor with 8 bins on SegNet contour. Then we evaluate the distance between each of the 8 patches on the left with their symmetrical patch on the right :

$$s_{sym}(r) = \exp\left(-\sum_{i=1}^{4}\sum_{j=1}^{2} \frac{d_{\chi^2}\left(H_e^{sym}(s(i,j)), H_e(i,j)\right)}{8\sigma_s}\right) \quad (4)$$

where $H_e(i,j)$ is the HOG descriptor with 8 bins of the patch (i,j). $H_e^{sym}(i,j)$ is the flipped version of vector $H_e(i,j)$. s is the axial symmetry of vertical axis and d_{χ^2} the χ^2 distance.

3.4.6 Semantic contrast cue

Facades are composed of semantic features such as windows, balcony, door, walls in specific proportions. Their proportions in one facade differ from the proportions of the surrounding region that can include other semantic features like portions of sky or roads.

$$s_{sem}(r) = 1 - \exp\left(-d_{\chi^2}\left(H_s^{b(r,\gamma)}, H_s^r\right)/\sigma_{sc}\right) \quad (5)$$

where H_s^r and $H_s^{b(r,\gamma)}$ are respectively the histogram of the semantic labels inside r and inside the band of thickness γ surrounding r.

Computation of all the cues is in constant time for one rectangle thanks to the use of integral images. This trick is detailed in [28] for the computation of sums in regions as well as local histograms. The quantification of the histograms and the number of patches have been chosen to keep that constant time reasonable (below 10^3 operations). The parameters α, β, γ have been learned on our training set (see section 3.5) so as to maximize the separability between positive and negative probability distributions of the cues values (Fig.

5, left column). The optimal values for these parameters are 5%, 30% and 20% of the dimensions of the rectangles, respectively. σ_c, σ_s, σ_{sc} are the standard deviation of the distances used respectively in the color, symmetry and semantic contrast cues.

3.5 Cues combination

Intersections between facade and non-facade probability distributions of cues values (Fig. 5, left column) mean that one cue alone cannot separate between facades and non-facades. To combine all these features into a single metric we use a multi-layer perceptron. It is composed of two hidden layers of 8 neurons. This neural network has been trained on positive and negative examples, taken from the rectangle sets generated by the sampling procedure presented in section 3.3, applied to all images of the learning database. To decide if a rectangle is a positive or a negative example, we used the commonly used metric "Intersection over Union" (IoU score s_{IoU}) [29]. An IoU threshold of 0.5 is often used in the literature to decide whether or not two image regions coincide. Moreover, an illustration in [29] shows that an IoU score of 0.5 already indicates a relatively high overlap. For these reasons, we took as positive examples rectangles that overlap the GT with $s_{IoU} \geq 0.5$, whereas negative examples are candidates with $s_{IoU} < 0.5$. This set of examples will be referred as our training set.

The final output of the perceptron can be seen as a probability score of being a facade. All candidates are sorted using this metric and a greedy algorithm keeps the best candidates that do not overlap strongly a higher-ranked rectangle (according to $s_{IoU} \geq 0.5$) as in [1]. Figure 6 shows the recall rate obtained on ZuBuD database as a function of the number of proposals. We compare the combination of our problem specific cues (contour, structure, symmetry and semantic contrast) to the combination of already-defined cues (contour, shape, color contrast) and the combination of all cues. A GT rectangle is counted retrieved when at least one of the selected candidates overlaps it with a $s_{IoU} \geq 0.5$. In practice, we use the 100 first proposals which corresponds to 85 % of recall. We can use much fewer proposals than state of the art methods for the same performance and about the same computational time (0.42s).

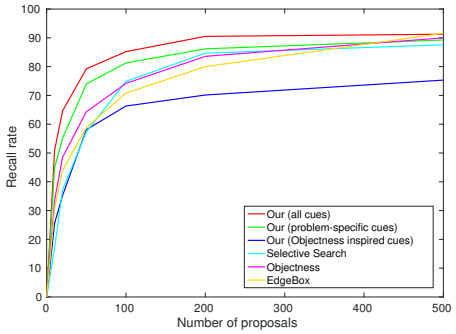

Figure 6: Recall rate as a function of the number of candidates chosen on ZuBuD test database. 3 versions of our facade proposal method with different cues combinations are shown : (in red) all the cues, (in green) only the problem-specific cues (contour, structure, symmetry, semantic contrast), (in blue) only the objectness-inspired cues (contour, shape, color contrast). We also compare our method to other object proposals methods.

4 Application : Facade Recognition

Our facade proposals method could be used to improve [2]. Indeed it could help to generate translational hypotheses and disambiguate situations with many adjacent facades lying on a single plane in the alignment step. However we choose to highlight another application of our method : facade recognition.

The problem we refer to as facade recognition and that we are trying to solve in this part is the following: taking a new single picture of a known urban place, we want to recognize the most prominent facades of this image. By recognition we mean not only detecting the facades but also match them to the correct reference in the facade database that represents our place. This problem is different from place recognition as we additionally intend to roughly locate the facades existing in the database in the considered image. Solving this problem opens the way towards AR applications in urban context such as practical or cultural information overlay on the building or may help to guide the users towards a desired building.

In the context of urban augmented reality the lack of GPS accuracy does not allow a seamless overlay information about the buildings in the image. However the GPS may help reduce the numbers of buildings to be matched to a dozen.

Our facade recognition method is threefold (Fig. 7):

1. **Generate a few number of facade candidates** through our facade proposal method.

2. **Classify these candidates into "facade" and "non-facade"** through a neural network using SPP descriptors [12] with surrounding context as inputs. Resorting to spatial pyramid pooling allows us to compute the feature maps from the entire image only once, avoiding repeatedly computing the convolutional features for each box proposal.

3. **Match the remaining facades with the facade database** of the current place using a semantic metric learned through a siamese neural network taking SPP descriptors as inputs.

4.1 Facade classification

The main difference between facade classification and general object classification is that facades cannot be described only from the inside of their bounding box. Indeed, a part of a facade may be visually a facade too (the entire facade cropped so that one floor is missing for example). The only way to avoid this multiple parts problem is to consider the surrounding visual context. A true facade shall look like a facade inside but its context may differ. We propose to build a descriptor by concatenating the SPP descriptor inside the rectangle with SPP descriptors of the surrounding band. The SPP descriptor of the inside is computed from the 5th convolution (Conv5) layer of our modified SegNet. We use a 3-levels spatial pyramid ($1 \times 1, 2 \times 2, 4 \times 4$) to pool the values of the 512 feature maps of Conv5. As the resolution of Conv5 is pretty low (23 $\times$ 30) and the surrounding band too thin we use a different scheme for the spatial pooling of the band. The surrounding band can be divided into 4 bands (top, down, left, right). For each band we use a 2-levels spatial pyramid (1×1, 1×4). Consequently the SPP descriptor of the inside has 10752 dimensions whereas the SPP descriptor of the band has 10240. Their 20992 concatenated vector is the input of a neural network classifier. The classifier is composed of 2 more hidden linear layers of size 4096. They have been trained on an augmented version of our training set. Indeed we have added synthetic data to our training set. These synthetic samples are plain facade images from ImageNet and GoogleStreetView pasted in urban context images. The positive and negative samples are defined in the same way as for the training of the cue combination.

4.2 Facade matching

At this step, the descriptor associated with each candidate is their inside SPP descriptor already computed during facade classification.

However distinguishing between two different facades and still be robust to few appearance changes of the same facade can be challenging. This problem is similar to fine-grained classification and

is generally solved by learning a similarity metric using a siamese network [5]. The idea is to map images in a low dimensional space so that the distance between points of the same class is small and is large for points from different classes.

Here, the inside SPP descriptor is the input of such a neural network with 2 layers of size 2048 trained in the siamese way on a third of the ZuBuD database. The positive pairs are generated from positive samples ($s_{IoU} \geq 0.5$) of different views of the same facade whereas negative pairs are generated from positive samples of different facades. The embedded space induced by this siamese network is then much smaller and tuned to distinguish facades that can be visually similar. We then compute the euclidian distance between the siamese network outputs of both the candidates and the facade references from the database. For each candidate we choose its closest neighbors in the facade reference database. To assure that the match is correct we apply crosschecking i.e. we intend the closest neighbors to be both-ways.

5 Experimental Results

5.1 Facade proposals

For testing we use the Zurich Buildings Database (ZuBuD). This database is composed of 1000 pedestrian street views of Zurich. It is divided into 200 scenes each one focusing on one building. The changes of viewpoint are severe and it is not rare for the facades to be occluded by trees, street lamps or electric lines. There is also a good diversity of buildings with different architectures from classic European style to more modern ones. For all of these reasons we evaluate our facade proposals method on that dataset. We compare our method to other object proposals methods through 3 different points : recall, precision and time (Tab. 1).

Recall is one of the most important measurement for an object proposals method. It is computed as the rate of object recovered among the n first proposals. If most general object proposals methods perform well on ZuBuD over $n = 500$ proposals with more than 90 % recall our method shows much better results below. Indeed with $n \leq 100$ proposals our method has a constant gain of more than 10 % compared to state of the art object proposals methods (Fig. 6).

This improvement can be explained by the use of cues that are more discriminative for facades than general objects. More precisely some assumptions that are made by general object proposals method are violated in urban environments. Selective Search is based on super-pixel merging. As the local context is not considered there is no way to distinguish between subparts and plain facades which leads to a drop in recall for few proposals. The main assumption of EdgeBox is that the edges are wholly enclosed inside proposals. In the case of adjacent facades, there are usually mutual edges between facades that overlap the proposal. In Objectness the way to sample the proposals in the first place is based on changes in the frequency distribution of natural images. But in urban environments the frequency distribution is already biased by verticals edges and strong repetitions. Eventually none of these methods take into account the rectification of the image whereas the symmetry and the structure cues use this information in our method. These poorly suited assumptions for urban environments affect the ranking of proposals. More specifically bounding boxes merging adjacent facades are usually ranked up top by general object proposals methods causing the true facade to be ranked further down. Thus more proposals are needed for a further task (e.g. detection or recognition) while $n = 100$ proposals are enough in our method.

The precision is computed here as the average value s_{IoU} of all the proposals that overlap the GT with $s_{IoU} \geq 0.5$ for $n = 100$. This measurement is not critical for object proposals but for facade recognition it is important to accurately locate the facade. Our better results in precision compared to others can be explained by the use of cues based on contours along the facade boundaries and surrounding context in both color and semantic.

Table 1: Statistics of facade proposals

	Selective Search	Objectness	EdgeBox	Our
Recall ($n = 100$)	74.89	74.18	70.81	85.16
Precision (s_{IoU})	0.59	0.63	0.61	0.68
Time (seconds)	0.28	1.42	0.35	0.42

The second measurement that actually matters for object proposals is time. All our cues are computed in constant time thanks to integral images and spatial pooling. That allows us to get a computational time which is better or in the same order of magnitude than the others. This time is compatible with augmented reality applications. All the codes are written in Matlab with critical parts executed in C. The computational times shown in Tab. 1 are the average times for $n = 100$ proposals on I7-3520M CPU with an Nvidia TITAN X GPU for the forward pass in SegNet.

5.2 Facade Recognition

The test set of our facade recognition application is the 2/3 of ZuBuD that have not been used for the metric learning. There are 937 GT facades tagged on this part of ZuBuD. These facades are grouped into 171 classes, each one representing a unique facade. Each class gathers images of the same facade but observed through different viewpoints. Consequently images from the same class have different rectification artifacts, different resolutions (Fig. 8). Due to occlusions they can be cropped versions of each other and some parts can be hidden or not. In addition to all this diversity intraclass different classes can be visually similar from one another. To represent the whole class we choose the facade reference the closest to the frontal view with the fewest occlusions.

In such challenging conditions classic matching approaches fail to match a facade to its correct reference (Tab. 2). We have evaluated the matching performance of our facade matching method compared to others in our facade recognition pipeline. For every approach we first execute our facade proposals method with $n = 100$ candidates. For each ground truth facade of the ZuBuD test set we select the candidate that overlaps it the most in the sense of s_{IoU}. Thus these facade candidates are the best we can expect using our facade proposals method no matter how well our detection step performs. For each of these facades we compute their descriptor and we search their closest neighbor in the 171 reference database for Euclidean distance. A match is correct if the class of its closest neighbors is the same as its ground truth class. Tab. 2 shows the results of different descriptors used for place recognition. Bag of Words (BoW) descriptor is the histogram of visual words (SIFT) inside the facade candidate. VGG is the 4096 output vector from VGG CNN[25] applied to the subimage of the facade candidate resized to fit the input of the network. SPP is the SPP descriptor from Conv5 of our modified SegNet and SPP (siamese) the output of the siamese network for that same vector.

Table 2: Statistics of facade matching

	BoW	VGG	SPP (SegNet)	SPP (siamese)
Correct match (%)	44.06	72.80	78.91	82.93
Time (seconds)	0.29	2.19	0.07	0.05

The very few number of SIFT features extracted in facades,

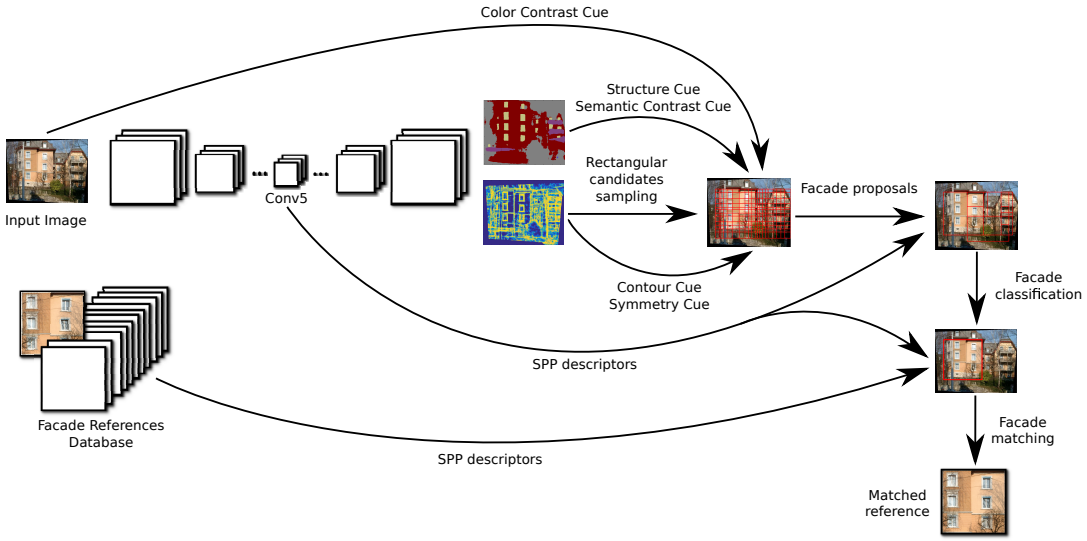

Figure 7: Overview of the whole facade recognition method.

Figure 8: Illustration of the diversity of facades that belong to the same class with rectification artifacts, occlusions and partial observation.

the similarity of their descriptors in repetitive structures can explain the poor results of BoW. Moreover without any spatial information BoW can only discriminate between facades through their proportion of visual words that is clearly not sufficient. The difference between VGG and SPP is mostly imputable to the fine tuning of our modified SegNet for buildings segmentation. Indeed the architecture of both networks is pretty similar and SPP is essentially a speed-up approximation of a true CNN descriptor. The Conv5 layer of our network yet contains more meaningful information about facades rather than an ImageNet trained VGG. Eventually metric learning through siamese networks helps for fined-grained classification.

We then evaluate the whole pipeline including the detection step. Example results are shown in Fig. 10. At the end of the facade proposals we have 100 facade candidates. As the matching is based on the closest neighbors we cannot apply it directly to each of the candidates. Indeed each candidate will always find a closest reference in the database even if no facade of that class is present in the image. We need to select the facades that are actually visible in the image and only them. The facade classification based on the inside and surrounding regions remove most of the candidates and the cross-checking assures that the match is correct. Thus they are few false detections 16.14 % with still a high recall of 71.13 % (SPP siamese) for the whole test set. These results outperforms BoW and are comparable to VGG even though the facade selection was biased to the best possible solution in their case. False detections do not necessarily mean the detection has completely failed. It can usually be explained by one of the following situations: the detected facade is too small but included in the GT facades (image 1 in Fig. 11), the detected facade groups several GT facades into one (image 3 in Fig. 11), the detected facade has been missed in GT tagging (image 1 in Fig. 11). However we have noticed one regular failure case that happened when small detected facade with poor photometric information match with a quasi-homogenous reference facade (image 1 in Fig. 11). The matching also can fail when there are two very similar facades in the database. For example, this situation can occur for facades coming from the same building (image 4 in Fig. 11).

We also evaluate the whole pipeline on a smaller set that comes from the Street part of Cambridge Relocalisation Dataset[5]. This set is composed of 80 images divided into 20 classes. That situation is more suitable for a real AR application using GPS localisation to prune the database. Unlike ZuBud there are no occlusions but the changes of viewpoint are more extreme. The statistics are pretty similar to ZuBuD with 73.38 % of recall and 19.34 % of false detections. Thus our method proves to be robust to severe changes of viewpoints as well as partial observations and occlusions (Fig. 12).

Another benefit of our pipeline for facade recognition is speed. As we use SPP-based descriptors we only need one single forward pass in our modified SegNet for the whole method. We reuse the outputs of the latter in different parts of the algorithm (rectangles sampling, cues computation,...) as well as its Conv5 feature maps for descriptors. Our method can be seen as an initialization for further more accurate pose estimation [15, 22]. As an initialization process it does not need to be executed at each frame but only at the beginning and when the tracking process fails. The time efficiency of the pipeline (0.45s) makes it compatible with AR applications as a new facade recognition step can be computed every 11 frames. We could imagine that detected facades could be tracked the rest of the time in such application. Thus the initialization could be processed server-side whereas the tracking could be done on the mobile device. Although the lack the GPU power is still a limitation for our method to perform in real-time on regular mobile-device, we believe that future generations of mobile devices will overcome this limitation (the SegNet inferences on an Nvidia TX1 embedded GPU requires 0.7s, contrary to 0.1s in our desktop setup). As most mobile-device manufacturers tend to develop embedded GPU in their devices hardware, we expect that in the future real-time CNN inference will become possible with commodity devices. This trend is also supported by major software corporations,

[5]http://mi.eng.cam.ac.uk/projects/relocalisation/#dataset

e.g., Facebook (Caffe2Go) and Google (Tensorflow Lite).

5.3 AR applications

Once facades in the image have been detected and identified it is already possible to overlay planar virtual objects onto them. It is sufficient to project the boundaries of the detected facades back into the original image through the inverse homography used to rectify the image. For example in Fig. 10 each colored area represents a facade that has been successfully detected and matched to its reference. These areas could be easily replaced by any meaningful information specific to their building (Fig. 1,b). However, if in addition to our facade references database each of these references are augmented with geometric information (such as a georeferenced frame and their real world dimensions) a complete 6DOF camera pose can be estimated (Fig. 1,c) for each recognized facade. Let us assume such geometrically-enriched model is available for a facade that has been detected and identified by our method. We already have the rotation R that transforms the camera frame into the local frame of the facade from the rectification step so only the translation T needs to be computed. That can be done by associating facade corners in the image with corresponding real world facade corners from the geometric information [23]. The real world dimensions are important to have a shared scale between detected facades and associated virtual objects (such as a 3D model of buildings). The georeferenced frame enables to transform the camera pose relative to the facade to an absolute pose for geolocalization purpose. As an example, we used a public 3D model[6] of the Nantes Event Center building where ISMAR 2017 will take place and we tried to recognize one of its facade on GoogleStreetView images showing the building from different viewpoints. The facade reference has been added to the ZuBuD references database used in the previous test. In all test images the facade has been successfully detected (always in the top-10 proposals) and recognized as part of the Nantes Event Center building. We choose to describe both applications depending on what information we suppose about the database. If the database is only made of facade images with associated planar virtual information (the ISMAR logo for Nantes Event Center building), we overlay it onto the detected facade (Fig. 9,1st row and Fig. 1,b). If we assume geometric information about the facades is available in the references database, we project the wireframe of the 3D model of the building associated to the recognized facade back into the image (Fig. 9,2nd row and Fig. 1,c).

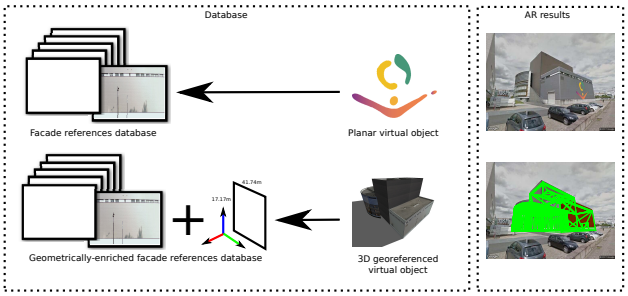

Figure 9: Examples of facade recognition for AR applications. 1st row shows the case of an image-only database with a facade reference and its associated feature overlaid in the images. 2nd row shows the case of a geometrically-enriched database with a 3D model and its wireframe version projected in the images.

6 Conclusion

We presented a fast facade proposals method that can be applied in a very efficient way to facades recognition and camera pose initial-

[6] http://www.3dwarehouse.sketchup.com

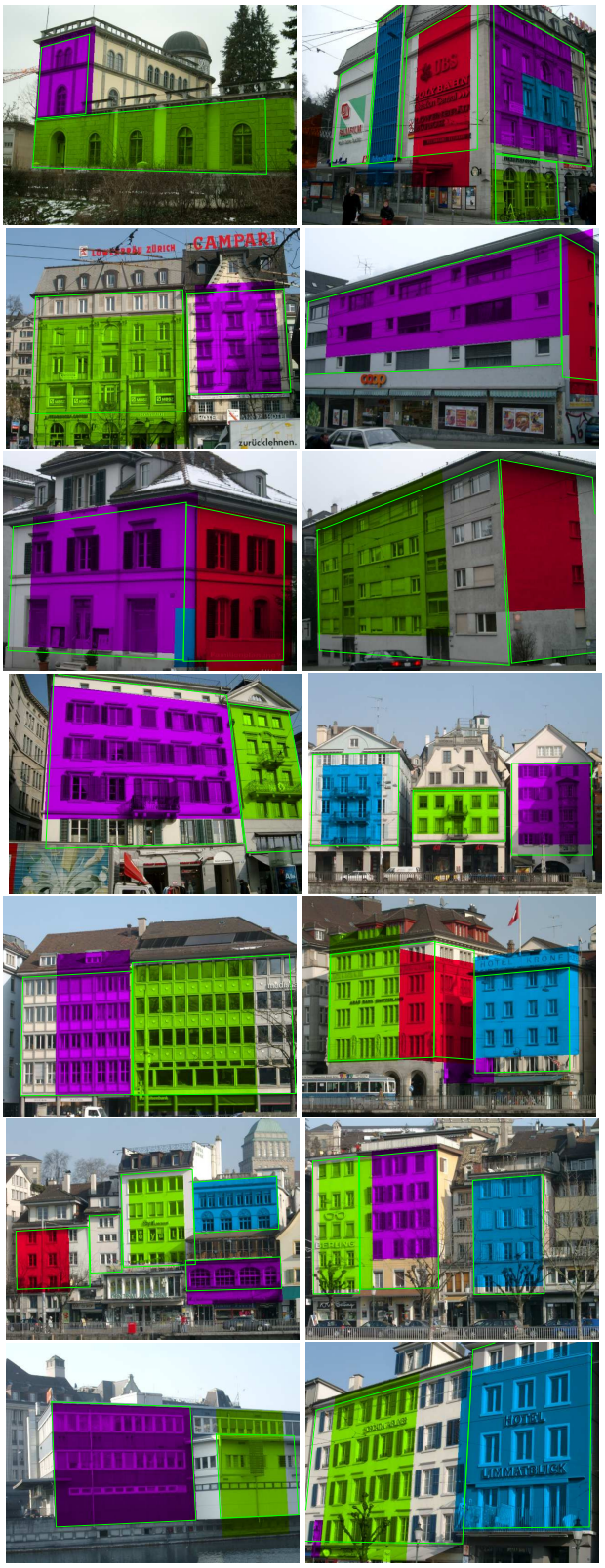

Figure 10: Example results of the whole facade recognition method obtained on the Zubud database. The green polygons are the ground truth facades. All the facades shown in this figure were correctly recognized.

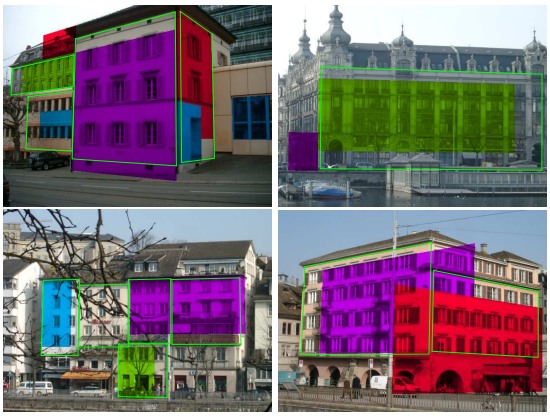

Figure 11: Example of failures with different cases of false detections (images 1, 2 and 3) and matching failure with two different facades matched to the same reference (image 4)

Figure 12: Example results of the whole facade recognition method obtained on the Cambridge database. All the facades shown in this figure were correctly recognized.

ization in urban environments. We demonstrated the relevance of combining ad-hoc features and deep learning framework together for object-specific localization and recognition. Though the invariance of CNN descriptors to small translations makes our method not currently suited for accurate pose estimation, it proposes a good initialization that could be refined with a further gradient-based image registration method based on deep learned-metric.

REFERENCES

[1] B. Alexe, T. Deselaers, and V. Ferrari. Measuring the Objectness of Image Windows. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(11):2189–2202, 2012.

[2] C. Arth, C. Pirchheim, J. Ventura, D. Schmalstieg, and V. Lepetit. Instant Outdoor Localization and SLAM Initialization from 2.5D Maps. In *Proceedings of the IEEE International Symposium on Mixed and Augmented Reality*, pages 1309–1318, Fukuoka, Japan, 2015.

[3] V. Badrinarayanan, A. Handa, and R. Cipolla. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Robust Semantic Pixel-Wise Labelling. *CoRR*, abs/1505.07293, 2015.

[4] D. M. Chen, G. Baatz, K. Koser, S. S. Tsai, R. Vedantham, T. Pylvanainen, K. Roimela, X. Chen, J. Bach, M. Pollefeys, B. Girod, and R. Grzeszczuk. City-scale Landmark Identification on Mobile Devices. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 737–744, Colorado Springs, USA, 2011.

[5] S. Chopra, R. Hadsell, and Y. LeCun. Learning a Similarity Metric Discriminatively, with Application to Face Verification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 539–546, San Diego, USA, 2005.

[6] H. Chu, S. Wang, R. Urtasun, and S. Fidler. HouseCraft: Building Houses from Rental Ads and Street Views. In *Proceedings of the European Conference on Computer Vision*, pages 500–516, Amsterdam, The Netherlands, 2016.

[7] I. Endres and D. Hoiem. Category Independent Object Proposals. In *Proceedings of the European Conference on Computer Vision*, pages 575–588, Berlin, Germany, 2010.

[8] C. Farabet, C. Couprie, L. Najman, and Y. LeCun. Learning Hierarchical Features for Scene Labeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1915–1929, 2013.

[9] A. Fond, M.-O. Berger, and G. Simon. Prior-based Facade Rectification for AR in Urban Environment. In *Proceedings of the IEEE International Symposium on Mixed and Augmented Reality workshop on Urban Augmented Reality*, pages 94–99, Fukuoka, Japan, 2015.

[10] B. Fröhlich, E. Rodner, and J. Denzler. A Fast Approach for Pixelwise Labeling of Facade Images. In *Proceedings of the International Conference on Pattern Recognition*, pages 3029–3032, Istanbul, Turkey, 2010.

[11] R. Gadde, V. Jampani, R. Marlet, and P. Gehler. Efficient 2D and 3D Facade Segmentation using Auto-Context. *CoRR*, abs/1606.06437, 2016.

[12] K. He, X. Zhang, S. Ren, and J. Sun. Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition. In *Proceedings of the European Conference on Computer Vision*, pages 346–361, Zurich, Switzerland, 2014.

[13] D. Hoiem, A. A. Efros, and M. Hebert. Automatic Photo Pop-up. In *Proceedings of the ACM SIGGRAPH*, pages 577–584, Los Angeles, USA, 2005.

[14] J. Hosang, R. Benenson, and B. Schiele. How Good are Detection Proposals, really? In *Proceedings of the British Machine Vision Conference*, Nottingham, England, 2014.

[15] J. Karlekar, S. Z. Zhou, W. Lu, L. Z. Chang, Y. Nakayama, and D. Hii. Positioning, tracking and mapping for outdoor augmentation. In *Proceedings of the IEEE International Symposium on Mixed and Augmented Reality*, pages 175–184, Seoul, Korea, 2010.

[16] J. Košecká and W. Zhang. Extraction, Matching, and Pose Recovery Based on Dominant Rectangular Structures. *Computer Vision and Image Understanding*, 100(3):274–293, 2005.

[17] H. Li, D. Song, Y. Lu, and J. Liu. A Two-view based Multilayer Feature Graph for Robot Navigation. In *Proceedings of the IEEE International Conference on Robotics and Automation*, pages 3580–3587, St. Paul, USA, 2012.

[18] F. Liu and S. Seipel. Detection of Facade Regions in Street View Images from Split-and-Merge of Perspective Patches. *Journal of Image and Graphics*, 2(1):8–14, 2014.

[19] J. Liu and Y. Liu. Local Regularity-Driven City-Scale Facade Detection from Aerial Images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3778–3785, Columbus, USA, 2014.

[20] A. Martinovic, M. Mathias, J. Weissenberg, and L. J. V. Gool. A Three-Layered Approach to Facade Parsing. In *Proceedings of the European Conference on Computer Vision*, pages 416–429, Florence, Italy, 2012.

[21] B. Micusík, H. Wildenauer, and J. Kosecka. Detection and Matching of Rectilinear Structures. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–7, Anchorage, USA, 2008.

[22] G. Reitmayr and T. Drummond. Going out: Robust Model-based Tracking for Outdoor Augmented Reality. In *Proceedings of the IEEE International Symposium on Mixed and Augmented Reality*, pages

109–118, Santa Barbara, USA, 2006.

[23] G. Simon, A. Fitzgibbon, and A. Zisserman. Markerless Tracking using Planar Structures in the Scene. In *Proceedings of the IEEE International Symposium on Mixed and Augmented Reality*, pages 120–128, Munich, Germany, 2000.

[24] G. Simon, A. Fond, and M.-O. Berger. A Simple and Effective Method to Detect Orthogonal Vanishing Points in Uncalibrated Images of Man-Made Environments. In *Proceedings of Eurographics*, Lisbon, Portugal, 2016.

[25] K. Simonyan and A. Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. *CoRR*, abs/1409.1556, 2014.

[26] N. Suenderhauf, S. Shirazi, A. Jacobson, F. Dayoub, E. Pepperell, B. Upcroft, and M. Milford. Place Recognition with ConvNet Landmarks: Viewpoint-Robust, Condition-Robust, Training-Free. In *Proceedings of Robotics: Science and Systems*, Rome, Italy, 2015.

[27] J. Uijlings, K. van de Sande, T. Gevers, and A. Smeulders. Selective Search for Object Recognition. *International Journal of Computer Vision*, 104(2):154, 2013.

[28] P. Viola and M. J. Jones. Robust Real-Time Face Detection. *International Journal of Computer Vision*, 57(2):137–154, 2004.

[29] C. L. Zitnick and P. Dollár. Edge Boxes: Locating Object Proposals from Edges. In *Proceedings of the European Conference on Computer Vision*, pages 391–405, Zurich, Switzerland, 2014.