A Deep Learning Pipeline for Semantic Facade Segmentation

Radwa Fathalla
http://www.aast.edu/cv.php?ser=36825

College of Computing and Information Technology
Arab Academy for Science and Technology
Alexandria, Egypt

George Vogiatzis
http://www.george-vogiatzis.org

School of Engineering and Applied Science
Aston University
Birmingham, UK

Abstract

We propose an algorithm that provides a pixel-wise classification of building facades. Building facades provide a rich environment for testing semantic segmentation techniques. They come in a variety of styles that reflect both appearance and layout characteristics. On the other hand, they exhibit a degree of stability in the arrangement of structures across different instances. We integrate appearance and layout cues in a single framework. The most likely label based on appearance is obtained through applying the state-of-the-art deep convolution networks. This is further optimized through Restricted Boltzmann Machines (RBM), applied on vertical and horizontal scanlines of facade models. Learning the probability distributions of the models via the RBMs is utilized in two settings. Firstly, we use them in learning from pre-seen facade samples, in the traditional training sense. Secondly, we learn from the test image at hand, in a way the allows the transfer of visual knowledge of the scene from correctly classified areas to others. Experimentally, we are on par with the reported performance results. However, we do not explicitly specify any hand-engineered features that are architectural scene dependent, nor do we include any dataset specific heuristics/thresholds.

1 Introduction

Facade parsing has attracted much research in recent years [2, 5, 16]. The problem has well known appearance challenges, such as occlusions, lighting/color variations, reflections, changes in architectural elements state (open and closed windows and doors), and the deterioration in building parts. On the other hand, most facades follow a Manhattan layout assumption, in addition to other persistent structural guidelines perceived by humans. These include the affinity of some structures to specific areas in the scene. Also, facades are characterized by the existence of repeated structures in grid-like arrangements. However, learning the guidelines and applying them to unseen data is made very difficult by the versatile nature of architectural designs.

An approach to facade splitting is employing a set of shape grammar rules specified manually by a human expert [15] or automatically learned [20]. Research has also been directed towards implementing a more flexible form of architectural guidelines. These guidelines are concerned with alignment, symmetry, similarity, co-occurrence and components layout. In [14], Martinovic *et al.* make use of these architectural principles in their final classification decision. They refine the output of a preceding pixel classification step by applying this set of restricting principles in an ad hoc procedure. Each principle is applied in isolation and in most part, as a matter of fulfilling a certain criterion is exceeding a manually specified threshold. Their work has been upgraded in [16]. While, maintaining the overall framework, they incorporate deformable part-based model detectors for object localization and perform max-margin learning of the CRF parameters in grid graphs. In addition, they refine the facade configuration through integer optimization. However, they resort once again to the heuristic weak architectural principles for post processing such as, rejecting a balcony hypothesis if it is not topped with a window and accepting running balconies only on specific floors of the building. In [11], Kozinski *et al.* specify a user-defined shape prior of grid form, in which they embed constraints of adjacency. They categorize the boundaries between structure pairs into 3 subtypes: straight, winding and irregular, and ones of containment in hierarchical form. Their final model is the one that acheives the minimal number of penalties over adjacency patterns, optimized through the Viterbi algorithm. The algorithm in [2], iteratively accesses the image, searching for specific structures in each iteration. Starting from the basic assumption that all pixels are wall ones, it then tries to replace this labeling with optimized local arrangements of window/balcony, roof/sky/chimney, and door/shop through dynamic programming runs. In [5] every pixel is represented by a vector of image features (such as: location, RGB values, and HOG features), in addition to contextual ones (such as: neighbourhood statistics, and bounding box features).Each feature vector is supplied independently to an ensemble of classifiers. It lacks the concurrency in classification of pixels of the arrangement and hence, it lacks the global optimality in the proper sense.

In this paper, we develop a pipeline which relies on 2 deep learning techniques, namely; Convolution Neural Network (CNN) [12] and RBM [6]. CNNs are acclaimed for achieving state-of-the-art results on the PASCAL VOC 2012 benchmark [1]. We utilize them in the classification module based on appearance qualities of the regions. The output is refined by another fine-grained classification module that uses RBMs to enforce contextual constraints. The difficulty in using context at pixel level classification to solve appearance ambiguities has been twofold. Firstly, representing the layout in a feature space that is more computationally efficient than the raw pixel space, while being able to preserve important scene characteristics. Secondly, the ability to assess the share of each pixel in the global layout, in a way that allows a local decision in fine grained vision tasks. Several efforts can be found in [19, 22]. Our use of the RBM is an attempt to tackle the aforementioned difficulties. We utilize its generative ability to restore the true structure of the scene. The algorithm maintains a global outlook while being able to fine-tune the final classifications at the pixel level. This is in contrast to the norm in the literature, which only refines classifications of preliminary whole structures. In addition, it builds its labeling on 2 models; the one based on experience from past data and a model of the captured layout of the scene at hand. This allows flexibility and extends the generalization ability of the trained machines.

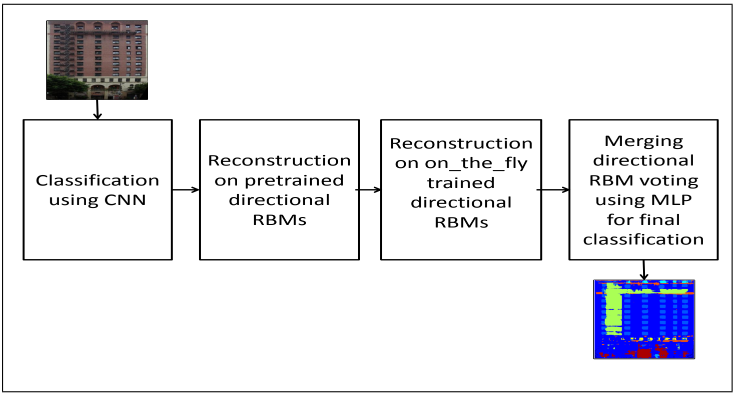
Classification using CNN
Reconstruction on pretrained directional RBMs
Reconstruction on on_the_fly trained directional RBMs
Merging directional RBM voting using MLP for final classification

Figure 1: A schematic showing system modules

2 Proposed Algorithm

The input to our algorithm is a RGB-valued facade image I. The aim is to provide an interpretation of the scene into meaningful architectural structures. Formally, the algorithm receives as input a set of image pixels $\mathbf{D} = \{\mathbf{d}_n\}_{n=1}^{N}$ in the $2d$ space. $N = r \times c$, where r and c are the vertical and horizontal dimensions of the image respectively. The algorithm classifies these data points by assigning them to a predefined set of labels $\mathcal{L} = \{L_m\}_{m=1}^{M}$, such that $\mathcal{L}$ holds indices to M architectural structures. We present a deep learning pipeline that utilizes both appearance and structural aspects of the scene. The core of our algorithm is the use of RBMs, to enforce architectural constraints deduced from past data and then to learn the structure of the test image at hand to allow it to make pixel labeling decisions based on the majority of its own pixels. The RBMs cascade is initially stimulated by predictions collected from a deep convolution network. Please refer to figure 1 for an overview of the algorithm.

2.1 Appearance cues

We have utilized the VGG-model [18] of a CNN, adapted to the task of pixel-wise classification [13]. The main features of the architecture is relying on small filter sizes that range from 1×1 to 3×3, to restrict the number of parameters to learn and prevent over-fitting. Also, there are deconvolution and interpolation layers to restore the original image size. In addition, lower level features and features from higher levels are combined through specialized *skip* layers and propagated ahead in the architecture. Learning the network parameters is based on minimizing the cross-entropy between the network outcome and ground truth.

The outcome of this phase is I^0 an appearance-based multinomial labeling map for the image, obtained through a softmax operation on the CNN classification scores.

2.2 Structure cues

Inspired by the work in [7], we opt for a generative probabilistic model for learning and enforcing architectural guidelines. We extend the RBM-based model from recognizing whole images to fine-grained recognition of image regions in a similar way to the Shape Boltzmann Machine (SBM) [4] and its extended version to multi-part objects (MSBM) [3]. However, our utilized visible nodes are binary instead of the SBM multinomial nodes. This is to preserve the original learning rule of the RBM formulation and retain its convergence properties [8]. In the following, we present a generic formulation of the utilized RBMs, throughout this work. The RBM consists of 2 sets of nodes; namely the visible **v** and hidden nodes **h**. The interconnections are symmetrical and the intra-connections are not allowed. The restriction imparts on the graphical model properties of tractability with respect to the calculated distributions. The model finds a joint probability distribution for **v** and **h** that can be represented as a Gibbs distribution of the form,

$$P(\mathbf{v},\mathbf{h};\theta) = \frac{1}{Z}\exp\left(-E\left(\mathbf{v},\mathbf{h};\theta\right)\right) \tag{1}$$

where, Z is the partition function

$$Z = \sum_{\mathbf{v}}\sum_{\mathbf{h}}\exp\left(-E\left(\mathbf{v},\mathbf{h};\theta\right)\right) \tag{2}$$

and E is the joint energy

$$E\left(\mathbf{v},\mathbf{h};\theta\right) = -\mathbf{h}^T W\mathbf{v} - \mathbf{a}^T\mathbf{v} - \mathbf{b}^T\mathbf{h} \tag{3}$$

W and $\{\mathbf{a},\mathbf{b}\}$ are the weight tensor and the set of visible and hidden biases, respectively. Collectively, they comprise the set of network parameters θ. During training, θ is optimized to maximize the log-likelihood of the visible data, by minimizing the discrepancy between this data and its reconstructed version. We use the FEPCD sampling technique introduced in [10] to approximate the derivative of the log probability of the training data, in the stochastic gradient descent process used to optimize θ. Similar to PCD [21], multiple persistent Markov chains are maintained to approximate the model-dependent expectations of the data. However, contrary to the PCD, there is a deterministic selection of the chains based on the free energy of the visible vector. The authors argue that the samples with the lower energy adhere better to the model's distribution, as they compute the likelihood gradient more accurately. Thus, the algorithm retains only half of the chains having the lowest energies.

In all settings, $v_i \in [0,1]$ and $h_j \in \{0,1\}$ and the conditional distributions defined over them are given by

$$p\left(v_i = 1|\mathbf{h}\right) = \sigma\left(a_i + \sum_i w_{ij}h_j\right) \tag{4}$$

$$p\left(h_j = 1|\mathbf{v}\right) = \sigma\left(b_j + \sum_i w_{ij}v_i\right) \tag{5}$$

where, $\sigma(\cdot)$ is the logistic function.

We use the probability $p_i = p\left(v_i = 1|\mathbf{h}\right)$ instead of sampling binary values. According to [8], this reduces the sampling noise and boosts the learning speed. In addition, the value p_i will be regarded as the likelihood of having a certain label L_m at some pixel.

2.3 Layout validation

We have trained 2 specialized RBMs: R^y for vertical and R^x horizontal scanlines. For each direction, the ground truth labeled image $G \in \Phi$ is resized to a fixed dimension (s_y for the vertical and s_x for the horizontal), while leaving the other dimension as a free parameter in

Algorithm 1 Structural Inference

```
Require: $I^0, R^y_\Phi, R^x_\Phi, B$
  $R^y_\Psi initialized, R^x_\Psi initialized$
  for each $k \in \{y, x\}$ do
    $I^0_k \leftarrow resize(I^0, s_k)$
    $\Gamma_{I^0_k} \leftarrow binarizeScanlines(I^0_k)$
    $\Gamma^\Phi_{I^0_k} \leftarrow reconstruct(R^k_\Phi, \Gamma_{I^0_k})$
  end for
  $T \leftarrow formMetaFeatures(\Gamma^\Phi_{I^0_y}, \Gamma^\Phi_{I^0_x})$
  $I^1 \leftarrow predict(B, T)$
  for each $k \in \{y, x\}$ do
    $I^1_k \leftarrow resize(I^1, s_k)$
    $\Omega_{I^1_k} \leftarrow binarizeScanlines(I^1_k)$
    $\hat{\Omega}_{I^1_k} \leftarrow augment(\Omega_{I^1_k})$
    $R^k_\Psi \leftarrow train(R^k_\Psi, \hat{\Omega}_{I^1_k})$
    $\Gamma^\Psi_{I^0_k} \leftarrow reconstruct(R^k_\Psi, \Gamma_{I^0_k})$
  end for
  $T \leftarrow formMetaFeatures(\Gamma^\Psi_{I^0_y}, \Gamma^\Psi_{I^0_x})$
  $I^2 \leftarrow predict(B, T)$
```

order to preserve the aspect ratio. We use nearest neighbour interpolation for the resizing. The result is two transformed images G_y and G_x. G_y produces $floor(s_y \cdot (c/r))$ scanlines of length s_y and G_y produces $floor(s_x \cdot (r/c))$ scanlines of length s_x, where r and c are the original height and width of the image. Basically, a scanline is a vector of pixel labels. The directional scanlines are accumulated into 2 training sets for the R^y_Φ and R^x_Φ RBMs.

Each scanline is binarized. This means, the q^{th} pixel on the scanline is represented by a one-hot mini-vector $o^q_{\mathcal{L}}$ with the value 1 at its label index. This is similar to the approach found in [23]. The mini-vectors of all pixels on a single scanline are aggregated in one flat vector. The visible data Ω_{G_y} and $\Omega_{G_x}$for the RBMs are the collections of these flat vectors in each direction. As such the machine, when trained on these scanlines, learns the associations between different labels at different aligned locations concurrently along columns and rows.

The approach of image subdivision can be found in [4]. It is normally carried out to keep the computation burden within tolerable limits and to escape severe resizing that might destroy the image layout. More importantly, it allows the RBM to focus on the highly stable pixel interactions which are mostly local. In our application, we opt for the scanline subdivisions as they hold the essence of architectural scene global semantics. They encode cues of structure order, neighbouring relations, equidistant repetitions, approximate location and alignment. In addition, as the training assumes independence between scanlines, we are implementing the weight sharing concept, commonly seen in CNNs, which achieves translation invariance. Thus, the location is no longer a coordinate value (even in normalized form) but rather a gestalt voting that emerges from the majority of pixels labels on the scanline. This tackles the scale-space difficulties encountered in location dependent approaches and boosts the algorithmic ability to deal with cropped images of facades. We regard our formulation based on scanlines as the first to tackle the problem of imposing layout constraints on scenes

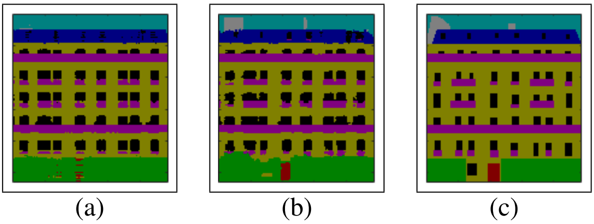

Figure 2: Final labeling of a sample facade image (a) without dataset augmentation, (b) with dataset augmentation. (c) The groundtruth map. It is clear that the door and chimney structures were correctly recovered in (b)

using RBM. In [3] the quadrant subdivision is confined to cropped images of single objects in focus (horse, face). The added complexity of scenes has multiple implications handled by scanlines. Severe resizing of the image in order to get the image quadrants to a size comparable to our scanlines, would devastate the layout of the building and render some structures unseen even by human eye. Generally, in context related applications, the challenge is using the largest scope of the image while maintaining the problem in reasonable size. Our formulation allows the decision at a single pixel to use knowledge from the whole span of the image width and height, thus achieving a higher degree of globalness and within tolerable computational burden. It encodes arrangement relations between more structures and allows repetition patterns to be more evident. Whereas, a quadrant subdivision will probably miss the symmetry of windows in a grid and miss encoding relations between structures inherent to different zones (sky and shop).

In our algorithm, we have 2 sources of learning. First, the R^y_Φ and R^x_Φ that build the posterior distribution from the training set, as explained. This is referred to as learning from seen data in the ground truth. The second source is the test image itself. This is achieved by RBMs R^y_Ψ and R^x_Ψ. We utilize the same machine architectures and the network parameters are initialized in the same manner. The only difference in the training is, the sets of directional scanlines are taken from a single layout validated test image, as will be shown. Learning from unseen data has been possible due to the existence of highly repetitive patterns in facades. Thus, the RBMs are able to construct the distribution upon the scanline parts that conform with the majority, while filtering out outliers.

Inference. The algorithm is presented in 1. At test time, we perform reconstructions of the CNN output on the trained RBMs. I^0 is subjected to a resizing into s_y and s_x dimensions by nearest neighbour interpolation. The resulting I^0_y and I^0_x images are divided into scanlines and the scanlines are binarized, as explained for G_y and G_x. These scanlines aggregated mini-vectors are clamped to the visible nodes of R^y_Φ and R^x_Φ. The reconstructions $\Gamma^\Phi_{I^0_k}$ in each direction k constitute of posteriors that reflect the visible node tendency to fire. And since the relative position of each label with respect to the pixel is fixed, the posteriors can easily be mapped to crisp predictions to get the multinomial maps I^1_k. For the n^{th}_k pixel, the assigned label

$$\hat{d}_{n_k} = \underset{L_m \in \mathcal{L}}{\arg\max}\, \bar{o}^{n_k}_{\mathcal{L}} \tag{6}$$

where $\bar{o}^{n_k}_{\mathcal{L}}$ is the reconstructed mini-vector for the pixel over all labels. I^1_k ($k \in \{y,x\}$) are merged into original image dimensions, resized again to fit each s_k and converted to binarized scanlines $\Omega_{I_y^1}$ and $\Omega_{I_x^1}$.

Algorithm 2 Augmentation procedure

```
Require: $\Omega_{I_y^1}, \Omega_{I_x^1}, \mathcal{L}$
  $\hat{\Omega}_{I_y^1} = \{\}$
  $\hat{\Omega}_{I_x^1} = \{\}$
  $dim_x \leftarrow countrows(\Omega^2_{I_x})$
  $dim_y \leftarrow countcols(\Omega^2_{I_y})$
  for each$L_m \in \mathcal{L}$ do
    for each$k \in \{y,x\}$ do
      $S \leftarrow getScanlinesofLabel(\Omega_{I_k^1}, L_m)$
      $t \leftarrow floor(dim_k/\|S\|)$
      $\hat{S} \leftarrow repeat(S,t)$
      $\hat{\Omega}_{I_k^1} \leftarrow append(\hat{\Omega}_{I_k^1}, \hat{S})$
    end for
  end for
```

Adaptation. R^y_Φ and R^x_Φ are used to enforce architectural guidelines. However, in the aforementioned procedure a rarely encountered facade scanline, but completely valid, is not guaranteed to have a low energy. This is due to the fact the RBM will place the probability mass on the frequently encountered scanlines. It will always produce the models that resemble what have been seen in its training set with sub-optimal adaptability to what is currently visualized in the image. Thus, we need an approach that increases the likelihood of certain labeled scanlines because of their abundance in the test image, regardless of their low frequency in the training ones. Hence, we train RBMs R^y_Ψ and R^x_Ψ on the validated test image scanlines. This time the RBMs will place the probability mass on what is frequent in the current image. In effect, we are extending the receptive field beyond the rows and columns to which the pixel directly belongs and propagating true labels from one area to another.

$\Omega_{I_k^1}$ are now augmented (see below) to form the training visible vectors $\hat{\Omega}_{I_k^1}$ for R^y_Ψ and R^x_Ψ. After learning the parameters of the RBMs, they are applied on the original scanlines obtained from the CNN output I_k^0 to produce directional posterior scores, which are then merged. These scoring maps are interpolated bilinearly to fit them in the r and c dimensions.

To sum up, the inference is dependent upon the correlation between the hypothesis perceived by appearance and the one suggested by common architectural layouts. That is to say, if a label is found at a certain pixel in I^0 and simultaneously in the set of putative structural labels, then it has a high chance of being assigned to the pixel. Otherwise, the most likely class will be propagated to it, based on the mass of scanlines that have a similar overall configuration in the test image.

Merging posteriors from directional RBMs. A final per-pixel decision can be taken by choosing the label with the maximum (maximum average) of the posteriors of both directions. However, we found a more sophisticated method based on higher-level reasoning on the RBM results for merging directional RBM outputs that boosted accuracies. After the training on R^k_Φ was done, we obtained the reconstructions of the training fold on R^k_Φ. A meta-feature vector was constructed per-pixel from the reconstructed posteriors in both directions, such that its length is $2 \cdot M$. The target classes are the true pixel labels obtained from the ground truth. We train B a backpropagation Multi-Layer Perceptron (MLP) with single hidden layer, on this data after massive sampling. At inference on the test fold, the meta-feature per-pixel is once again constructed but this time from the reconstructions on R^k_Ψ and ran on B to deduce the final labeling.

$\Omega_{I_k^1}$ **augmentation.** Imbalanced data is a widely recognized problem for classification techniques [9]. It makes a minority class more prone to be misclassified, due to its relative scarcity in the training set. In our experiments, we noticed that this was highly likely to affect small structures (such as door and chimney) when reconstructing the scanlines on R^y_Ψ and R^x_Ψ (figure 2). This is due to the fact a dataset of scanlines built from a single image would have such structures in extremely low counts. Interestingly, this did not occur when the scanlines were tested on R^y_Φ and R^x_Φ, despite being minority classes with the same ratio as in the test image. We realized that the representation power of a class is not only dependent on the relative count of its instances in the set, but can also be attributed to the size of the class in absolute terms. This phenomenon has been called the *lack of density* and its implication is explained in [17]. We believe the overfitting problem in very small training sets aggravates the imbalance, such that the machine has limited generalization ability not only beyond its training set, but even beyond the majority models in its training set.

We carry out an arbitrary augmentation procedure explained in algorithm 2. The procedure lead to increasing the count of small classes and achieving more balanced class-to-class ratios. However, it did not by any means accomplish priors equalization, because adding scanlines for one class inevitably increases other classes as well. Overfitting is an issue when learning from a single image. However, the objective in the first place is not boosting the generalization ability of the image-specific RBM as it will not be applied to unseen data but only reapplied on its training data. As such, we are exploiting the denoising ability of the RBM to conform outlier scanlines to the majority. Learning on RBMs proceeds in batches with an update of the set of parameters after the processing of each batch has ended. Scanlines with minority classes need to be represented in each batch inorder to prevent them from being filtered out as outliers and not contributing to the built conditional probability distribution of $\mathbf{v}$ and $\mathbf{h}$. For this reason, we need an augmentation phase such that scanlines with minority classes are cloned and added to the training set.

3 Evaluation

We tested our proposed algorithm on the *ECP-Monge* dataset [20] and the *CMP* dataset [24]. The *ECP-Monge* contains 104 rectified images of facades in Hausmannian style. There are 8 structures specified in the groundtruth maps. We use the corrected ground truth [14]. The *CMP* dataset is considered more challenging as it contains 378 samples with 12 structures from various (often difficult to model) styles.

In the experiments, s_y and s_x were unified for all images and set to 300 and 200, respectively. In all settings of RBMs, R^k_Φ and R^k_Ψ, the number of hidden nodes was set equal to the number of visible ones. Also, we trained all RBMs for 50 epochs. Our formulation based on the most stable layout representative, the scanline, allowed the RBMs to converge within this unprecedented small number of learning epochs. For the CNN, we retrained the VGG-model

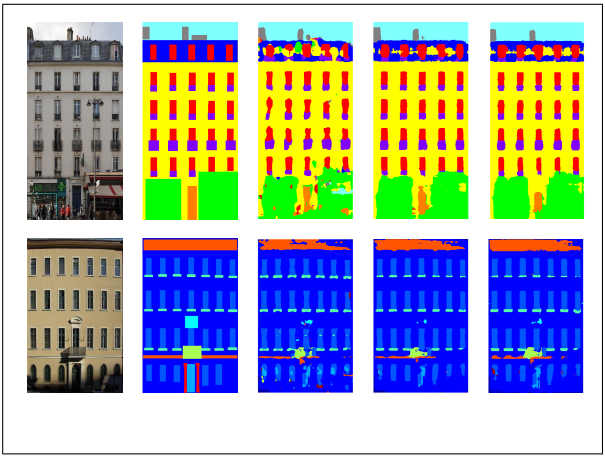

Figure 3: Sample results- *first column:* Original image; *second column:* Groundtruth map; *third column:* CNN output; *fourth column:* DPF-Φ output; *fifth column:* DPF-Ψ output.

on each dataset for 250 epochs, while maintaining the original parameters of learning rate, momentum and weight decay of the pretrained net.

As a performance measure, we utilize the pixel accuracy. It is calculated as follows; $TP/(TP+FN)$. True Positives TP and False negatives FN are determined overall the set of image pixels. We report our results based on 5-fold cross validation to ensure fair comparison between our algorithm Deep Facade Parsing (DPF-Ψ) and related work. We present the results in table 1.

[5] is the highest reported accuracy on the *ECP-Monge* dataset, to-date. It achieves 90.0% in the phase based on image features (equivalent to our appearance-based module) and 91.4% after structural improvement. Their results are reported for the older version of ground truth images based on 7 classes. The disregarded class is for the chimney structure. This kind of minority classes is often a bottleneck for the algorithms. In the literature, the lowest class accuracies were seen for doors and chimneys. Even in their own reported results, the door was the class of lowest accuracy. Thus, one can not be sure what the true overall accuracy will be, if the chimney structure was included as in our case. Same reasoning applies for [11].

The results show that we are on-par with state-of-the-art algorithms in terms of accuracy. More importantly, our algorithm highlights the benefit of context in image analysis. We report one of the highest accuracy gains, after inclusion of layout cues, defined as $A_2 - A_1$, without the dataset tailored refining rules found in [2, 16]. In figure 3, we display results of a selection of samples. Please, refer to the `supplementary materials` for the rest of the results.

As a self-test, we examine 2 variants of our algorithm to evaluate the different aspects proposed in its pipeline. These are:

- *variant 1-* Same as DFP-Ψ with the per-pixel classification obtained through maximizing the posterior based on the average of both directions to get the labeling. This is used to assess the goodness of the MLP as a merging criterion. For the *ECP-Monge* dataset the pixel accuracy was 90.87 and for the *CMP* dataset the figure was 67.70.

Dataset	*ECP-Monge*		*CMP*	
Method	A_1	A_2	A_1	A_2
[24]	59.60	84.20	33.20	60.30
[16]	84.75∗	88.02∗	-	-
[2]	86.71	90.34	-	-
[11]	90.10∗	91.30∗	-	-
[5]	90.80∗	91.40∗	66.20	68.10
DPF-Ψ	86.45	91.31	61.46	69.02

Table 1: Overall pixel accuracies based on appearance cues A_1 and when combined with layout cues A_2. The references marked with $*$ are included for completeness of results and are not suitable for direct comparison.

- *variant 2*- DFP-Φ which involves running the test image on R^y_Φ and R^x_Φ only, without the adaptation phase and its complementary augmentation. The results for *ECP-Monge* dataset and the *CMP* dataset were 90.49 and 65.54, respectively.

- *variant 3*- Same as DFP-Ψ but without augmenting the dataset for training on R^y_Ψ and R^x_Ψ. This variant was run only in a pilot experiment on the *ECP-Monge* dataset on 1 testing fold. There was no need to examine the rest of the folds because the pixel accuracy was consistently worse than I^1 (result of MLP B run on R^k_Φ outputs) yielding an accuracy of 87.63.

4 Conclusion

We have presented a pipeline for facade parsing. It relies on the state-of-the-art techniques of computer vision, and acheives on-par results with related research efforts. We do not include any ad hoc post-processing steps, nor do we manually specify any architecture-based features. The pipeline is initialized with classifications from the VGG-16 convolution model customized to semantic segmentation. The results are further improved through a probabilistic shape prior captured by trained RBMs. We present a novel idea to learn from test images, to increase the generalization ability of the algorithm. We illustrate the importance of dataset augmentation for severely small imbalanced datasets, resulting from a single test image. Our future work is targeted at experimenting with deeper architectures of Boltzmann machines. In addition, we intend to adapt our pipeline to images of articulated objects. Formulations such as ours will help boost the RBM as a powerful tool in structural modeling beyond single object of focus.

References

[1] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *CoRR*, abs/1606.0, 2016. URL http://arxiv.org/abs/1606.00915.

[2] Andrea Cohen, Alexander G Schwing, and Marc Pollefeys. Efficient Structured Parsing of Facades Using Dynamic Programming. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.

[3] S Eslami and Christopher Williams. A Generative Model for Parts-based Object Segmentation. In F Pereira, C J C Burges, L Bottou, and K Q Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 100–107. Curran Associates, Inc., 2012. URL http://papers.nips.cc/paper/4774-a-generative-model-for-parts-based-object-segmentation.pdf.

[4] S. M. Eslami, Nicolas Heess, Christopher K. Williams, and John Winn. The shape boltzmann machine: A strong model of object shape. *Int. J. Comput. Vision*, 107(2): 155–176, April 2014. ISSN 0920-5691. doi: 10.1007/s11263-013-0669-1. URL http://dx.doi.org/10.1007/s11263-013-0669-1.

[5] Raghudeep Gadde, Varun Jampani, Renaud Marlet, and Peter V Gehler. Efficient 2D and 3D Facade Segmentation using Auto-Context. *CoRR*, abs/1606.0, 2016. URL http://arxiv.org/abs/1606.06437.

[6] G E Hinton and R R Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507, July 2006. doi: 10.1126/science.1127647. URL http://www.ncbi.nlm.nih.gov/sites/entrez?db=pubmed&uid=16873662&cmd=showdetailview&indexed=google.

[7] Geoffrey E Hinton. To Recognize Shapes First Learn to Generate Images. In *Computational Neuroscience: Theoretical Insights into Brain Function*. Elsevier, 2007.

[8] Geoffrey E Hinton. *A Practical Guide to Training Restricted Boltzmann Machines*, pages 599–619. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012. ISBN 978-3-642-35289-8. doi: 10.1007/978-3-642-35289-8_32. URL http://dx.doi.org/10.1007/978-3-642-35289-8_32.

[9] Chen Huang, Yining Li, Chen Change Loy, and Xiaoou Tang. Learning Deep Representation for Imbalanced Classification. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[10] Mohammad Ali Keyvanrad and Mohammad Mehdi Homayounpour. Deep Belief Network Training Improvement Using Elite Samples Minimizing Free Energy. *CoRR*, abs/1411.4, 2014. URL http://arxiv.org/abs/1411.4046.

[11] Mateusz Kozinski, Raghudeep Gadde, Sergey Zagoruyko, Guillaume Obozinski, and Renaud Marlet. A MRF Shape Prior for Facade Parsing With Occlusions. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[12] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-Based Learning Applied to Document Recognition. In *Proceedings of the IEEE*, volume 86, pages 2278–2324, 1998. URL http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.42.7665.

[13] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully Convolutional Networks for Semantic Segmentation. *CoRR*, abs/1411.4, 2014. URL http://arxiv.org/abs/1411.4038.

[14] Andelo Martinović, Markus Mathias, Julien Weissenberg, and Luc Van Gool. A Three-Layered Approach to Facade Parsing. In Andrew Fitzgibbon, Svetlana Lazebnik, Pietro Perona, Yoichi Sato, and Cordelia Schmid, editors, *Computer Vision – ECCV 2012*, volume 7578 of *Lecture Notes in Computer Science*, pages 416–429. Springer, 2012. ISBN 978-3-642-33785-7.

[15] Markus Mathias, Andelo Martinović, Julien Weissenberg, and Luc Van Gool. Procedural 3D Building Reconstruction Using Shape Grammars and Detectors. In *3DIMPVT*, pages 304–311, 2011.

[16] Markus Mathias, Andjelo Martinović, and Luc Van Gool. ATLAS: A Three-Layered Approach to Facade Parsing. *International Journal of Computer Vision*, 118(1):22–48, 2016. ISSN 1573-1405. doi: 10.1007/s11263-015-0868-z. URL http://dx.doi.org/10.1007/s11263-015-0868-z.

[17] S J Raudys and A K Jain. Small sample size effects in statistical pattern recognition: recommendations for practitioners. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(3):252–264, March 1991. ISSN 0162-8828. doi: 10.1109/34.75512.

[18] Karen Simonyan and Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. *CoRR*, abs/1409.1, 2014. URL http://arxiv.org/abs/1409.1556.

[19] Fatemeh Tabib Mahmoudi, Farhad Samadzadegan, and Peter Reinartz. Context Aware Modification on the Object Based Image Analysis. *Journal of the Indian Society of Remote Sensing*, 43(4):709–717, 2015. ISSN 0974-3006. doi: 10.1007/s12524-015-0453-5. URL http://dx.doi.org/10.1007/s12524-015-0453-5.

[20] Olivier Teboul, Iasonas Kokkinos, Loïc Simon, Panagiotis Koutsourakis, and Nikos Paragios. Shape grammar parsing via Reinforcement Learning. In *CVPR*, pages 2273–2280. IEEE Computer Society, 2011. ISBN 978-1-4577-0394-2. URL http://dblp.uni-trier.de/db/conf/cvpr/cvpr2011.html#TeboulKSKP11.

[21] Tijmen Tieleman. Training Restricted Boltzmann Machines Using Approximations to the Likelihood Gradient. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, pages 1064–1071, New York, NY, USA, 2008. ACM. ISBN 978-1-60558-205-4. doi: 10.1145/1390156.1390290. URL http://doi.acm.org/10.1145/1390156.1390290.

[22] Antonio Torralba. Contextual Priming for Object Detection. *International Journal of Computer Vision*, 53(2):169–191, 2003. ISSN 1573-1405. doi: 10.1023/A:1023052124951. URL http://dx.doi.org/10.1023/A:1023052124951.

[23] Stavros Tsogkas, Iasonas Kokkinos, George Papandreou, and Andrea Vedaldi. Semantic Part Segmentation with Deep Learning. *CoRR*, abs/1505.0, 2015. URL http://arxiv.org/abs/1505.02438.

[24] Radim Tyleček and Radim Šára. *Spatial Pattern Templates for Recognition of Objects with Regular Structure*, pages 364–374. Springer Berlin Heidelberg, Berlin, Heidelberg, 2013. ISBN 978-3-642-40602-7. doi: 10.1007/978-3-642-40602-7_39. URL http://dx.doi.org/10.1007/978-3-642-40602-7_39.