A Three-Layered Approach to Facade Parsing

Anđelo Martinović[1] Markus Mathias[1]
Julien Weissenberg[2] Luc Van Gool [1,2]

[1]ESAT-PSI/VISICS, KU Leuven

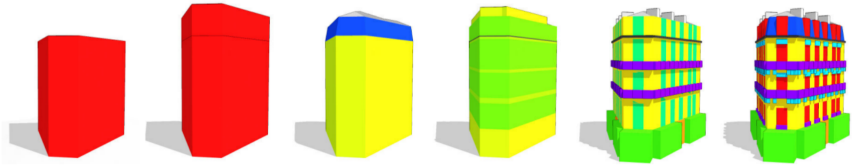
[2]Computer Vision Laboratory, ETH Zurich

We aim to improve the state of the art in facade parsing

From an image ... ... to its labeling

We do not use shape grammars!

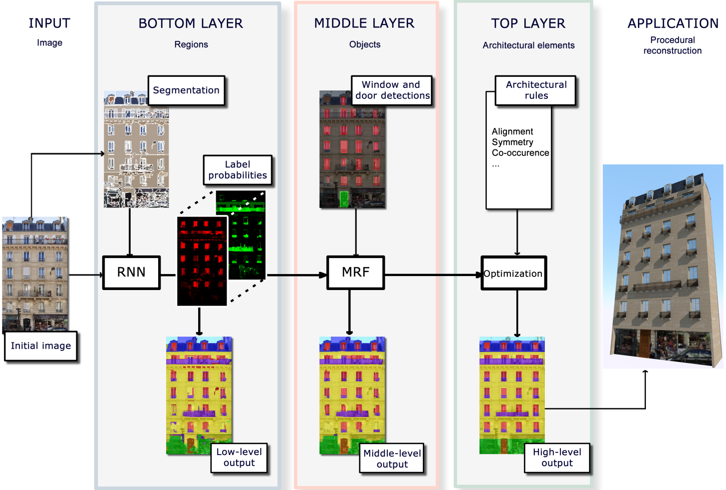
- State-of-the-art methods in facade parsing assume that an appropriate shape grammar is available [1].



- We do not use shape grammars as priors, and still achieve superior performance.

[1] Teboul, Kokkinos, Simon, Koutsourakis, Paragios: "Shape grammar parsing via Reinforcement Learning", CVPR, (2011).

A Three-Layered Approach



Bottom layer - segments

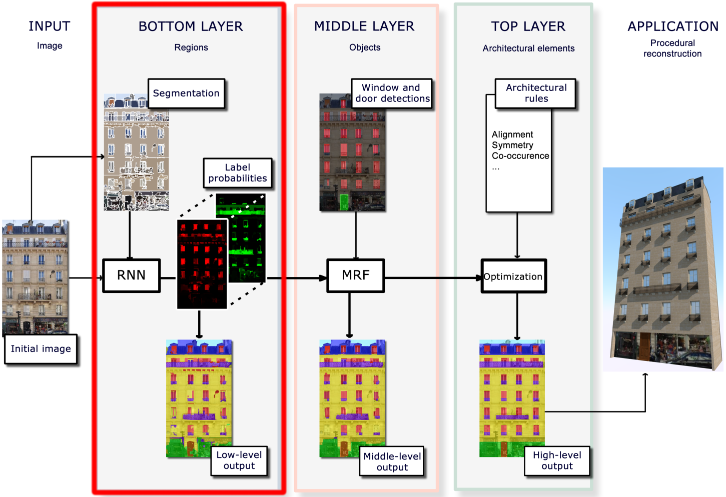
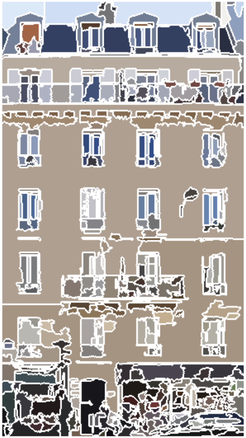
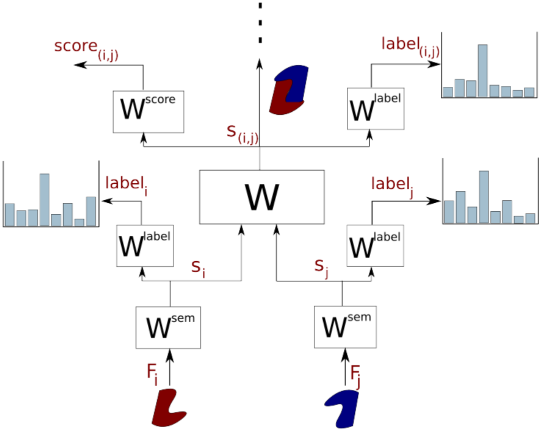


Image preparation

- We segment the image using mean-shift.
- The appearance (color and texture), geometry, and location features are extracted for each region.
  - STAIR Vision Library
- This results in 225-dimensional feature vectors.



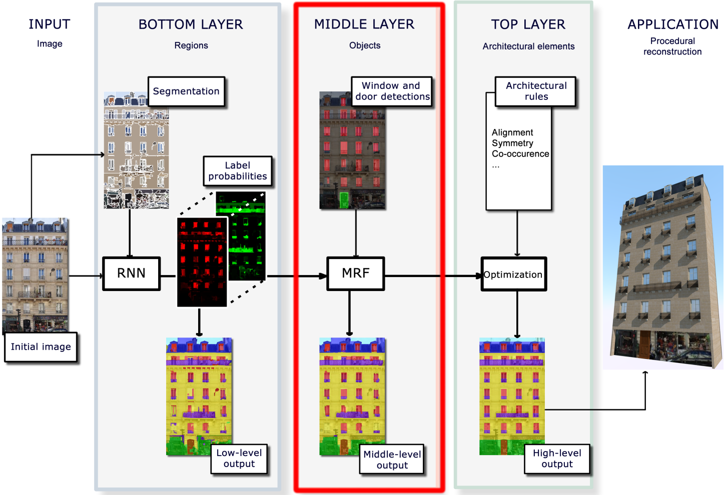
Recursive Neural Network



[6] Socher et al., "Parsing Natural Scenes and Natural Language with Recursive Neural Networks", ICML (2011).

Bottom Layer Output

Middle layer - objects



Window and Door Detection

Incorporating Detector Knowledge With MRFs

Energy minimization with graph cuts

- Potts model

$$E(l) = \sum_{x_i} \phi_s(l_i \mid x_i) + \lambda \sum_{x_i} \sum_{x_j \sim x_i} \phi_p(l_i, l_j \mid x_i, x_j) \quad (1)$$

- Pairwise potentials

$$\phi_p(l_i, l_j \mid x_i, x_j) = \begin{cases} 0, & \text{if } l_i = l_j \\ 1, & \text{otherwise} \end{cases} \quad (2)$$

- Unary potentials

$$\phi_s(l_i \mid x_i) = -\log p(l_i \mid RNN(x_i)) - \sum_k \alpha_k \log p(l_i \mid D_k(x_i)) \quad (3)$$

Incorporating Detector Knowledge With MRFs

Energy minimization with graph cuts

- Potts model

$$E(l) = \sum_{x_i} \phi_s(l_i \mid x_i) + \lambda \sum_{x_i} \sum_{x_j \sim x_i} \phi_p(l_i, l_j \mid x_i, x_j) \quad (1)$$

- Pairwise potentials

$$\phi_p(l_i, l_j \mid x_i, x_j) = \begin{cases} 0, & \text{if } l_i = l_j \\ 1, & \text{otherwise} \end{cases} \quad (2)$$

- Unary potentials

$$\phi_s(l_i \mid x_i) = -\log p(l_i \mid RNN(x_i)) - \sum_k \alpha_k \log p(l_i \mid D_k(x_i)) \quad (3)$$

Incorporating Detector Knowledge With MRFs

Energy minimization with graph cuts

- Potts model

$$E(l) = \sum_{x_i} \phi_s(l_i \mid x_i) + \lambda \sum_{x_i} \sum_{x_j \sim x_i} \phi_p(l_i, l_j \mid x_i, x_j) \tag{1}$$

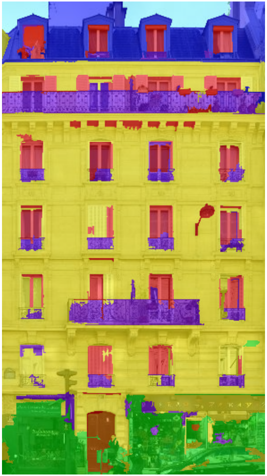
- Pairwise potentials

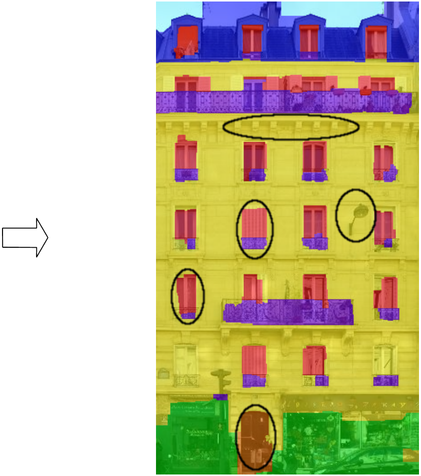
$$\phi_p(l_i, l_j \mid x_i, x_j) = \begin{cases} 0, & \text{if } l_i = l_j \\ 1, & \text{otherwise} \end{cases} \tag{2}$$

- Unary potentials

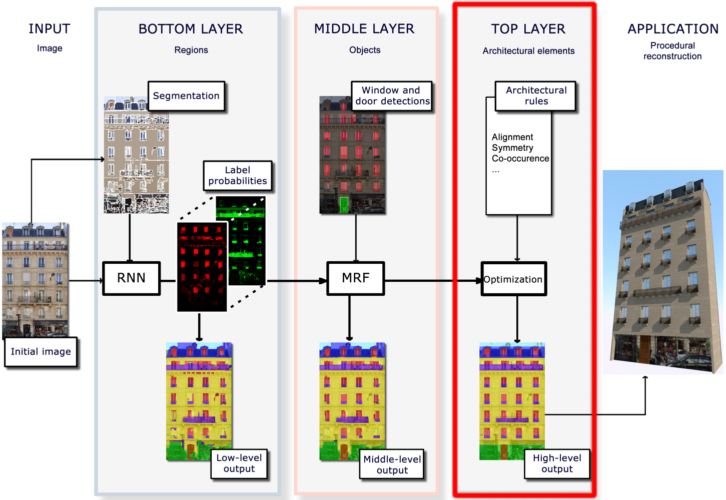
$$\phi_s(l_i \mid x_i) = -\log p(l_i \mid RNN(x_i)) - \sum_k \alpha_k \log p(l_i \mid D_k(x_i)) \tag{3}$$

From Bottom To Middle Layer Output





Top layer - architectural elements

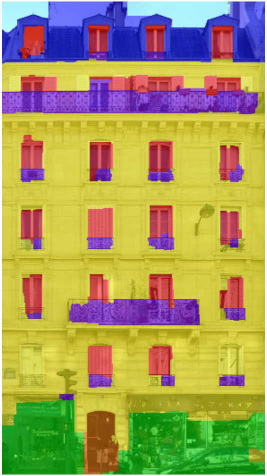


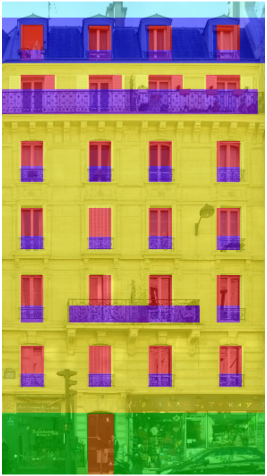
Weak Architectural Principles

- Soft constraints instead of fixed grammar structure
- Only enforced if there is enough image support

Principle	Alter	Add	Remove
Vertical and horizontal (non)alignment	✓	-	-
Window similarity	-	✓	-
Facade symmetry	-	✓	✓
Element co-occurence	-	✓	✓
Equal width/height in a row or column	✓	-	-
Door hypothesis	✓	✓	✓
Vertical region order	✓	-	-

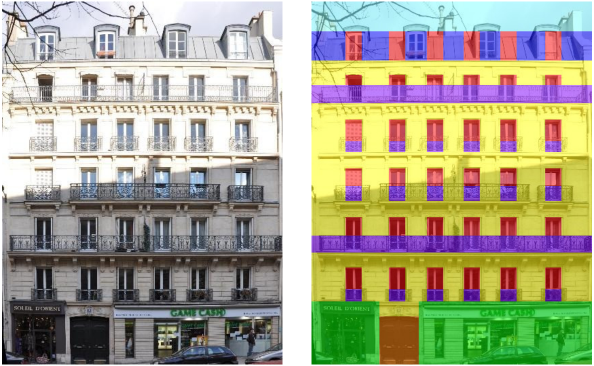
From Middle To Top Layer Output





Ecole Centrale Paris Facades Database [2]

- Contains 104 rectified and cropped Haussmannian facades.



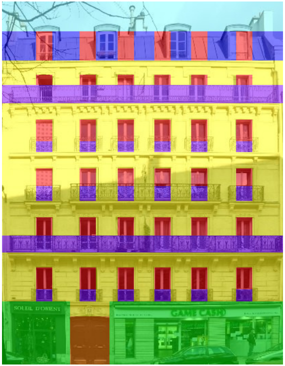
window | wall | balcony | door | roof | sky | shop

[2] Teboul, O. , "Ecole Centrale Paris Facades Database" (2010).

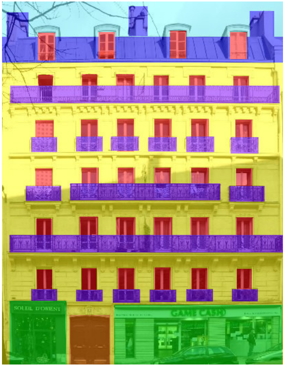
Ecole Centrale Paris Facades Database

- Original labeling is plausible, but imprecise.
- We provide more precise annotations (available online).

Old annotation



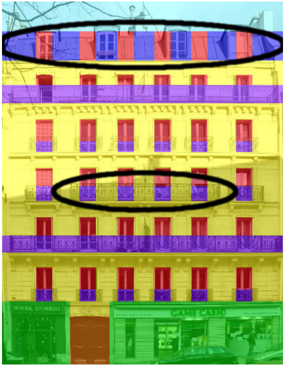
New annotation



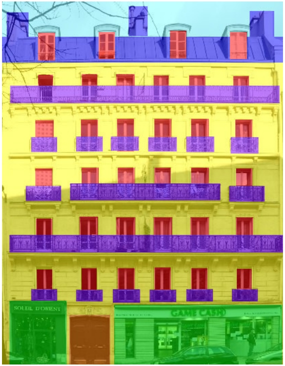
Ecole Centrale Paris Facades Database

- Original labeling is plausible, but imprecise.
- We provide more precise annotations (available online).

Old annotation



New annotation



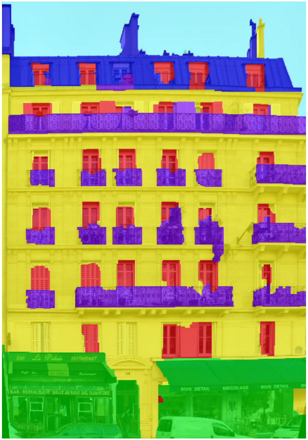
Results - ECP Dataset

Class	Baseline[4]	Layer 1	Layer 2	Layer 3
window	62	62	69	**75**
wall	82	91	**93**	88
balcony	58	**74**	71	70
door	47	43	60	**67**
roof	66	70	73	**74**
sky	95	91	91	**97**
shop	88	79	86	**93**
Pixel acc.	74.71	82.63	**85.06**	84.17

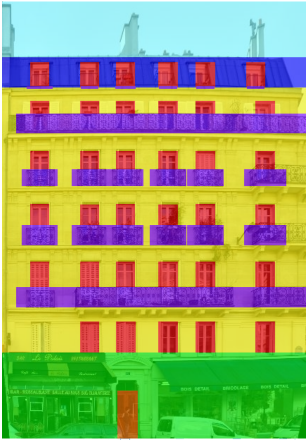
[4] Teboul, O., "Shape Grammar Parsing: Application to Image-based Modeling" (2011).

Pixel Accuracy vs Visual Effect

Pixel accuracy: 89.48%



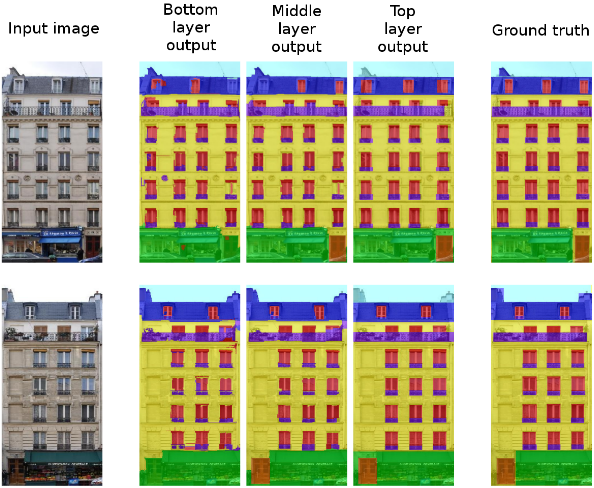
Pixel accuracy: 87.82%



Results - ECP Dataset

Class	Baseline[4]	Layer 1	Layer 2	Layer 3
window	62	62	69	**75**
wall	82	91	**93**	88
balcony	58	**74**	71	70
door	47	43	60	**67**
roof	66	70	73	**74**
sky	95	91	91	**97**
shop	88	79	86	**93**
Pixel acc.	74.71	82.63	**85.06**	84.17
Class acc.	71.14	72.86	77.46	**80.71**

Example Outputs - ECP Dataset



eTRIMS Database [3]

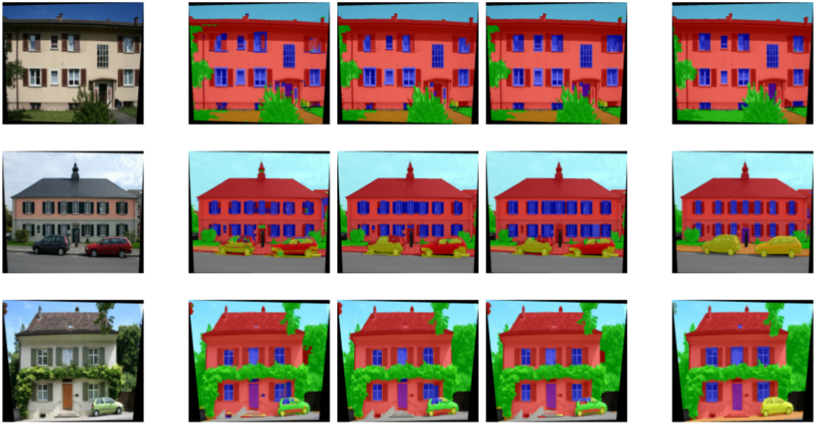
- Contains 60 images of various building styles.
- We perform automatic rectification.

building car door pavement road sky vegetation window

[3] Korč, F. and Förstner, W., "eTRIMS Image Database for Interpreting Images of Man-Made Scenes" (2009).

Example Outputs - eTRIMS Dataset

Input image | Bottom layer output | Middle layer output | Top layer output | Ground truth



Example Outputs - Procedural Models

Summary

- We developed a novel three-layer approach for facade parsing.
- We significantly outperform the state-of-the-art on two facade parsing datasets.
- We utilize the concept of weak architectural knowledge.

- Outlook
  - So far, the inferred procedural models are instance-specific.
  - We want to generalize between buildings of the same style.
  - As we no longer depend on grammars as priors, can we instead induce them from the data?

Questions?

Anđelo Martinović

http://homes.esat.kuleuven.be/~amartino/

Available online: updated ECP annotations, paper manuscript, supplementary material, spotlight video

References

- [1] Teboul, O. and Kokkinos, I. and Simon, L. and Koutsourakis, P. and Paragios, N. , "Shape grammar parsing via Reinforcement Learning" (2011).
- [2] Teboul, O. , "Ecole Centrale Paris Facades Database" (2010).
- [3] Korč, F. and Förstner, W., "eTRIMS Image Database for Interpreting Images of Man-Made Scenes" (2009).
- [4] Teboul, O., "Shape Grammar Parsing: Application to Image-based Modeling" (2011).
- [5] Yang, M.Y. and Förstner, W. , "Regionwise Classification of Building Facade Images", Springer (2011).
- [6]Socher et al. , "Parsing Natural Scenes and Natural Language with Recursive Neural Networks", ICML (2011).

Results - eTRIMS Dataset

The results for eTrims were obtained by automatically rectifying both the input images and the ground truth labelings. Our results were computed in the rectified space. As previous work did not perform any rectification, we repeated the evaluation by "unrectifying" our output labeling and comparing to the original ground truth. The results obtained in this way are actually better by ~1% than reported in the paper.

Class	Baseline[5]	Layer 1	Layer 2	Layer 3
building	71	88	**91**	87
car	35	**69**	**69**	**69**
door	16	**25**	18	19
pavement	22	**34**	33	**34**
road	35	**56**	55	**56**
sky	78	**94**	93	**94**
vegetation	66	**89**	**89**	88
window	75	71	74	**79**
Pixel acc.	65.8	81.87	**83.16**	81.63
Class acc.	49.75	**65.85**	65.4	65.6