Spatial Pattern Templates for Recognition of Objects with Regular Structure

Radim Tyleček and Radim Šára

Center for Machine Perception
Czech Technical University in Prague

Abstract. We propose a method for semantic parsing of images with regular structure. The structured objects are modeled in a densely connected CRF. The paper describes how to embody specific spatial relations in a representation called *Spatial Pattern Templates* (SPT), which allows us to capture regularity constraints of alignment and equal spacing in pairwise and ternary potentials.
Assuming the input image is pre-segmented to salient regions the SPT describe which segments could interact in the structured graphical model. The model parameters are learnt to describe the formal language of semantic labelings. Given an input image, a consistent labeling over its segments linked in the CRF is recognized as a word from this language. The CRF framework allows us to apply efficient algorithms for both recognition and learning. We demonstrate the approach on the problem of facade image parsing and show that results comparable with state of the art methods are achieved without introducing additional manually designed detectors for specific terminal objects.

1 Introduction

The recent development in the areas of object detection and image segmentation is centered around the incorporation of contextual cues. Published results confirm the hypothesis that modeling relations between neighboring pixels or segments (superpixels) can significantly improve recognition accuracy for structured data. The first choice one has to make here is to choose the neighbor relation, or in other words, which primitive elements participate in constraints on labels. The constraints are usually specified with a formal language of spatial arrangements. A common choice for the relation is the adjacency of element pairs in the image plane, such as 4 or 8-neighborhood of pixels in a grid, which supports the language model [1]. This can be extended in various directions: In 'depth' when we have more concurrent segmentations, or in cardinality when we connect more elements together. Generally speaking, in this paper we will take a closer look on this design process and introduce a concept called *Spatial Pattern Templates* (SPT).

A convenient framework to embed such patterns into are probabilistic graphical models, where image elements correspond to nodes and edges (or higher-order cliques) to the relations among them. In such a graph, our pattern templates

correspond to cliques or factors, as they describe how a given joint probability factorizes. We choose Conditional Random Fields (CRF [9]) as a suitable model, which allows us to concentrate on the element relations and not to care much about how the data are generated. Specifically, we propose pattern templates to deal with regular segmentations and call them *Aligned Pairs (AP)* and *Regular Triplets (RT)*.

We identify regular segmentations as those where object geometry, shape or appearance exhibit symmetry, particularly translational, which manifests in alignment and similarity. Such principles often apply to images with man-made objects, even though such phenomena are also common in the nature. Urban scenes have some of the most regular yet variable segmentations and their semantic analysis is receiving more attention nowadays, as it can aid other computer vision tasks such as image-based urban reconstruction. We design our method with this application in mind, specifically targeting parsing of facade images (a multi-class labeling problem).

In this task, we exploit the properties of largely orthogonal facade images. We start by training a classifier to recognize the patches given by unsupervised segmentation. Based on the initial segments we build a CRF with binary relative location potentials on *AP* and ternary label consistency potentials on *RT*. For intuition, this can be seen as a process where all segments jointly vote for terminal labels of the other segments, with voting scheme following the chosen spatial patterns. The concept of template design, its embedding in the CRF and implementation for regular objects with *Regular Triplets* and *Aligned Pairs* are the contributions of this paper.

2 Related work

Contextual models. Relative location prior on label pairs is used in [4] for multi-class segmentation. Every segment votes for the label of all other segments based on their relative location and classifier output. Ideally, such interactions should be modeled with a complete graph CRF, where an edge expresses the joint probability of the two labels given their relative location, but this would soon make the inference intractable with the growing number of segments. Instead Gould et al. [4] resort to a voting scheme and use CRF with pairwise terms for directly adjacent segments only. In our approach, we include the discretized relative location prior in a CRF but limit the number of interactions by choosing a suitable pattern template.

CRFs are popular for high-level vision tasks also thanks to the number of algorithms available for inference and learning [11]. However, useful exact algorithms are known only for a specific class of potential functions (obeying *submodularity*). Kohli et al. [5] fit in this limitation with a robust version of a generalized Potts model, which softly enforces label consistency among any number of elements in a high order clique (pixels in segments). We can use this model for *RT*, but because the pairwise relative location potentials may have arbitrary form, we cannot apply the efficient α-expansion optimization used in [5].

Structure learning. A number of methods for learning general structures on graphs have been recently developed [3, 12, 13]. They learn edge-specific weights in a fully connected graph, which is directly tractable only when the number of nodes n is small (10 segments and 4 spatial relations in [3]) due to edge number growing with $\mathcal{O}(n^2)$. Scalability of the approach has been extended by Schmidt et al. by block-wise regularization for sparsity [12] (16 segments) and subsequently also for higher-order potentials with a hierarchical constraint [13] (30 segments). Since we deal with ~ 500 segments, this approach cannot be directly applied and, as suggested in [12], a restriction on the edge set has to be considered. The SPT can be here seen as a principled implementation of this restriction to keep the problem tractable.

Facade parsing. There are two major approaches to the facade parsing problem. Top-down approach relies on the construction of a generative rule set, usually a grammar, and the result is obtained stochastically as a word in the language best matching the input image [14]. So far there are no methods for automatic construction of the grammars and they do not generalize well outside of the style they were designed for. Learning is possible for very simple grammars (e.g. grid [17]) but it cannot express other structural relations.

Bottom-up approaches instead combine weak general principles, which are more flexible and their parameters can be learned. Regularity of spacing, shape similarity and alignment is used in [18] to find weakly regular structures, but the model cannot be simply extended for more classes than one (*window*). The hierarchical CRF [8], which aggregates information from multiple segmentations at different scales, has been applied to facades in [19], where binary potentials model consistency of adjacent labels within as well as across segmentations. Here neighboring segments with similar appearance are more likely to have the same label (contrast-sensitive Potts model). The recent three-staged method [10] combines local and object detectors with a binary Potts CRF on pixels. The result is further sequentially processed to adjust the labels according to the alignment, similarity, symmetry and co-occurence principles, each of them applied with a rather heuristic procedure. Additional principles are designed for a specific dataset and in fact resemble grammatical rules. In contrast, our method accommodates the general assumption of regularity in a principled and general way as a part of the model, which is based on the CRF and can benefit from the joint learning and inference.

3 Spatial Pattern Template model

Initially we obtain a set of segments S in the input image with a generic method such as [2], tuned to produce over-segmentation of the ground truth objects. The image parsing task is to assign labels $L = \{l_i \in C\}_{i=1}^N$ of known classes $C = \{c_j\}_{j=1}^K$ to given segments $S = \{s_i \subset \operatorname{dom} I\}_{i=1}^N$ in an image I. Let $X = \{x_i \subset I\}_{i=1}^N$ be the image data of segments S. With segments corresponding to nodes in a graph and labels L being the node variables, we construct

a CRF with potentials taking the general form of

$$p(L|X,S) = \frac{1}{Z(X,S)} \prod_{q \in \mathcal{Q}} e^{-\sum_{j \in \phi(q)} \theta_j \varphi_j(\mathbf{l}_q | \mathbf{x}_q, \mathbf{s}_q)}, \tag{1}$$

where $\mathcal{Q}$ is the set of cliques, φ_j are potential functions from a predefined set $\phi(q)$ defined for a clique q. The φ_j is a function of all node variables in collections $\mathbf{l}_q, \mathbf{x}_q, \mathbf{s}_q$, their weights are θ_j and $Z(X,S)$ is the normalizing partition function. The design of a specific CRF model now lies in the choice of cliques $\mathcal{Q}$ defining **topology** on top of the segments and their potential functions φ_j, which act on all node variables in the clique and set up the **probabilistic model**.

3.1 Spatial templates for data-dependent topology

As a generalization of the *adjacency*, used i.e. in [19], we can think of other choices for the graph topology that may suit our domain by including interactions between distant image elements, which are 'close' to each other in a different sense. As mentioned in Sec. 2, the scale of the problem does not allow us to reach complete connectivity. To allow dense connectivity while keeping the problem tractable, we need to restrict the number of cliques (edges). We describe this restriction with a *template* and, with the geometrical context in mind, we limit ourselves to *spatial* templates, which assign segments to cliques based on their geometrical attributes (shape, location). In principle other attributes (appearance) could be used in the template too. The meaning of this representation is to provide a systematic procedure for automatic learning of which interactions are the most efficient ones for the recognition task at hand.

In order to describe the process of designing a complex data dependent topology for a CRF, we first have to decompose the process behind clique template design into individual steps:

1. The first step is the specification of **core attribute relation functions** $\delta_i : A^n \to \mathbb{R}$ based on the domain knowledge. The relations act on easily measurable attributes A of n-tuples of segments. *Example: Positions of two points in a plane as attributes $A_x, A_y \in \mathbb{R}^2$ and their signed distances in directions x and y as the relations δ_x, δ_y.*
2. The ranges of relations δ_i are **discretized** to ordered sets Δ_i and $d_i : A^n \to \Delta_i$ becomes the discrete counterpart of function δ_i. *Example: The signed distance is divided into three intervals, $\Delta_x = \{\text{left}, \text{equal}, \text{right}\}$, $\Delta_y = \{\text{below}, \text{equal}, \text{above}\}$.*
3. In the next step the Cartesian product of m relation ranges Δ_i gives domain $D = \Delta_1 \times \cdots \times \Delta_m$, where subsets define logical **meta relations** (*and, or, equal*). *Example: Three intervals on two axes give 3^2 combinations in $D_{xy} = \Delta_x \times \Delta_y$.*
4. The **spatial template** is a subset $\Omega \subset D$ representing a concrete relation. The template is specified by an indicator function $\omega : D \to \{0,1\}$ representing the allowed combinations. *Example: For alignment in one direction we set $\omega_{xy} = 1$ when $d_x = \text{equal}$ or $d_y = \text{equal}$, otherwise $\omega_{xy} = 0$.*

The template design may be viewed as a kind of declarative programming framework for model design, a representation that can incorporate the specific knowledge in a generic way with combinations of core relations δ_i. Each spatial template is related to one potential function φ_j in (1).

In summary, the result of this process describes which subsets of segments S labeled L should be jointly modeled in a graphical model; which of these are effective is subject to learning. Figure 2 shows how the segments correspond to nodes and their subsets define factors in $p(L|X, S)$. In this work we introduce two templates suitable for regular segmentations.

Aligned Pairs (AP) Out of all pairs of segments u, v we choose those which are aligned either vertically or horizontally. It is useful to connect segments not directly adjacent when the labels in such pairs follow some pattern, i.e. windows are aligned with some free space in between.

The specification follows the spatial template design steps: 1) Based on the position attribute we choose horizontal and vertical **alignment** δ_h, δ_v with $\delta_h : (s_u, s_v) \to \mathbb{R}$ and $\delta_h = 0$ when the segments are exactly aligned, otherwise according to Fig. 1 (analogically δ_v for vertical). 2) Quantized d_h, d_v take values from Δ_a according to Fig. 3 evaluated on segment bounding boxes. 3) Combinations of horizontal and vertical alignment are then represented by joining d_h, d_v in a discrete domain $D_{AP} = \Delta_a^2$ limited by maximum distance. 4) Finally we specify the AP template with $\omega_{AP} = 1$ in the blue region in Fig. 1.

Note that *adjacency* (4-neighborhood) is a special case of AP when we specify $\omega_{AP} = 1$ only for four specific values in D_{AP} (directly above/under/left/right, red squares in Fig. 1). Similarly values of $|d_h| \leq 5$ together with $|d_v| \leq 5$ correspond to *overlap* or *nesting* of segments.

Regular Triplets (RT) Here we combine two *Aligned Pairs* in a triplet u, v, w with regular spacing, in which the v is the shared segment. Including triplets allows to express a basis for repetitive structures (rows, columns) of primitive objects of the same label (*window*, *balcony*).

1) In addition to position alignment δ_h, δ_v we introduce ternary relation functions for **size similarity** $\delta_s : (s_u, s_v, s_w) \to \mathbb{R}$ (relative difference in size of segments) and **regular spacing** $\delta_r : (s_u, s_v, s_w) \to \mathbb{R}$ (relative difference in free space between segments). 2) Based on them we define binary function $d_s : (s_u, s_v, s_w) \to \{0, 1\}$ to be 1 when $|\delta_s| < 0.1$ and similarly $d_r : (s_u, s_v, s_w) \to \{0, 1\}$ to be 1 when $|\delta_r| < 0.1$. 3) All functions $d_h(s_u, s_v)$, $d_v(s_u, s_v)$, $d_h(s_v, s_w)$, $d_v(s_v, s_w)$, $d_s(s_u, s_v, s_w)$ and $d_r(s_u, s_v, s_w)$ are then joined in a six-dimensional domain $D_{RT} = \Delta_a^4 \times \{0, 1\}^2$. 4) Finally we specify $\omega_{RT} = 1$ in the subspace of D_{RT} where $d_s = 1, d_r = 1$ and values of d_h, d_v indicate that the three segments are pair-wise aligned in the same direction (horizontal or vertical).

3.2 Probabilistic model for label patterns

Given the fixed set of segments S, we will now make use of the SPT topology to model regular contextual information with a CRF for the graphical model.

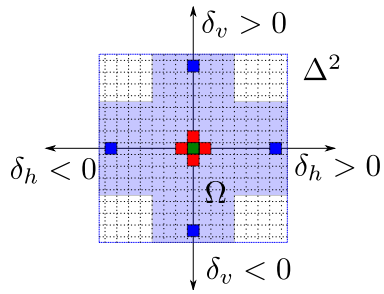

Fig. 1. Spatial template Ω is a subspace in the domain D_{AP} given by relation functions δ_h, δ_v. The center corresponds to the exact alignment in both axes. If segment u (green) is located in the center, other squares (red for *adjacency*, blue belong to *Aligned Pairs*) correspond to discrete relative positions of segment v.

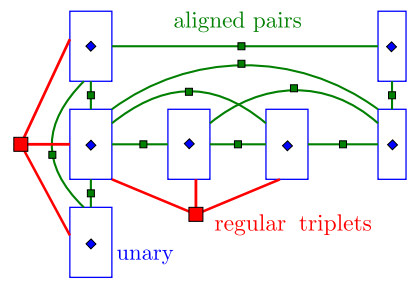

Fig. 2. Factor graph for regular SPT. Segments S are shown as blue rectangles s_i (i.e. corresponding to *window* frames), factors are solid squares. *Aligned Pairs* connect only segments in mutual relative position specified by the template in Fig. 1. *Regular Triplets* then combine two aligned and equally spaced pairs together.

For clarity we rewrite (1) in a convenient form

$$p(L|X,S) \;\propto\; \prod_{u\in S} e^{\varphi_1(\nu_u)} \times \prod_{(uv)\in AP} e^{\varphi_2(\nu_u,\nu_v)} \times \prod_{(uvw)\in RT} e^{\varphi_3(\nu_u,\nu_v,\nu_w)}, \quad (2)$$

where $\nu_i = (l_i|s_i, x_i)$ are variables related to node i and $\varphi_1, \varphi_2, \varphi_3$ are unary, pair-wise (AP) and ternary (RT) potential functions (factors) respectively. We will now discuss features used in these factors.

The unary potentials $\varphi_1(\nu_i) = \log p(l_i|s_i, x_i)$ are outputs of a multi-class classifier evaluated on the features for an image patch x_i of the segment s_i. The feature vector $f(x_i)$ is extracted from the image data by appending histogram of gradients (HoG), color (HSV), relative size, position, aspect ratio and 2D auto-correlation function.

Pairwise potentials for AP are restrictions on the template learned for concrete label pairs. They are based on a discretized version of the relative location distribution [4], similar form is used in [15] for *adjacency*. It is the statistical function

$$\varphi_2(\nu_u, \nu_v) = \theta_{2,d_h,d_v} \;\log p(l_u, l_v \mid d_h, d_v), \quad (3)$$

where d_h are the values of horizontal alignment $d_h(s_{u1}, s_{v1})$ specified in Fig. 3, analogically d_v for vertical. The **pattern** of labels l_u, l_v is the empirical distribution in the given relative locations d_h, d_v computed as the second order co-occurrence statistics of the labels for pairs of segments observed in a training set. The co-occurrence frequencies are obtained from a training set for each pair of class labels and are accumulated for all values in the spatial template domain Ω_{AP}. Figure 4 shows the resulting histograms of AP in Fig. 1.

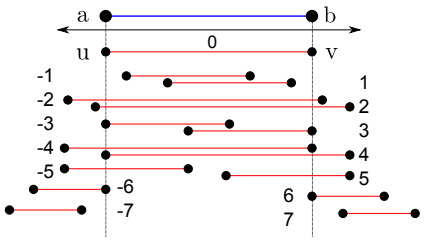

Fig. 3. Given interval (a, b) the figure shows the values Δ_a of alignment relation function d_a for a set of intervals (u, v), ranging from 0 (aligned) to ± 7 (no overlap). More free space between intervals corresponds to higher absolute values $(8, 9, 10, \dots)$ in Δ. Positions are considered equal within 10% tolerance of the interval length.

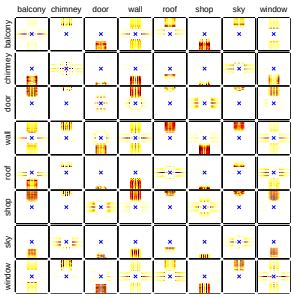

Fig. 4. Discrete relative co-occurrence location histogram $p(l_u, l_v, d_h, d_v)$ for label pairs in the *ECP-Monge* dataset. It holds information such as 'sky is usually above windows' or 'balconies are aligned vertically with windows'. Dark colors correspond to high frequency, blue cross marks $d = 0$ (equality).

Ternary potentials models regularity by encouraging some labels in RT to have the same value (i.e. *window*) in

$$\varphi_3(\nu_u, \nu_v, \nu_w) = \begin{cases} \theta_{3,c} & \text{if } l_u = l_v = l_w = c, \\ \theta_{3,0} & \text{otherwise,} \end{cases} \tag{4}$$

which is a generalized Potts model [5] and $\theta_{3,c}$ is a learned class-specific parameter. We do not use the complex ternary co-occurrence statistic with this potential because there is not enough data for its training. To facilitate efficient learning, we convert ternary potentials into pairwise by adding a hidden variable for each ternary factor φ_3.

Piece-wise parameter learning. The unary potential classifiers are trained independently to reduce the number of free parameters in the joint CRF learning process. For binary potentials (including the reduced ternary potentials) we use pseudo-likelihood learning procedure to obtain values of the potential weights θ. This process corresponds to structure learning within the domain Ω_{AP} limited by the SPT topology, resulting in $\theta_2 \to 0$ where the relation does not contribute to the discriminative power of the CRF. In practice this amounts to learning ~ 200 parameters based on likelihood in 50 sampled images, each of them with approximately 500 label variables, 3000 pair and 100 triplet factors. The training process takes several hours to complete (8 cores, 2 GHz) using Mark Schmidt's UGM library (`www.di.ens.fr/~mschmidt/Software/UGM.html`).

Inference. Because some of our potentials have a general form, exact inference is not possible and we use an approximate algorithm [6] to compute the marginal distributions of the labels, with run time around 30 s per image.

Method	SPT (proposed)				Three Layers [10]		SG [14]	HCRF [19]
Classifier	*SGT*	SVM			RNN		RF	RDF
Spatial pattern		NC	AP	**APRT**	NC	Adjacency	BSG	HAdj
Prob. model		-	Cooc	**Cooc**	-	Potts	SG	CS-Potts
ECP-Monge (8)	88.5	59.6	79.0	**84.2**	82.6	85.1	74.7	-
eTrims (8)	93.7	56.7	77.4	**82.1**	81.1	81.9	-	65.8
CMP Facade (12)	84.8	33.2	54.3	**60.3**	-	-	-	-

Table 1. Pixel-wise accuracy comparison on facade datasets (number of classes in brackets). *Abbreviations*: *SGT*=Segments with Ground Truth labels, *NC*=No Context, *AP*=Aligned Pairs, *RT*=Regular Triplets, *Cooc*=Coocurence, *BSG*=Binary Split Grammar, *HAdj*=Hieararchical Adjacency, *RNN*=Recursive Neural Network, *RF*=Randomized Forest, *SG*=Shape Grammar, *HCRF*=Hierarchical CRF.

4 Experimental results

We have validated our method on two public datasets annotated into 8 classes (like *wall, window, balcony etc.*). In addition in this paper we introduce a new large facade dataset.

The public *ECP-Monge* dataset is available from [14] (we use corrected ground truth labellings from [10]). It contains 104 rectified facade images from Paris, all in uniform Hausmannian style. Next, the public *eTrims* database [7] contains 60 images of buildings and facades in various architectural styles (neo-classical, modern and other). We rectified them using vanishing points.

We have compiled a new publicly available larger *CMP Facade* database [16] with ~ 400 images of greater diversity of styles and 12 object classes.

Figure 5 shows parsing results for different contextual models, additional results can be found in the report [16]. Table 1 provides their pixel-wise accuracy and comparison with other methods based on 5-fold cross validation. We have used method [2] to extract averagely 500 segments (independently on the image resolution) and show it under *SGT*, where ground truth labels of pixels within each segment have been collected and the most frequent label among them selected for the entire segment. The result is the maximum achievable accuracy with this segmentation, inaccurate localization of the segment borders is currently the main limiting factor (we are 4.3% below the limit on *ECP-Monge*).

The main observation is that contextual information improves the accuracy averagely by 20% when statistics on *AP* is used, and by further 4% when *RT* are included. The *RT* help mostly with window and balcony identification, thanks to the statistics of these labels following regular pattern in the dataset. The qualitative improvement is noticeable, even when their effect on the total pixel-wise accuracy is small, which is a sign it is not a very suitable measure. A more sophisticated local classifier could make the structural part of the model almost unnecessary, as observed in [10], but such model may be overly reliant on a good training set and perhaps prone to overfitting.

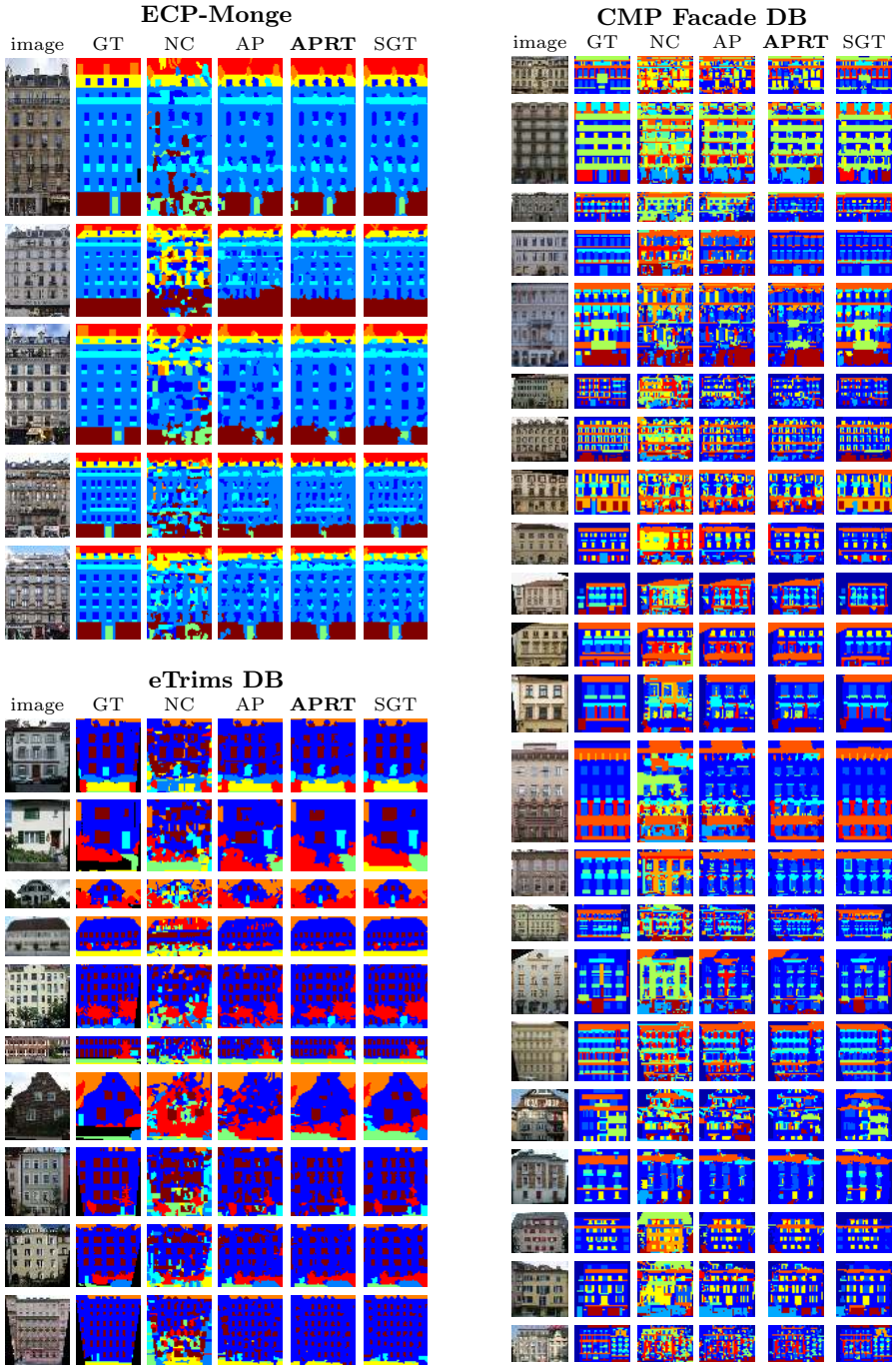

Fig. 5. Selected visual results on facade dataset, our result with full model is under *APRT*, (note it cannot be better than *SGT*). See legend in Tab. 1 for abbreviations.

5 Conclusion

We have introduced the concept of *Spatial Pattern Templates* for contextual models. The proposed *Aligned Pairs* and *Regular Triplets* templates have been found useful for segmentation of regular scenes by increasing accuracy of facade image parsing. Our next interest is to improve the quality of the segment extraction to increase accuracy of their borders.

Acknowledgement: This work was supported by the Czech Science Foundation under Project P103/12/1578.

References

1. Čech, J., Šára, R.: Languages for constrained binary segmentation based on maximum a posteriori probability labeling. IJIST 19(2), 69–79 (2009)
2. Felzenszwalb, P.F., Huttenlocher, D.P.: Efficient graph-based image segmentation. IJCV 59(2), 167–181 (2004)
3. Galleguillos, C., Rabinovich, A., Belongie, S.: Object categorization using co-occurrence, location and appearance. In: Proc. CVPR (2008)
4. Gould, S., Rodgers, J., Cohen, D., Elidan, G., Koller, D.: Multi-class segmentation with relative location prior. IJCV 80(3), 300–316 (2008)
5. Kohli, P., Ladicky, L., Torr, P.: Robust higher order potentials for enforcing label consistency. IJCV 82(3), 302–324 (2009)
6. Kolmogorov, V.: Convergent tree-reweighted message passing for energy minimization. Trans. PAMI 28(10), 1568–1583 (2006)
7. Korč, F., Förstner, W.: eTRIMS image database for interpreting images of man-made scenes. Tech. Rep. TR-IGG-P-2009-01 (2009)
8. Ladicky, L., Russell, C., Kohli, P., Torr, P.: Associative hierarchical CRFs for object class image segmentation. In: Proc. ICCV. pp. 739–746 (2009)
9. Lafferty, J., McCallum, A., Pereira, F.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: Proc. ICML (2001)
10. Martinovic, A., Mathias, M., Weissenberg, J., Gool, L.J.V.: A three-layered approach to facade parsing. In: Proc. ECCV. pp. 416–429 (2012)
11. Nowozin, S., Gehler, P., Lampert, C.: On parameter learning in CRF-based approaches to object class image segmentation. In: Proc. ECCV. pp. 98–111 (2010)
12. Schmidt, M., Murphy, K., Fung, G., Rosales, R.: Structure learning in random fields for heart motion abnormality detection. In: Proc. CVPR (2008)
13. Schmidt, M., Murphy, K.: Convex structure learning in log-linear models: Beyond pairwise potentials. In: Proc. AISTATS (2010)
14. Simon, L., Teboul, O., Koutsourakis, P., Paragios, N.: Random exploration of the procedural space for single-view 3D modeling of buildings. IJCV 93(2) (2011)
15. Tighe, J., Lazebnik, S.: Understanding scenes on many levels. In: Proc. ICCV. pp. 335–342 (2011)
16. Tyleček, R.: The CMP facade database. Research Report CTU–CMP–2012–24, Czech Technical University (2012), `http://cmp.felk.cvut.cz/~tylecr1/facade`
17. Tyleček, R., Šára, R.: Modeling symmetries for stochastic structural recognition. In: Proc. ICCV Workshops. pp. 632–639 (2011)
18. Tyleček, R., Šára, R.: Stochastic recognition of regular structures in facade images. IPSJ Trans. Computer Vision and Applications 4, 12–21 (2012)
19. Yang, M., Förstner, W.: A hierarchical conditional random field model for labeling and classifying images of man-made scenes. In: Proc. ICCV Workshops (2011)