ABSTRACT

Building Energy Modeling (BEM) is a time-consuming and costly process, primarily due to the extensive manual inputs required for model development. The definition of geometry inputs for models of existing buildings is particularly difficult. We present an automatic procedure for the analysis and segmentation of building facades in low-resolution aerial photographs in order to recover their semantic structure as well as to synthesize geometric models for reuse in the BEM process. An iterative approach for facade parsing is developed for the detection of facade structures which is robust to image noise and under-sampling artifacts as they are common in aerial photographs. In each step, a further refined shape grammar representing the high-level facade topology is inferred from the low-level facade element detection. Then, the low-level detection is improved by using feedback from the previously generated high-level structure. Based on this iterative analysis, we recover high-level facade topology by deriving a compact procedural description which takes into account symmetries and repeating patterns corresponding to floors and windows. The facade structure is encoded into common BEM formats. Moreover, synthesis of realistic facade textures is achieved by replacing low-resolution segments from the aerial images with best matching high-resolution pictures from a repository of facade elements.

INTRODUCTION

Generation of BEMs currently requires extensive manual input (Bazjanac et al., 2011; Bazjanac, 2001). Recent work focuses on linking Building Information Modelling (BIM) to BEM for accelerating the model development process by importing data contained within BIM into BEM tools (O’Donnell et al., 2013; Yan et al., 2013; Jeong et al., 2015). Such approaches significantly improve the process of generating BEMs by drawing on existing BIM data sources and extending models for energy simulation.

Unfortunately, geometric data coverage of real buildings is still very limited. As there is a growing abundance of aerial image data available, it is important to investigate techniques that are capable of extracting high level information in order to create detailed geometric and contextual models. Generally, automating the generation of building geometry models for the use in BEM software is a critical factor in encouraging building practitioners to use such tools throughout the life cycle of a building.

A variety of related facade segmentation and reconstruction methods (Müller et al., 2007; Xiao et al., 2008, 2009; Wu et al., 2010; Shen et al., 2011; Musialski et al., 2012; Martinovic and Van Gool, 2013; Ham and Golparvar-Fard, 2014, 2013) were developed for obtaining or recovering detailed geometric structures of building facades captured from high-resolution photographs, e.g., street-view photos. However, current methods are not sufficiently robust to be directly applied to aerial facade images. The reason for this is the image noise level in the available photo material which does not allow for a proper handling of strongly distorted facade elements (windows, doors, balconies, etc). In order to overcome the shortcom-
Listing 1: Simple example of the generated CityGML code

```
<core:CityModel>
  <core:cityObjectMember>
    <bldg:Building>
      <!-- LOD3 building envelope -->
      <bldg:boundedBy>
        <gml:LinearRing>
          <gml:pos>0.0 0.0 0.0</gml:pos>
          <gml:pos>1.9 0.0 0.0</gml:pos>
          <gml:pos>1.9 2.6 0.0</gml:pos>
          <gml:pos>0.0 2.6 0.0</gml:pos>
          <gml:pos>0.0 0.0 0.0</gml:pos>
        </gml:LinearRing>
      </bldg:boundedBy>
    </bldg:Building>
    <bldg:Window>
      <gml:LinearRing>
        <gml:pos>0.1 0.1 0.0</gml:pos>
        <gml:pos>0.4 0.1 0.0</gml:pos>
        <gml:pos>0.4 0.4 0.0</gml:pos>
        <gml:pos>0.1 0.1 0.0</gml:pos>
        <gml:pos>0.1 0.1 0.0</gml:pos>
      </gml:LinearRing>
    </bldg:Window>
  </core:cityObjectMember>
</core:CityModel>
```

Listing 2: Simple example of the resulting IDF code

```
BuildingSurface:Detailed,
  Zn009:Wall004, !- Name
  Wall, !- Surface Type
  EXTERIOR, !- Construction Name
  ...
  0.0,0.0,0.0, !- Vertex 1 {m}
  1.9,0.0,0.0, !- Vertex 2 {m}
  0.0,2.6,0.0, !- Vertex 3 {m}
  1.9,2.6,0.0, !- Vertex 4 {m}
  FenestrationSurface:Detailed,
  Zn009:Wall004:Win001, !- Name
  Window, !- Surface Type
  SCWINDOW, !- Construction Name
  ...
  0.1,0.1,0.0, !- Vertex 1 {m}
  0.4,0.1,0.0, !- Vertex 2 {m}
  0.1,0.4,0.0, !- Vertex 3 {m}
  0.4,0.4,0.0, !- Vertex 4 {m}
```
Figure 3: Automatic analysis and segmentation of the aerial facade photography and final shape grammar output as well as conversion into BEM formats

**METHOD**

This section provides insights into the different steps of our method. A detailed outline of the system is shown in Figure 3.

1. **Low-Level Structures Segmentation**

The input to our framework are low-resolution oblique facades captured in urban aerial images (see Figure 1). In the preprocessing, the 2D aerial image of urban areas will be automatically registered to the corresponding 3D city cadastral map (Habbecke and Kobbelt, 2010). Then the single oblique facade will be extracted and rectified based on the 3D coordinates of the building footprint specified by the cadastral map. Afterwards, the front-view images of the low-resolution building facades are processed by our iterative facade parsing procedure. In the beginning of each step, the low-level facade structure of semantic elements will be detected to infer the high-level facade topology. Then the feedback from the high-level structure is used to improve the low-level structure. In the following sections 1.1, 1.2, we will introduce the technique of low-level facade structure detection.

1.1 **Binary Structure Subdivision**

In this stage we aim at extracting the low-level structures of floor or column positions in given facade. The irregular facade structure will be decomposed into a set of regular regions by our binary structure subdivision procedure. The optimal splitting position and direction of each facade region are selected adaptively.

Enhanced Profile Accumulation A weak prior knowledge of urban buildings, originally observed by Müller et al. (2007) and Shen et al. (2011): horizontal (vertical) splitting planes should be placed where vertical (horizontal) lines are rare and horizontal (vertical) lines are dense, is widely used for extracting facade splitting planes. However, the low-level supporting evidence of facade structures, such as object boundaries, are strongly distorted by the noise and under-sampling artifacts of aerial imagery. In order to overcome this deficiency, the following metric functions 1 and 2 are proposed to signal the presence of horizontal or vertical splitting planes by introducing the high-level structural knowledge inferred by object detectors, such as a facade Hough forest (Gall and Lampitski, 2009).

\[
\begin{align*}
\text{Hor}(Y_i) &= \sum_{I_{x,y} \in \text{Img}} F_g (I_{x,y}, Y_i) \times G \left( d \left( I_{x,y}, Y_i \right) \right) \times \delta_x \left( I_{x,y} \right) \\
&\quad - \lambda \times \sum_{I_{x,y} \in \text{Img}} B_g (I_{x,y}, Y_i) \times G \left( d \left( I_{x,y}, Y_i \right) \right) \times \delta_y \left( I_{x,y} \right) \\
\text{Ver}(X_i) &= \sum_{I_{x,y} \in \text{Img}} F_g (I_{x,y}, X_i) \times G \left( d \left( I_{x,y}, X_i \right) \right) \times \delta_y \left( I_{x,y} \right) \\
&\quad - \lambda \times \sum_{I_{x,y} \in \text{Img}} B_g (I_{x,y}, X_i) \times G \left( d \left( I_{x,y}, X_i \right) \right) \times \delta_x \left( I_{x,y} \right)
\end{align*}
\]
sures the 2D distance between pixel position \((x, y)\) and the specified split plane, e.g., \(Y_i\) along Y-axis.

\[
G(d) = \frac{1}{2\pi\sigma^2} e^{-d^2/2\sigma^2} (\sigma \geq 0.5m)
\]

is a one-dimensional Gaussian kernel used to weighting pixel effects based on the distance between each pixel \(I_{x,y}\) and the splitting plane. \(d_x (I_{x,y})\), \(d_y (I_{x,y})\) (Function 3 and 4) are the partial differential functions used to suppress noise along \(x\) or \(y\)-axis. \(\lambda\) is a weighting parameter with value 0.2.

\[
d_x (I_{x,y}) = \max \left( \frac{\partial I_{x,y}}{\partial x}^2 - \alpha \times |\nabla I_{x,y}|^2, 0 \right) (3)
\]

\[
d_y (I_{x,y}) = \max \left( \frac{\partial I_{x,y}}{\partial y}^2 - \alpha \times |\nabla I_{x,y}|^2, 0 \right) (4)
\]

\(|\nabla I_{x,y}|\) is the gradient magnitude at position \((x, y)\), \(\alpha \in [0, 1]\).

Function 1 and 2 enhance the detection of facade splitting by placing horizontal (vertical) splitting planes at the background region, e.g., wall elements, where foreground objects and vertical (horizontal) lines, are rare and background wall elements and horizontal (vertical) lines, are dense.

A vertical accumulation profile, determined by function 1, is shown in Figure 4 left. Each valley of the profile represents a possible horizontal splitting plane (see Figure 4 middle) of given facade.

**Binary Structure Subdivision:** Extracting multiple splitting planes from all of the good peaks or valleys detected in the accumulation profile of given facade is a commonly used technique in profile-based facade subdivision approaches (Lee and Nevatia, 2004; Müller et al., 2007). However, after analyzing many facade structures, we find this splitting procedure, e.g., adaptive facade partition proposed by Shen et al. (2011), will also fail on irregular structures (see Figure 5).

Inspired by the work Zhang et al. (2013), we propose the following new strategy for handling complex facade partition which adaptively decomposes an irregular facade into a set of regular sub-facade regions (see Figure 6): 1) Generate the accumulation profile of given facade level along one image direction, e.g., X-axis; 2) Find each valley of the profile; 3) The valley of minimum value is selected as a candidate splitting plane of the specific image direction (see Figure 5 bottom left); 4) Redo the same procedure 1) to 3) for another direction, e.g., Y-axis, to detect another candidate position; 5) Compare profile values of these two candidates, the valley with smaller value is selected as the optimal splitting plane of current region, and its direction is also determined as the optimal splitting direction; 6) Binary subdivide current region into two sub-regions according to the optimal splitting plane (Figure 6a) and 7) Repeat step 1) to 6) for new divided sub-regions recursively until the termination condition is satisfied (see Figure 6d), e.g., no possible valley can be detected in the profile or the subdivision depth of current region reaches the maximum depth we allow.

**Figure 4:** From left to right: Y-axis accumulation profile \(\text{Hor}(Y_i)\), possible splitting positions along Y-axis, foreground merit map \(F_g(I_{x,y})\).

**Figure 5:** Adaptive partitioning of urban facades (Shen et al., 2011). From left to right: accumulation profiles of the given facade along X or Y-axis (the red circle indicates the minimum valley), splitting position detected in each direction.

**Figure 6:** Adaptive Binary Structure Subdivision
1.2 Facade Elements Detection

After the binary structure subdivision, an irregular facade is decomposed into a set of regular facade regions. The goal of this step is to subsequently detect positions and dimensional sizes of semantic elements in each decomposed regular region. The ISM-based Hough forest classification (Gall and Lempitsky, 2009; Gall et al., 2011) is applied to enhance the low level element detection.

**Randomized Hough forest:** As shown in Figure 7, a small dataset of low-resolution aerial facades is annotated for training a randomized Hough forest. For each image patch i in the training data, a vector \( P = (I_i, C_i, d_i^c) \) is extracted to represent the image feature \( I_i \), class label \( C_i \in \{ \text{object, wall, outlier} \} \) and the offset \( d_i^c = (x_i - x_c, y_i - y_c) \) between patch center \((x_c, y_c)\) and the corresponding object center \((x_c, y_c)\).

![Figure 7: Low-resolution training dataset. Top: single-view low-resolution facade images. Bottom: annotated segmentation of these low-resolution facades. Each color represents a particular semantic label, i.e., yellow means wall elements, black denotes outliers, other colors represent facade elements.](image)

Then the set of extracted patches \( P_{set} = \{ P = (I_i, C_i, d_i^c) \} \) is used to train a randomized Hough tree \( T_i \) by recursively dividing testing patches into the left or right child node based on the binary test minimizing the following two uncertainties: 1) Class-label uncertainty, \( U_1(P_{set}) = -\sum_{C_j \in \{ \text{object, wall, outlier} \}} P(C_j | P_{set}) \log(P(C_j | P_{set})) \) which measures the impurity of the class labels \( C_j \); and 2) Offset uncertainty, \( U_2(P_{set}) = \sum_{\text{offset vectors } \vec{d}_i} P(\text{offset vectors } \vec{d}_i | P_{set}) \) measuring the impurity of offset vectors \( \vec{d}_i \). The same training procedure for generating randomized hough tree \( T_i \) is repeated on given patches \( P_{set} \) to train the randomized Hough forest \( T_{set} \), which is a set of hough trees \( T_{set} = \{ T_i \} \).

The randomized Hough forest \( T_{set} \) will be used as a low-level pixel-wise classifier to determine the class label \( C_i \) of each patch \( P_i \) sampled in the input facade image. The probability of patch \( P_i \) belonging to label \( C_i \) is measured by the formula:

\[
P(C_i|I_i) = \sum_{T_i} \frac{1}{|\{T_i\}|} \times P(C_i|I_i, T_i)
\]  

where \( P(C_i|I_i, T_i) \) is the conditional probability of assigning label \( C_i \) to patch \( P_i \) based on the discriminative feature \( I_i \). \( |\{T_i\}| \) is the total number of hough trees contained in the randomized Hough forest \( T_{set} \). So the merit function \( F_g(I_i, y) = P(C_i = \text{object} | I_i, y) \).

**Multiple Object Detection:** After predicting the labels of each sampling patch of given facade, the probability of corresponding object center \( h_i \) (a hypothesis) voted by patch \( P_i \) in given facade is determined by the formula:

\[
P(h_i \in H | I_i) = \sum_{T_i} \frac{1}{|\{T_i\}|} \times \frac{1}{|L_i|} \times P(C_i | I_i, T_i)
\]

where \( H = \{ p_i - \vec{d}_j \} \), \( D_i = \{ d_i^c \} \) denotes a set of offset vectors contained in the leaf \( L_i \) where patch \( P_i \) arrived at when it is tested on the hough tree \( T_i \). \( p_i = (x_i, y_i) \) represents the central position of patch \( P_i \). Then \( H \) represents a set of potential object positions voted by formula 6. Indeed, \( |L_i| \) denotes the total number of training patches arrived at leaf \( L_i \). Then based on the hypothesis space \( H \), voted by formula 6, a set of hypotheses, whose probabilities are above the voting threshold, will be detected as semantic elements of given facade.

2 High-Level Facade Topology Optimization

The low-level facade structure detection might be affected by the severe distortions of the low-resolution facade imagery. Indeed, the high-level topology inferred by the low-level structure might fall in a local optimal solution. In this stage, the high-level facade topology will be refined by grammar-based structure optimization, and the facade symmetry is then detected to recover errors that exist in the low-level structure. In the following sections 2.1, 2.2, we will introduce the technique applied for the high-level facade topology optimization.

2.1 Global Topology Optimization

As observed in many low-resolution facades, we find a serious shortcoming commonly exists in bottom-up facade splitting systems, such as Lee and Nevatia (2004), Müller et al. (2007) and Shen et al. (2011). Their system relies heavily on the first splitting step and there is no way for recovering from severe errors in this stage, e.g., over-splitting planes. Postprocessing results of their system, e.g., symmetry detection, splitting rectification, are not sufficient if the detection of first splitting planes is not perfect.

In order to resolve this problem, we introduce a global optimization procedure to refine the high-level topology derived from low-level structure detection: Based on the low-level semantic information detected (see
where $s$ is the initial high-level topology (see Figure 8b). In this hierarchical tree structure, each node is considered as a non-terminal state $s = (I_i, L_i)$ representing a region that is split by our binary structure subdivision. $I_i$ is the image region covered by $s$, and $L_i$ is the shape label. Each leaf is considered a terminal state representing a semantic shape, e.g., doors, windows, or wall elements, defined in the shape-grammar.

Figure 8a), the corresponding binary shape-grammar tree of the given facade is generated to represent the initial high-level topology (see Figure 8b). In this hierarchical tree structure, each node is considered as a non-terminal state $s = (I_i, L_i)$ representing a region that is split by our binary structure subdivision. $I_i$ is the image region covered by $s$, and $L_i$ is the shape label. Each leaf is considered a terminal state representing a semantic shape, e.g., doors, windows, or wall elements, defined in the shape-grammar.

Therefore, the generation of the correct grammar tree of given facade will be formulated as solving the Markov decision process (MDP) (Formula 7) in order to find the global optimal topology that fits the bottom cues detected, like the foreground merit map determined by function $F_g(I, Y)$.

$$V^\pi(s) = \sum_a \pi(s,a) \times \left[ R(s,a) + V^\pi(s') \right]$$ (7)

where $\pi(s,a) = P(a|s)$ denotes the sampling probability for taking splitting action $a = (D_a, P_a)$ at state $s$. $D_a$ is the split direction of action $a$ and $P_a$ is the corresponding split position. $\pi(s,a)$ is determined by the value of accumulation profile $\text{Hor}(Y_i)$ or $\text{Ver}(X_i)$, e.g., the more close to profile valley, the higher sampling probability. $R(s,a)$ is the immediate reward for taking action $a$ at state $s$. $s'$ is the next state after taking action $a$ at $s$. So $a^* = \arg\max(V^\pi(s))$ is the optimal action learned to be taken on current state $s$. The set $\{a^*\}$ for each $s$ in the tree structure represents the global optimal topology learned from bottom cues.

### 2.2 Enhanced Facade Symmetry Detection

Based on the globally optimized high-level facade topology, the symmetry information of repeated facade structures is detected for recovering errors caused by the low-level analysis, such as missing detecting of facade elements due to the severe distortions of aerial facades (see Figure 8c). An enhanced facade symmetry metric, combining structure as well as texture information, is proposed for overcoming the effect of image noise and under-sampling artifacts.

**Preprocessing:** The optimized positions of facade elements at current region (see Figure 8c) is transferred into a binary structural map (see Figure 8d). In this map, each white pixel belongs to an object, and all black pixels belong to wall elements. Furthermore, the corresponding splitting planes of given facade region will be easily extracted from the profile accumulation of this binary structure map.

**Symmetry Detection:** $\mathbf{r}_i$ denotes the rectangular region $(x_i, y_i, w_i, h_i)$ of a detected element. $S = \{\mathbf{r}_i\}$ represents a set of facade elements that are split into the same facade slice, such as floor, column, etc. Then the structural symmetry between two facade slices $S_i$ and $S_j$, at the same depth of subdivision is formulated as follows:

$$R_{ij} = \frac{R(S_i \cap T(S_j))}{R(S_i \cup T(S_j))} = \sum R(\mathbf{r}_i \cap T(\mathbf{r}_j)) \over \sum R(\mathbf{r}_i \cup T(\mathbf{r}_j))$$ (8)

Where $T(S_j)$ is a translational movement of $S_j$ specified by the optimal rigid transformation $T$ along the splitting axis. $R_{ij} \in [0,1]$.

$$P_{ij} = \text{NCC}(S_i, S_j)$$ (9)

where $P_{ij}$ is the Normalized Cross Correlation (NCC) between these two image regions. Then the final similarity between $S_i$ and $S_j$ is formulated by the following function:

$$\text{Similarity}(S_i, S_j) = \alpha \times R_{ij} + (1-\alpha) \times P_{ij}$$ (10)

where the weighting parameter value of $\alpha \in [0,1]$ is configured based on the quality of input aerial images.

**Group Structure Refinement:** After the symmetry detection, facade slices of similar structures will be classified into the same group. Based on this group information, the boundary positions of each group member can be efficiently rectified by minimizing the energy function of boundary alignment. The detail algorithm is introduced in Shen et al. (2011).

**RESULTS**

In order to evaluate our system, we collected 900 single-view, low resolution aerial facades, selected from more than 2000 buildings with varying architectural structures using different resolutions. 1) The average resolution of our aerial facades is approximately the magnitude of $100 \times 100$ pixels (see Figure 9 and 10); 2) Most of the low level facade elements in our aerial photographs contain noise and sampling artifacts which can affect the segmentation quality; 3) Some parts of the tested building facades are covered by neighboring buildings or contain obstacles such as plants and shadows.

**Facade Segmentation Results**

Figure 9 shows a collection of perfect segmentations from our benchmark dataset whereas Figure 10 shows segmentation affected by local occlusions and distortions. The raw input images are shown on top and the segmentation along with the exact detected windows...
are shown below. 1) Each detected facade element is marked by a rectangular shape of white color; 2) The symmetry information of high-level structures is identified by different color groups and 3) If the dimensional size of a facade slice is smaller than 0.3m, our system will terminate the parsing on this small region in order to improve the general efficiency.

Figure 9: Aerial facades segmentation (dimensions in pixel). Top: front-views of rectified low-resolution facades. Bottom: perfect segmentation results. Rectangles represent detected facade elements. Colorful regions denote different symmetry groups.

Figure 10: Aerial facades segmentation (dimensions in pixel): high-level topologies detected on given facades are affected by local occlusions or distortions.

Table 1: Root-mean-square errors (RMSE) for all real and reconstructed windows and doors.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>RMSE abs.</th>
<th>RMSE pct.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Width</td>
<td>27.144cm</td>
<td>2.502%</td>
</tr>
<tr>
<td>Height</td>
<td>71.949cm</td>
<td>6.631%</td>
</tr>
<tr>
<td>Horizontal position</td>
<td>39.228cm</td>
<td>3.616%</td>
</tr>
<tr>
<td>Vertical position</td>
<td>40.948cm</td>
<td>3.774%</td>
</tr>
</tbody>
</table>

Complete example and validation Figure 11 shows a complete example of IDF and CityGML file generation for BEM as well as a comparison between real elements and reconstructed elements. 1) Depth information of the building was manually added by the authors and the roof section has been omitted as our method currently only deals with the facade structure; 2) The suitable high-resolution geometries are successfully recovered; 3) The visual quality is enhanced by automatically applying new textures from our database. The total width of our test building is 1085cm. Comparing the real parameters of our reconstructed doors and windows with the estimated parameters results in the root-mean-square errors \( \text{RMSE}(\hat{\theta}) = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (\hat{\theta}_i - \theta_i)^2} \) shown in Table 1.

Figure 11: Full example of a reconstructed IDF and CityGML file and comparison of original and reconstructed facade elements.

CONCLUSION

In this paper, we investigated how the BEM generation process can be improved through an automatic facade reconstruction based on aerial photographs. The method is robust to image noise and under sampling artifacts as they are common in aerial photographs and can be used as a powerful complement to current 3D reconstruction methods. It not only allows for an improved level of detail creating the geometry models of BEM for existing buildings but also provides an opportunity to reconstruct entire city quarters for inclusion in energy simulation. As we do not know the ground truth of the aerial facade images in our benchmark, we manually evaluate the segmentation quality of the
low-resolution aerial facades (see Figure 10 and 9). In the statistical results, 675 facades are perfectly parsed by our system, while 102 façade results are erroneous, and 123 façade detection failed due to the image quality being too low.

FUTURE WORK
The method can be extended in such a way, that it takes into account multi-view satellite images. Photos from different angles provide richer information about the façade structure and can improve the robustness against image noise and obstacles.

REFERENCES


