Orientation-selective building detection in aerial images

Andrea Manno-Kovacs[a,*], Tamas Sziranyi[a]

[a]*Distributed Events Analysis Research Laboratory, Institute for Computer Science and Control, MTA SZTAKI, Hungarian Academy of Sciences, H-1111, Kende u. 13-17, Budapest, Hungary*

Abstract

This paper introduces a novel aerial building detection method based on region orientation as a new feature, which is used in various steps throughout the presented framework. As building objects are expected to be connected with each other on a regional level, exploiting the main orientation obtained from the local gradient analysis provides further information for detection purposes. The orientation information is applied for an improved edge map design, which is integrated with classical features like shadow and color. Moreover, an orthogonality check is introduced for finding building candidates, and their final shapes defined by the Chan-Vese active contour algorithm are refined based on the orientation information, resulting in smooth and accurate building outlines. The proposed framework is evaluated on multiple data sets, including aerial and high resolution optical satellite images, and compared to six state-of-the-art methods in both object and pixel level evaluation, proving the algorithm's efficiency.

Keywords: orientation selectivity, modified Harris for edges and corners, building detection, active contour

1. Introduction

Automatic building detection is currently a relevant topic in aerial image analysis, as it can be an efficient tool for accelerating many applications, like urban development analysis and map updating, also providing great

[*]Corresponding author. Tel.: +36 12796106.

Email addresses: andrea.manno-kovacs@sztaki.mta.hu (Andrea Manno-Kovacs), sziranyi@sztaki.mta.hu (Tamas Sziranyi)

support in crisis and disaster management, and in aiding municipalities in long-term residential area planning. Large, continuously changing areas have to be monitored periodically, requiring a huge effort if performed manually. Therefore, there is a high interest for automatic processes to facilitate such analysis.

A wide range of publications is available in remote sensing for building detection for sparsely located building objects, often based on shape estimation or contour outlining. Earlier works like Huertas and Nevatia (1988) introduced a technique for detecting buildings with rectangular components and shadow information. Line based segmentation techniques, like Lin and Nevatia (1998) were based on the extraction of line segments, processed with various methods. Following this principle, Unsalan and Boyer (2005) proposed an extension, where the street network was extracted from the segmented images and houses were detected based on graph theoretical algorithms. In the same manner, Sirmacek and Unsalan (2008) – denoted by *BoxFit* in the experiments – fused shadow and invariant color features with edge information in a two-step process. First, a building candidate was defined based on color and shadow features, then a rectangle was fitted using a Canny edge map. This sequential method was very sensitive to the deficiencies of both steps: inappropriate shadow and color information causing false candidates and inexact edge maps causing inaccurate detections. As the proposed method uses similar information sources, we can highlight the impacts of our contributions by direct comparisons during the evaluation.

Following the region-based trend, Song et al. (2006) introduced a segment-merge technique (*SM*), which considered building detection as a region level task and assumed buildings to be homogeneous areas (considering either color or texture). First, a building model prior was constructed with texture and shape features from a training building set. After selecting building-like regions, shape and size constraints were used to merge such regions into building candidates, followed by shadow and geometrical rules to finalize candidates. However, the basic assumptions influenced the success of the whole approach: when buildings could not be distinguished from the background by using color and texture features, further steps would also fail. Moreover, they assumed simple building models, so complex shapes could not be reconstructed. The orientation of a candidate building region seed was introduced as a useful feature, defining potential rectangle orientations.

A point process based technique was introduced in Ortner et al. (2008), which used stochastic geometry based on the superposition of segment and rectangle processes. The work of Katartzis and Sahli (2008) is based on a stochastic image interpretation model and applies a Markov random field model to describe the dependencies between the available hypotheses.

Latest publications can be grouped into hierarchical and graph model based approaches: A hierarchical approach was introduced in Benedek et al. (2012), using a multitemporal Marked Point Process (MPP) model combined with a bi-layer Multiple Birth and Death (*bMBD*) optimization for rectangular building detection. Object-level features (exploiting low level features) were integrated into a configuration function, which was then evaluated by a bMBD stochastic optimization process. The result of the process was a group of rectangles, representing detected buildings. Although the hierarchical approach of the method was able to handle diverse objects, it was limited by the applied features (gradient, color, shadow), therefore it had problems when detecting objects with weak features, like non-red roofs. Moreover, the applied strictly rectangular templates prevented the accurate detection of complex shapes.

A graph model based algorithm for polygonal building shape detection was developed in Izadi and Saeedi (2012), employing lines, line intersections, and their relations. Ok et al. (2013) introduced the *GrabCut* partitioning algorithm for building extraction. The algorithm first investigated the shadow evidence to select potential building regions by applying a fuzzy landscape generation approach to model the directional spatial relationship between buildings and their shadows (which motivated our orientation selective fusion step in the proposed method). Then a pruning process was developed to eliminate non-building objects. Finally, *GrabCut* partitioning detected the building regions. However, a drawback of the algorithm was its sensitivity to the shadow extraction step, therefore it had problems when detecting buildings without shadow or having only fragmented shadow parts.

The method presented here is based on the fact that feature point detectors can be applied efficiently for man-made object detection, also indicated in Martinez-Fonte et al. (2005), where Harris and SUSAN detectors, published in Harris and Stephens (1988) and Smith and Brady (1997), were validated for distinguishing man-made versus natural structures. The principle was followed also in Peng et al. (2005) using the Tomasi and Kanade corner detector and in Sirmacek and Unsalan (2009), introducing a graph construction approach (denoted by *SIFT-graph*) for urban area and building detection using SIFT keypoints proposed in Lowe (2004). The method applied a light and a dark template to represent buildings. First, SIFT feature points were extracted from the image, followed by graph based techniques to detect urban areas. The given templates helped to divide the point set into separate building subsets and to define the locations of the different objects without any shape estimation. However, in many cases, not all building objects could be represented by only two templates, moreover, the given features were not always enough to distinguish the buildings from the background.

Cui et al. (2008) and Cote and Saeedi (2013) introduced a method using Harris corner points; Sirmacek and Unsalan (2011) tested directional Gabor filter based feature points, Harris corner points and the FAST points of Rosten et al. (2010) for extracting different local feature vectors (LFV), estimating a joint probability density function for urban area detection, assuming that around such points there is a high probability for urban characteristics. This technique motivated our previous work in Kovacs and Sziranyi (2013) for introducing the Modified Harris for Edges and Corners (MHEC) feature point set for efficient urban area detection.

The main contribution of the present approach is the application of orientation as a novel feature in a direction-selective framework for detecting buildings. Earlier approaches like Sirmacek and Unsalan (2008) and Ok et al. (2013) usually dealt with orientation in terms of the illumination angle. Others applied orientation based techniques for indirect solutions, like Unsalan and Boyer (2004) to identify line support regions. Ortner et al. (2008) introduced the alignment interaction in the MRF energy term. Cui et al. (2012) followed a novel interpretation of orientation information, by using perpendicular building borders to define dominant directions. They were looking for lines in Hough space with orientations indicating potential straight building borders. Similarly, Benedek et al. (2012) extracted a local gradient orientation density function, published in Kumar and Hebert (2003), to find the main orientations characterizing a building and measuring the orthogonality of the candidate area.

However, these approaches concentrated only on one building and its neighborhood to perform the directional analysis, while our proposed method handles orientation as a region level feature and combines it with other features throughout the approach (see Fig. 1):

- The orientation feature is calculated for the whole image and an **improved edge map** is defined, which emphasizes edges only in the main orientations. This improved edge map is fused with color and shadow features, resulting in accurate localization of building candidates.

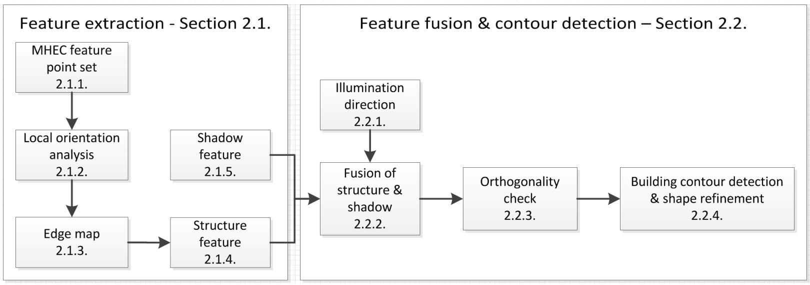

Feature extraction - Section 2.1.
MHEC feature point set 2.1.1.
Local orientation analysis 2.1.2.
Edge map 2.1.3.
Structure feature 2.1.4.
Shadow feature 2.1.5.
Feature fusion & contour detection – Section 2.2.
Illumination direction 2.2.1.
Fusion of structure & shadow 2.2.2.
Orthogonality check 2.2.3.
Building contour detection & shape refinement 2.2.4.

Figure 1: Main steps of the proposed method and the corresponding section numbers of the paper.

- An **orthogonality checking** of the possible building blobs is applied to find remaining candidates with limited feature evidence.
- A **directional refinement step** is performed after the Chan-Vese active contour based building detection phase: the shapes of the final blobs are refined by a novel directional operator to efficiently smooth the irregularities of the active contour outline.

2. Proposed framework

In this paper, a novel framework, called Orientation Selective Building Detection (**OSBD**) is proposed, based on the exploitation of multiple features: color, shadow map, and a novel orientation descriptor. Our guiding principle is that, when detecting buildings, orientation is a valuable information, as the alignment of buildings usually depends on the structural properties of their surroundings (e. g. the road network). Therefore the main edges of neighboring buildings should have similar orientations. The main contribution of this paper is to introduce region orientation as a novel feature for building detection and to use this information in a direct manner, unlike previous works e. g. Ortner et al. (2008) where only alignment interaction was calculated between building candidates, avoiding the use of the orientation value itself.

Moreover, we exploit orientation information in multiple steps: the directional descriptor also helps in the verification of the building candidates in the detection step. An orthogonality check is created to validate whether the candidate is a real building or some other image structure. Finally, a directional morphological operator is applied for the remaining building blobs to smooth the final outline, resulting in more accurate detection results.

The main steps of the proposed method with the corresponding subsections are shown in Figure 1. The two main parts are Feature extraction (Sec. 2.1) and Feature fusion and contour detection (Sec. 2.2). As a first step, a feature point set is extracted which is based on the modification of the Harris corner detector (Sec. 2.1.1), proposed in Kovacs and Sziranyi (2012a). This point set is used as a directional sampling set to compute orientation statistics in Section 2.1.2. Local orientation information is calculated as the main orientation of gradients in the close proximity of the feature points, and extended to the whole image to produce a directional map. Using this map, dominant directions describing the urban regions are defined, helping the construction of a more accurate edge map in Section 2.1.3 by specifying the favorable edge orientations. Joint edge and color information (later called as structure information in Section 2.1.4) is then integrated with shadow information (Sec. 2.1.5) in Section 2.2.2 using illumination direction (Sec. 2.2.1) to verify the possible building candidates and filter out false positive color and edge blobs. After a novel orthogonality check (Sec. 2.2.3), building shapes are detected with the Chan-Vese non-parametric active contour algorithm and the final outlines are refined by a proposed directional morphological operator, described in Section 2.2.4. The experimental evaluation (Section 3) includes parameter analysis and detailed experiments indicating the method's superiority over state-of-the-art approaches.

2.1. Feature extraction

As the first step, a novel orientation feature is extracted based on the Modified Harris for Edges and Corners (MHEC) feature point set. Orientations, describing the building areas, are applied to create an accurate edge map, complementing color features to form efficient structure information. Shadow features are then integrated with structure information like in Benedek et al. (2012) and Ok et al. (2013), introducing a novel orientation inspired framework for building candidate localization.

2.1.1. Modified Harris for edges and corners (MHEC)

The MHEC feature point set was first introduced in Kovacs and Sziranyi (2012a) and was proven to be efficient for urban area detection in Kovacs and Sziranyi (2013). The proposed algorithm adapts the R_{mod} (Eq. 1) modification of the original characteristic function of Harris and Stephens (1988).

$$R_{\mathrm{mod}} = \max(l_1, l_2), \tag{1}$$

where l_1 and l_2 denote the eigenvalues of the Harris matrix. Eigenvalues distinguish different regions: both of them are large in corner regions, only one of them is large in edge regions and both of them are small in homogeneous regions. Therefore the R_{mod} function separates homogeneous and non-homogeneous regions efficiently.

The advantage of the improved detector is an automatic balanced recognition of corners and edges. Therefore, it is an efficient tool for characterizing contour-rich regions, such as urban areas in aerial images.

Feature points are calculated as local maxima of R_{mod}. A pixel $p_i = (x_i, y_i)$ is the element of the P feature point set, if it has the largest $R_{\mathrm{mod}}(p_i)$ value compared to its neighbors in a surrounding $b_i = \{[x_i - 1, x_i + 1] \times [y_i - 1, y_i + 1]\}$ window and its $R_{\mathrm{mod}}(p_i)$ value exceeds a given $T_{\max}$ threshold:

$$P = \left\{ p_i : R_{\mathrm{mod}}(p_i) > T_{\max} \text{ AND } p_i = \underset{r \in b_i}{\operatorname{argmax}} R_{\mathrm{mod}}(r) \right\}. \tag{2}$$

Here, the $T_{\max}$ threshold is adaptively calculated by Otsu's method Otsu (1979) for each image.

The extracted MHEC point set for a sample aerial image is in Figure 2(d), showing points in both corner (like buildings) and edge (like roads) regions, points having higher density in urban areas. Only a few points are situated in non-urban areas.

Our previous work Kovacs and Sziranyi (2013) compared the MHEC point set to other feature point detectors, like SIFT of Lowe (2004), FAST of Rosten et al. (2010) or SUSAN of Smith and Brady (1997) and revealed that the MHEC point set represents urban areas efficiently, therefore the features extracted from the local neighborhood of these points contain valuable information for describing urban areas.

2.1.2. Local orientation analysis

A novel, orientation based concept for building detection was introduced in Kovacs and Sziranyi (2012b) which was extended for handling multidirectional areas in Manno-Kovacs and Sziranyi (2013). The idea was to calculate main gradient orientations in the small neighborhood around the feature points and by collecting such data, define the orientation histogram of the image.

Let us denote the gradient vector by ∇g_i with $\|\nabla g_i\|$ magnitude and φ_i^{∇} orientation for the i^{th} point (pixel) p_i. By denoting the $n \times n$ neighborhood around the point in image I with $W_n(i)$ (where n depends on the resolution), the weighted density of φ_i^{∇} is as follows:

$$\lambda_i(\varphi) = \frac{1}{N_i} \sum_{r \in W_n(i)} \frac{1}{h} \cdot \|\nabla g_r\| \cdot k\left(\frac{\varphi - \varphi_r^{\nabla}}{h}\right), \tag{3}$$

with $N_i = \sum_{r \in W_n(i)} \|\nabla g_r\|$ and $k(.)$ kernel function with h bandwidth parameter (See Figure 2).

The main orientation for the i^{th} feature point is defined as:

$$\varphi_i = \underset{\varphi \in [-90,+90]}{\text{argmax}} \{\lambda_i\}. \tag{4}$$

The process is illustrated in Figure 2 for a selected ith feature point, marking its neighborhood with a white rectangle. The neighborhood with $n = 15$ size is then enlarged in Figure 2(b) and the corresponding $\lambda_i(\varphi)$ local gradient orientation density function is in Figure 2(c). The maximum value of the $\lambda_i(\varphi)$ function, assigning $\varphi_i = -51$ orientation value for the point is marked in red.

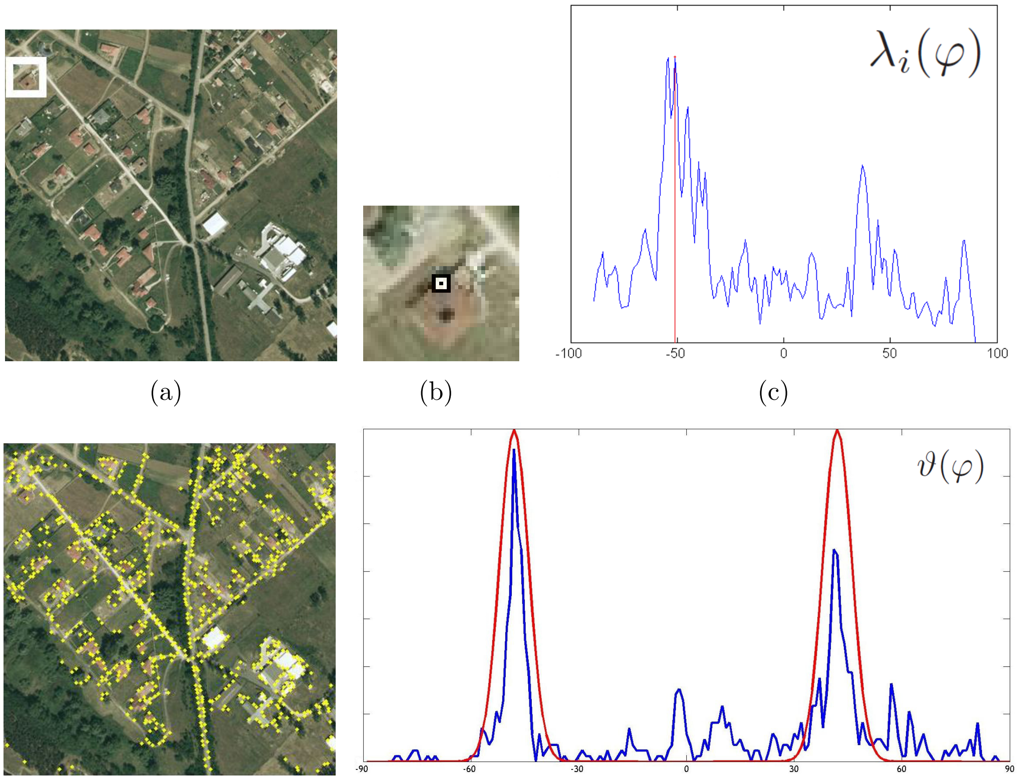
As buildings usually have orthogonal edges, in such case the $\lambda_i(\varphi)$ function is supposed to have two main peaks with a 90 degree difference. To test this assumption, the function was correlated with a bimodal Mixture of Gaussian (MG) function, where the two components had 90 degree difference between them, and the candidate was categorized based on the rate of the correlation.

In this approach λ_i is applied in a region-level context. The orientation histogram is computed for the whole image. After calculating the dominant direction for all K feature points, the histogram function ϑ is defined:

$$\vartheta(\varphi) = \frac{1}{K} \sum_{i=1}^{K} H_i(\varphi), \tag{5}$$

where $H_i(\varphi)$ is a logical function:

$$H_i(\varphi) = \begin{cases} 1, & \text{if } \varphi_i = \varphi \\ 0, & \text{otherwise.} \end{cases} \tag{6}$$



(a) (b) (c)

(d) (e)

Figure 2: Local orientation analysis: (a) is the original image, denoting the neighborhood ($n = 15$) of the ith feature point by a white rectangle; (b) is the cropped image showing only the investigated neighborhood of the point; (c) shows the $\lambda_i(\varphi)$ local gradient orientation density function for the feature point, with the main orientation $\varphi_i = -51$ marked in red; (d) shows all the K MHEC feature points in yellow and (e) is the calculated $\vartheta(\varphi)$ orientation histogram of the urban area in blue and the correlated bimodal Gaussian function in red, indicating main orientations of the urban area.

Figure 2(d) shows all feature points in yellow and the calculated $\vartheta(\varphi)$ orientation histogram is marked with blue in Figure 2(e). As the histogram is calculated for the whole image, unlike in Benedek et al. (2012) where the aim was to decide whether a building is situated around a point or not, the $\vartheta(\varphi)$ orientation histogram is expected to have multiple dominant peak pairs with 90 degree difference between them, caused by perpendicular edges of different buildings groups at various locations. Therefore the ϑ histogram has to be correlated to an unknown number of bimodal MGs.

To estimate the optimal number of MGs we introduce the *Iterative Bimodal Mixture of Gaussian Matching* (IBMGM) process (see Algorithm 1). In every iteration, a bimodal Gaussian function is correlated to the $\vartheta(\varphi)$ data function. The rate of correlation is measured by α:

$$\alpha(m) = \int \vartheta(\varphi)\eta_2(\varphi, m, d_\vartheta)\, d\varphi, \tag{7}$$

where $\eta_2(.)$ denotes a two-component MG, with m and $m + 90$ mean values and d_ϑ standard deviation, which was set by training. The orthogonal directions represented by this MG are marked by $\theta, \theta_{\mathrm{o}}$. The first dominant direction can be obtained as the value at the maximum correlation:

$$\theta_1 = \underset{m\in[-90,+90]}{\operatorname{argmax}} \left\{\alpha(m)\right\}. \tag{8}$$

Algorithm 1 Iterative Bimodal Mixture of Gaussian Matching (IBMGM) process

```
Input: $\vartheta(\varphi)$ orientation histogram
while $\alpha_{\mathrm{j}-1} = \alpha_{\mathrm{j}}$ AND $CPR \leq \epsilon$ do
  1. Correlate $\eta_2(.)$ MG to data $\vartheta(\varphi)$;
  2. Calculate $\alpha_j$;
  3. Update CPR and $\vartheta(\varphi)$;
  if $\alpha_j \geq \alpha_{j-1}$ OR $\alpha_j \leq \alpha_{\mathrm{th}}$ then
    $\alpha_{j-1} = \alpha_j$;
  end if
end while
Result: Main orientations: $\theta_1, \ldots, \theta_q$
```

The corresponding orthogonal direction, the other peak of the two-component MG:

$$\theta_{\mathrm{o},1} = \begin{cases} \theta - 90, & \text{if } \theta \geq 0 \\ \theta + 90, & \text{otherwise.} \end{cases} \tag{9}$$

Thus, in every iteration the most correlating MG is extracted. The α_j value is calculated in the jth iteration to measure the correlation, along with the number of the total correlated points (CP_j) to follow the overall likelihood. CP_j is calculated as the sum of the $\vartheta(\varphi)$ histogram values, involved in the actual η_2 MG. Bins having η_2 MG value more than 1% of the amplitude are selected. Therefore, the full width h of the Gaussian curve with A amplitude was measured at the height of $0.01 \cdot A$. The first component of the η_2 MG is:

$$\eta_{2,1}(\varphi) = A \cdot e^{\frac{-\varphi^2}{2d_\vartheta^2}}. \tag{10}$$

We calculate the width h at height value $0.01 \cdot A$ as:

$$0.01 \cdot A = A \cdot e^{\frac{-h^2}{2d_\vartheta^2}}, \tag{11}$$

$$h = \sqrt{2log(100)} \cdot d_\vartheta \approx 3 \cdot d_\vartheta. \tag{12}$$

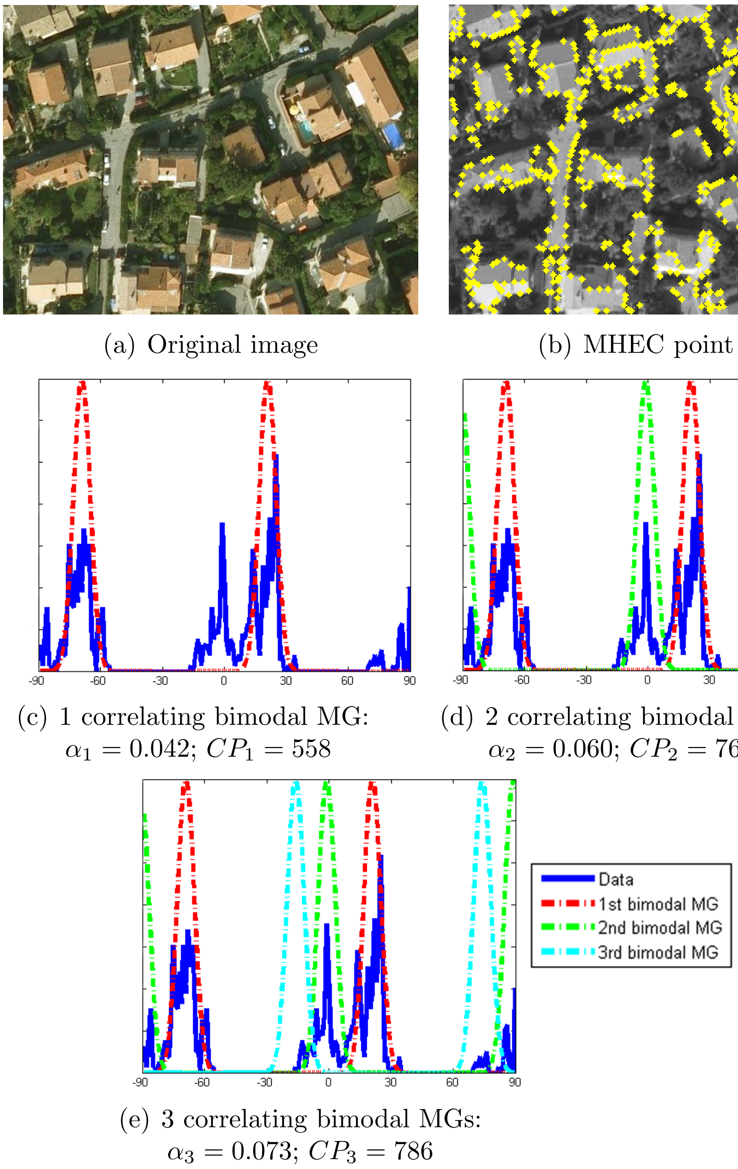
This means that the bins involved in η_2 MG in the jth iteration are: $[\theta_j - 3\ d_\vartheta, \ldots, \theta_j + 3\ d_\vartheta]$ and $[\theta_{\mathrm{o},j} - 3\ d_\vartheta, \ldots, \theta_{\mathrm{o},j} + 3\ d_\vartheta]$. The Correlated Point Ratio (CPR_j) is calculated as: $CPR_j = CP_j/K$, where K is the total number of feature points. In every iteration, the CPR_j value is updated and the iterative process is stopped if this value exceeds an ϵ threshold. The behavior of the α_{th} and ϵ IBMGM parameters is analyzed in Section 3.2.

This iterative process is responsible for picking the optimal number of MGs, and prevents the extraction of orientations which are not significant enough. The histogram data is also updated in each iteration, eliminating the already involved bins, as follows:

$$\vartheta(\theta_j - 3\ d_\vartheta, \ldots, \theta_j + 3\ d_\vartheta) = 0, \tag{13}$$

$$\vartheta(\theta_{\mathrm{o},j} - 3\ d_\vartheta, \ldots, \theta_{\mathrm{o},j} + 3\ d_\vartheta) = 0. \tag{14}$$

Figure 3 shows the iterations of the IBMGM process for defining the number of dominant directions (q). The calculated MHEC points (790 in total) are shown in Fig. 3(b). The correlating bimodal MGs and the belonging α_j and CP_j parameters are in Figs. 3(c)-3(e). We can see that the α_j parameter is increasing continuously and the CPR_j parameter reaches a high ratio in the second step, representing $CPR_2 = 768/790 \approx 0.97$ of the point set. The third MG (Fig. 3(e)) is included to illustrate the behavior in the next iteration: although α_j is still increasing, the newly correlated point set is small, containing only $CP_3 - CP_2 = 18$ points. Therefore, the estimated number of main orientations is $q = 2$, with peaks at $\theta_1 = 22$ ($\theta_{\mathrm{o},1} = -68$) and $\theta_2 = 0$ ($\theta_{\mathrm{o},2} = 90$).



(a) Original image

(b) MHEC point set

(c) 1 correlating bimodal MG: $\alpha_1 = 0.042$; $CP_1 = 558$

(d) 2 correlating bimodal MGs: $\alpha_2 = 0.060$; $CP_2 = 768$

(e) 3 correlating bimodal MGs: $\alpha_3 = 0.073$; $CP_3 = 786$

Figure 3: Correlating increasing number of bimodal mixture of Gaussians (MGs) with the ϑ orientation density function (marked in blue). The measured α_j and CP_j parameters are represented for each j step. The third component is determined insignificant, as it covers only 18 MHEC points. Therefore the estimated number of main orientations is $q = 2$.

To separate the influence of MHEC and the orientation-selective framework, we performed the local orientation analysis for different feature point detectors in Sec. 3.1. Results showed that the main orientations representing the urban area are not sensible to the applied feature point detector and remain almost constant in all cases.

2.1.3. Orientation selective edge feature

After obtaining the main orientations, this information can be applied to construct an improved edge map by only including edges in the main directions. This map is later combined with other lower level features (like color and shadow). Efficient edge detection is especially important in the case of buildings with weak color (e.g., gray or black roofs).

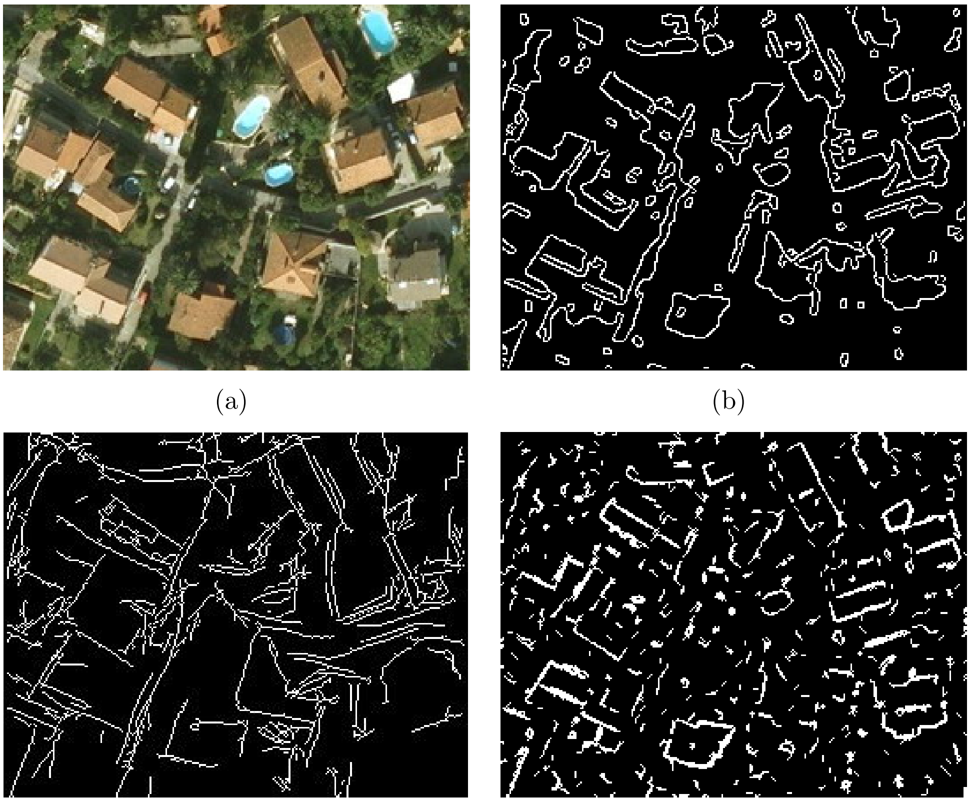
There are different approaches applying directional information, like Canny edge detection Canny (1986) using the gradient orientation, or Perona (1998) which is based on anisotropic diffusion, but they cannot handle cases with multiple orientations (like corners). Other single orientation methods exist, like Mester (2000) and Bigun et al. (1991), but their main issue is that they calculate orientation on the pixel-level and lose the scaling nature of orientation, therefore they cannot be used for edge detection in a higher level interpretation (like for object detection). Edge detection methods like shearlets of Yi et al. (2009) are using histogram bins instead of fixed directions. In the present case, edges constructed by joint pixels have to be enhanced, thus the applied edge detection method has to be able to handle the extracted orientation values. Moreover, as we are looking for building contours, the algorithm must handle corner points as well.

The Morphological Feature Contrast (MFC) operator was introduced in Zingman et al. (2014) for extracting isolated structures while suppressing textured details of the background. For doing so, the following operators were introduced for dark and bright features:

$$\psi^{+}_{MFC}(f) = \left| f - \gamma_{r_2}\delta_{r_1}(f) \right|^{+}, \tag{15}$$

$$\psi^{-}_{MFC}(f) = \left| \delta_{r_2}\gamma_{r_1}(f) - f \right|^{+}, \tag{16}$$

where γ is a morphological opening, δ is a morphological closing and r_1, r_2 denote the size of the square-shaped morphological structuring element (SE). The parameters of the SEs are chosen so that r_1 should be greater than the maximal distance between details of texture to be suppressed and r_2 should


(a)

(b)

(c)

(d)

Figure 4: Edgemaps extracted with different approaches for (a) original image (extracted edges in white, background is black): (b) for Canny edge detector Canny (1986); (c) for shearlet based edge detection Yi et al. (2009); (d) for MFC operator Zingman et al. (2014). Results show that the applied MFC-based technique is able to emphasize edges efficiently, while generating low number of false hits.

be greater than the size of the features to be extracted (chosen according to the resolution). Detailed parameter analysis is included in Sec. 3.2.

The motivation for using MFC is its ability to extract object boundaries (edge features) from textured backgrounds with high accuracy (Fig. 4(d)). Moreover, after applying the MFC operator, a subsequent filter γ_{lin}, obtained by the point-wise maximum of morphological openings with linear SE, extracts narrow linear structures with respect to the defined orientation of the SE. Therefore, it can be used for fast edge extraction in the defined main orientations. Figure 4 shows the results of different edge extraction algorithms, comparing Canny, shearlet and MFC, showing that MFC produces less false detections than the shearlet based method and that it includes more of the real building edges than Canny.

2.1.4. Structure feature

Previous methods applied the roof color feature as an evidence for building candidates. The u component of the CIE Luv colorspace was typically used with an adaptive Otsu thresholding published in Otsu (1979) to get a B_c binary color map (see Figure 5(b)). This color map is designed for roof colors with significant red component, typically for orange, red, brown roofs. However, in case of other colors without significant red channel, like gray or black roofs, this color feature does not provide enough information and additional features are needed to aid the detection.

To compensate for the drawbacks of the color feature, the improved MFC-based edgemap is fused with B_c and will be called *structure feature.* Figure 5(c) shows the fused structure feature map B_t, indicating that non-red roofs are also represented fairly for further detection.

2.1.5. Shadow feature

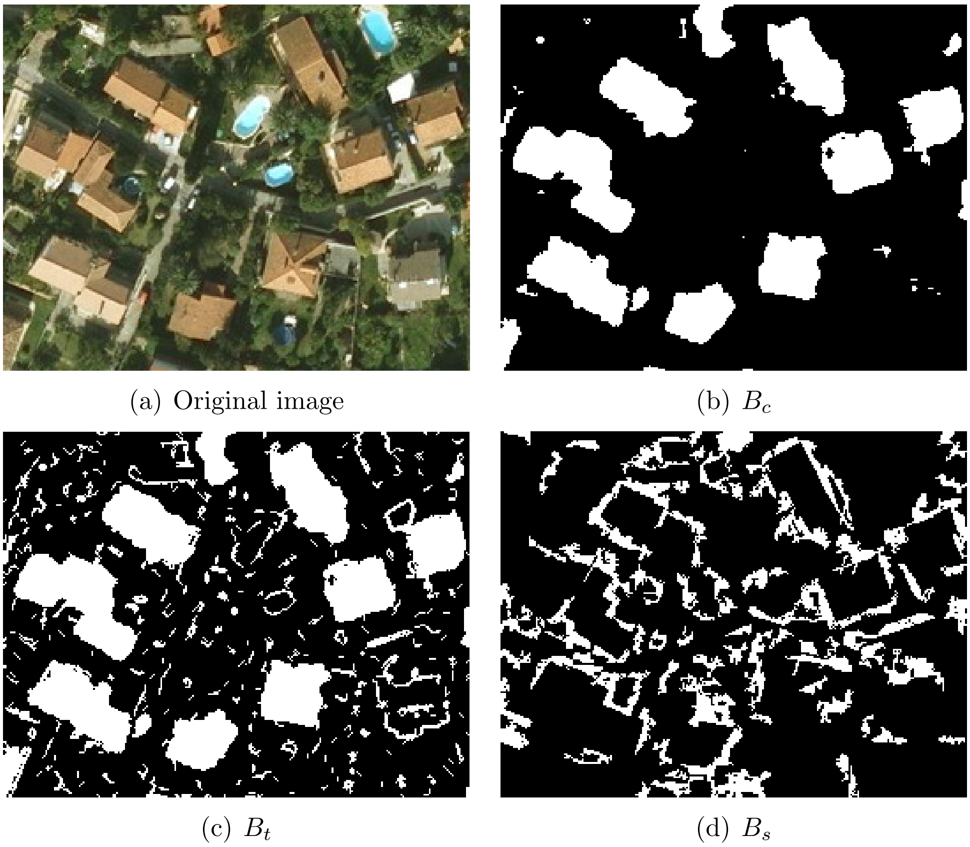
When the color feature is not relevant, the shadow feature can provide useful information about building objects. Moreover, it may also assist in distinguishing false color-based hits from real buildings. Shadow evidence has been used for building detection in recent works (Benedek et al. (2012); Ok et al. (2013)).

A shadow mask can be extracted by filtering pixels from the dark grayish and blueish color domain. Following the recommendations of Tsai (2006), we have tested many color spaces (like YIQ, YC$_b$C$_r$ and HSV) and finally applied the YC$_b$C$_r$ color space for shadow detection.

After applying *spectral ratioing*, the *ratio image* was calculated as follows:

$$R_{sh} = \frac{C_r + 1}{Y + 1}, \tag{17}$$

which enhances the increased hue property of shadows with low luminance. After an Otsu thresholding, we get the binary shadow map B_{sh}. As the ratio image is sensitive to greenish colors, B_{sh} may also contain vegetation areas which have to be eliminated.



(a) Original image

(b) B_c

(c) B_t

(d) B_s

Figure 5: Color, structure and shadow features: (a) is the original image; (b) is the color map, the detected color blobs are marked in white, while background is black; (c) shows the structure feature given as the fusion of color and edge information; (d) is the binary shadow map (Eq. 20).

Normalized Difference Vegetation Index (NDVI), calculated as a comparison of near-infrared (NIR) and red (R) channels is widely applied for vegetation extraction:

$$R_{NDVI} = \frac{NIR - R}{NIR + R}. \tag{18}$$

However, in the lack of the NIR channel, the NDVI index has been modified for RGB aerial images in the following way:

$$R_{veg} = \frac{G - R}{G + R}. \tag{19}$$

The B_{veg} binary vegetation map will be a binarization of R_{veg} using Otsu's method.

The final B_s binary shadow map (Fig. 5(d)) is given after eliminating the vegetation by a logical subtraction:

$$B_s = B_{sh} - B_{veg}. \tag{20}$$

2.2. Feature fusion and building contour extraction

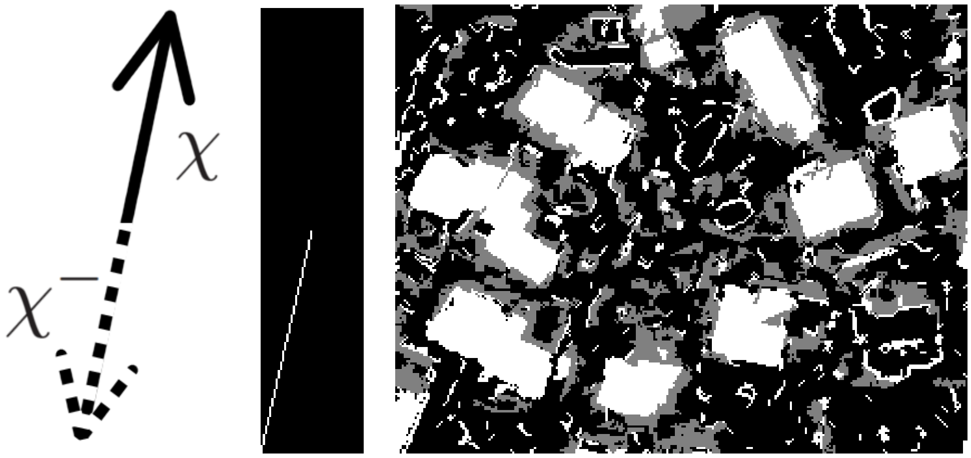
On the next level of the segmentation process we integrate the extracted structure and shadow features for obtaining building candidate blobs. Furthermore, candidates supported by only one feature are also investigated. The main steps of the feature fusion and final building contour detection are the following:

1. Orientation-selective structuring element for fusing shadow and structure features, using illumination direction;
2. Fusion of structure and shadow features to get coherent blobs for building candidates;
3. Orthogonality checking to find the remaining candidate blobs;
4. Calculation and refinement of building contours for the localized candidates.

2.2.1. Application of illumination direction

Illumination direction (denoted by χ in the paper) connects different features enabling the definition of their relations. While the structure information usually indicates the exact location of the building, shadow evidence indicates the presence of the building based on its dimensions (size and height). Moreover, in the lack of the structure feature, shadow information combined with illumination direction may estimate the possible location of a building.

Earlier methods Benedek et al. (2012); Ok et al. (2013) also applied the illumination direction, which was either provided in the image metadata or could be calculated automatically as in Sirmacek and Unsalan (2008) after considering the influencing facts such as building height and off-nadir position. In our algorithm, the illumination direction is assumed to be available, therefore we do not go into details about its calculation. A supplementary metadata file is supposed to be given with image acquisition details about date and time, the solar illumination angles (azimuth and elevation) and viewing geometry, from which the direction of solar illumination can be computed.



(a) (b) (c)

Figure 6: Initialization for the feature fusion process for Figure 5(a): (a) is the original and reverse illumination direction ($\chi = 78°$) ; (b) shows the anisotropic structuring element; (c) Shadow (B_s, gray) and structure (B_t, white) features visualized together.

Using the illumination direction, we define a direction-selective structural element which is applied for the fusion of shadow and structure evidence of the same building candidate. A similar idea was proposed in Aksoy and Cinbis (2010), but it was only used for estimating the building location from shadow evidence. In our case, the proposed direction-selective structuring element has a major role.

The direction-selective or anisotropic structuring element is constructed as a linear element directed along the reverse illumination direction (χ^-), and an anisotropic kernel is created with this linear element's origin in the center, denoted by S_{χ^-}. For a sample $\chi = 78°$ illumination direction, the created kernel in the reverse illumination direction is shown in Figure 6(b). The 0° direction is the horizontal axis and the value is calculated counterclockwise. The size of the kernel (the length of the linear element is 13) is chosen adaptively according to the resolution of the image, which is tested in Sec. 3.2

2.2.2. Fusion of structure and shadow features

The B_t structure and B_s shadow feature maps can be seen in Figure 6(c), showing the structure feature in white and the shadow feature in gray. If a structure blob and a shadow blob have the proper relation, we fuse them to create a building candidate. This proper relation is defined by the orientation-selective structural kernel. We investigate the following morphologically modified shadow map:

$$B_s^\chi = (B_s \oplus S_{\chi^-}) \ominus S_\chi, \tag{21}$$

which means that the blobs of B_s are shifted in the opposite illumination direction, with a morphological dilation ($\oplus$) in the χ^- direction, followed by a morphological erosion ($\ominus$) in the χ illumination direction. According to the resolution, a certain shadow size is expected in case of buildings, therefore the smaller blobs of B_s are eliminated; we only deal with shadow blobs having at least 20 connected pixels. This also guarantees to find important shadow blobs in a densely built area, where a shadow of one building falls on the other building. Sometimes color space transformation or spectral ratioing errors might occur, resulting in large, continuous shadow regions. Thus, we also remove blobs with more than 2000 pixels. These thresholds are based on image resolution and selected after training, which is discussed in Sec. 3.2.

In the first step of the fusion process, a building candidate is given by the ith separate shadow blob of B_s (marked as $B_{s,i}$) and the corresponding structure blob of B_t, marked as $B_{t,j}$ if the following condition is satisfied:

$$B_{t,j} = \operatorname*{argmax}_{k \in \{1 \ldots N_t\}} \left| B_{t,k} \cap B_{s,i}^\chi \right|, \tag{22}$$

which means that the jth blob of B_t corresponds to the ith blob of B_s if it has the largest intersection with $B_{s,i}^\chi$. N_t denotes the total number of separate blobs in B_t. An example for the fusion step can be seen in Figure 7, where the original $B_{s,i}$ and the shifted $B_{s,i}^\chi$ shadow maps can be seen for a sample ith shadow blob and the structure map for the corresponding jth blob, together creating a BC building candidate in Figure 7(d). Iterating the fusion step over all shadow blobs gives the opportunity to join broken shadow parts of the same building.

2.2.3. Orthogonality checking of candidate blobs

The fusion step provides evidence for structure blobs with size over the threshold. Smaller structure and shadow blobs may form false candidates with higher probability, as they might indicate noise and might be adjacent only by chance. Also, there might be some structure blobs without any shadow evidence at all. To eliminate false candidates and to find the remaining true ones, we extended the fusion step with a filter function, which is

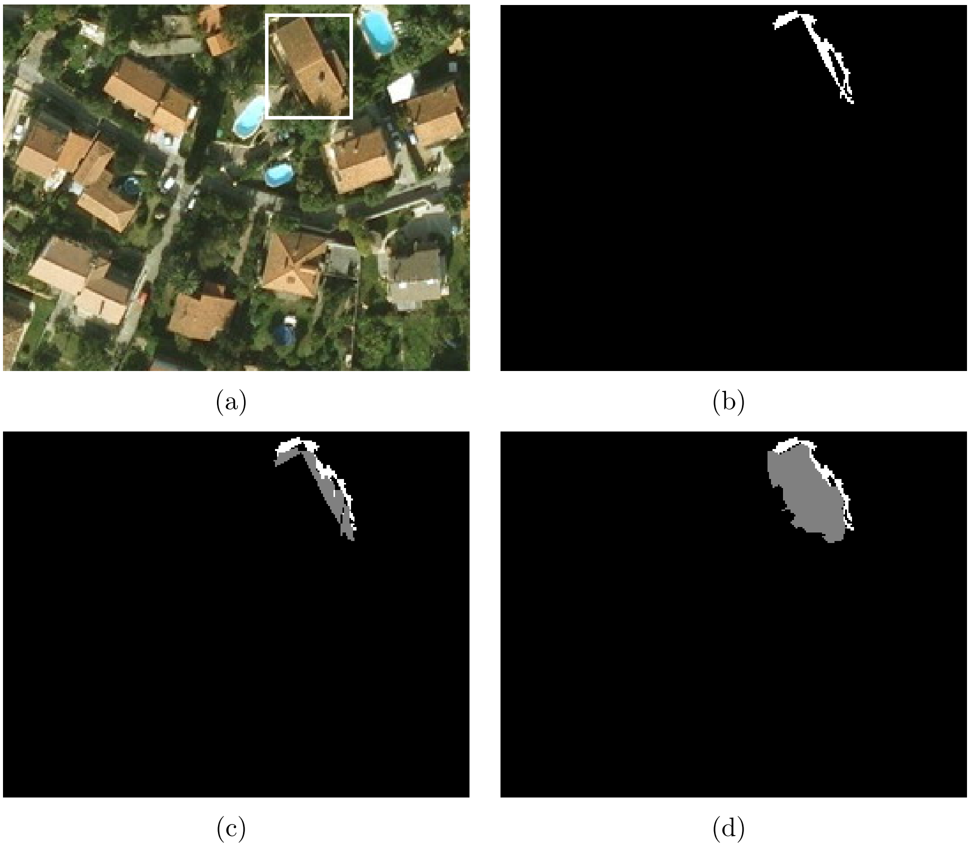


Figure 7: Fusion process: (a) is the original image with the marked sample area; (b) shows the sample shadow blob ($B_{s,i}$), (c) shows the shifted shadow blob ($B^{\chi}_{s,i}$) in gray together with the original $B_{s,i}$ in white; (d) shows the BC building candidate: shadow part (white) together with the corresponding $B_{t,j}$ structure blob (gray).

performed for candidates that have a structure part smaller than 100 pixels. This consideration is for accelerating the process and the threshold is based on the analysis of building sizes in different image resolutions.

The filter function was created to measure the orthogonality of a candidate. We used an approach similar to the local orientation analysis in Section 2.1.2, investigating only the blob area: the $W_n(i)$ window of Eq. 3 is replaced with the total area of the BC building candidate in the I image. The α_{BC} correlation of λ_{BC} local gradient orientation density information of BC to a bimodal Gaussian function is:

$$\alpha_{BC}(m) = \int \lambda_{BC}(\varphi)\eta_2(\varphi, m, d_\vartheta)\ d\varphi, \tag{23}$$

please see Eq. 7 for comparison. The mean value of the η_2 function (Eq. 8) corresponding to the maximum α_{BC} is denoted by m_{BC} in this case.

Previously, in Benedek et al. (2012) the orthogonality of a region was measured by α_{BC} (Eq. 23), but Figure 8 shows that α_{BC} can also have high values for false objects, like road parts. To compensate for such drawbacks of α_{BC}, we have to measure the balance between the two correlated peaks of the bimodal Gaussian function, instead of an overall correlation. Therefore, the correlation for the two peaks ($\alpha_{BC,1}$, $\alpha_{BC,2}$) has to be calculated separately, and compared:

$$\alpha_{BC,1}(m_{BC}) = \int \lambda_{BC}(\varphi)\eta_{2,1}(\varphi, m_{BC}, d_\vartheta)\ d\varphi, \tag{24}$$

$$\alpha_{BC,2}(m_{BC}) = \int \lambda_{BC}(\varphi)\eta_{2,2}(\varphi, m_{BC}, d_\vartheta)\ d\varphi, \tag{25}$$

where $\eta_{2,1}$ and $\eta_{2,2}$ denote the Gaussian components corresponding to η_2. Thus, the orthogonality ratio is defined as:

$$Q_{BC} = \frac{\min(\alpha_{BC,1}, \alpha_{BC,2})}{\max(\alpha_{BC,1}, \alpha_{BC,2})}. \tag{26}$$

According to the Q_{BC} value the blob is either defined as a candidate, or it is supposed to be a false hit and eliminated. As it is shown in Figure 8, Q_{BC} (Eq. 26) provides additional information for α_{BC} about the balance of the correlation and distinguishes building blobs (2. building candidate) and other objects (1. building candidate) more efficiently. When selecting the Q_{BC} value accepted for buildings, we investigated the side ratios of typical buildings. Such objects usually have an elongated shape, with a certain ratio between their widths and lengths. For the general case, we use an acceptance ratio threshold of 0.5.

Figure 9 shows the steps of the building candidate localization. On the left, the resulting candidates of the fusion are presented. We can see that the grayish building in the lower right part of the image is missed by the feature fusion, as its structure information is poor. During the subsequent

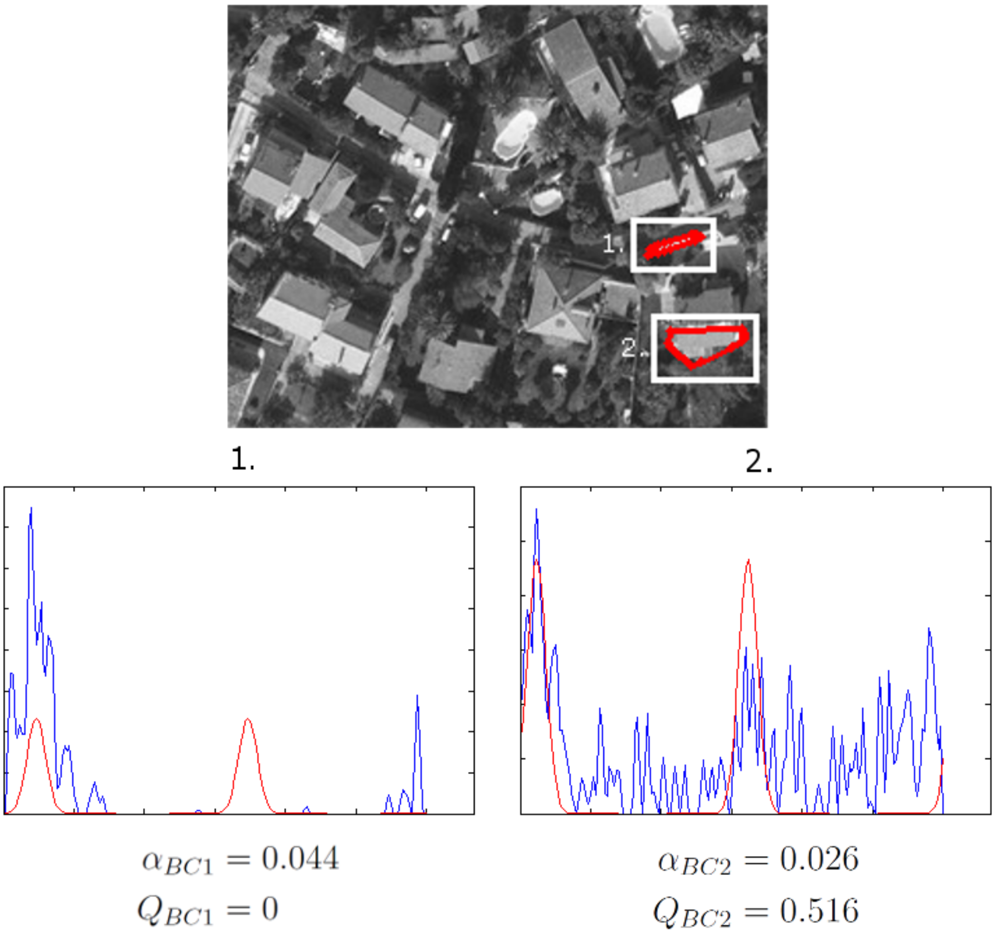


Figure 8: Orthogonality check for candidates. Two selected structure blobs are analyzed: 1. is a road part, 2. is a building part. While the α is high for the 1. candidate, the Q value shows that only one orientation is present, the orthogonality ratio is zero. For the 2. candidate, the α value is lower, but Q indicates the higher rate of orthogonal edges.

orthogonality check, the smaller blobs are also analyzed and due to the high Q_{BC} value (see Figure 8) the building is localized in two parts: one of them is a joint structure and shadow blob (upper), the other is solely a structural hit (lower). The result of the orthogonality blob check is shown in Figure 9(b), where the overall results of the candidate localization process can be seen.

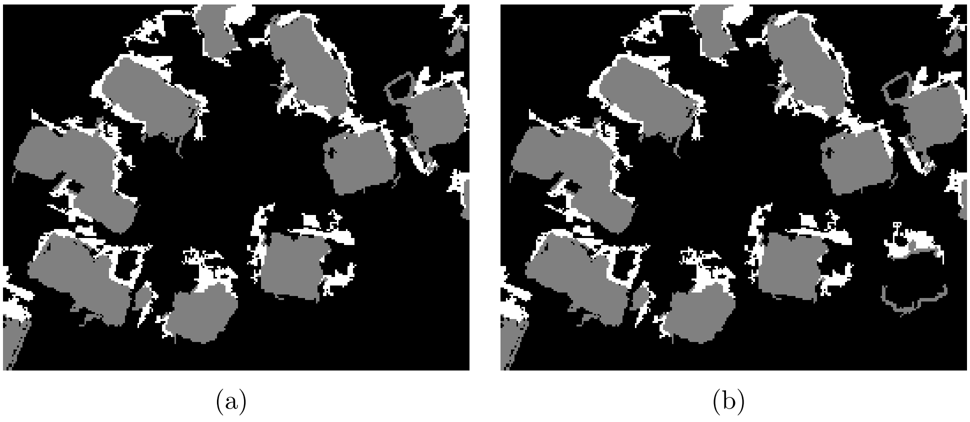


Figure 9: Result of the building candidate localization: (a) is the result after the fusion of shadow and structure blobs, (b) shows the additional candidates as well, given by the orthogonality check.

2.2.4. Building contour detection and shape refinement

After localizing the building candidates, their accurate contour has to be detected. Instead of applying a shape template or only providing a pixel-level location, the present approach estimates the real building outline, by applying the non-parametric Chan-Vese active contour algorithm from Chan and Vese (2001) similarly to Cote and Saeedi (2013). As the application of this iterative technique in not a novel approach for detecting building contours, we will not go into details. The specialty of our method is in the initialization of the Chan-Vese contours: the areas of the candidate blobs are used. For every candidate, the contour is initialized with the convex hull of the structural part's outline (gray in Fig. 9(b)). However, the active contour may result in diverse and cluttered outlines as shown in Figure 10(b), thus, an orientation selective morphological refinement process is introduced as a novel contribution.

To describe this refinement process, we refer to Section 2.1.2 where the main directions of the urban area have been calculated with the IBMGM iterative method. The obtained directionally classified point set for the sample image is shown in the first row in Figure 10. The center of the selected building candidate is in the green area having $[\theta_j, \theta_{o,j}] = [-24, 66]$ dominant directions. An orientation selective morphological operator is created based on the dominant directions characterizing the area: Figure 10(c) shows these

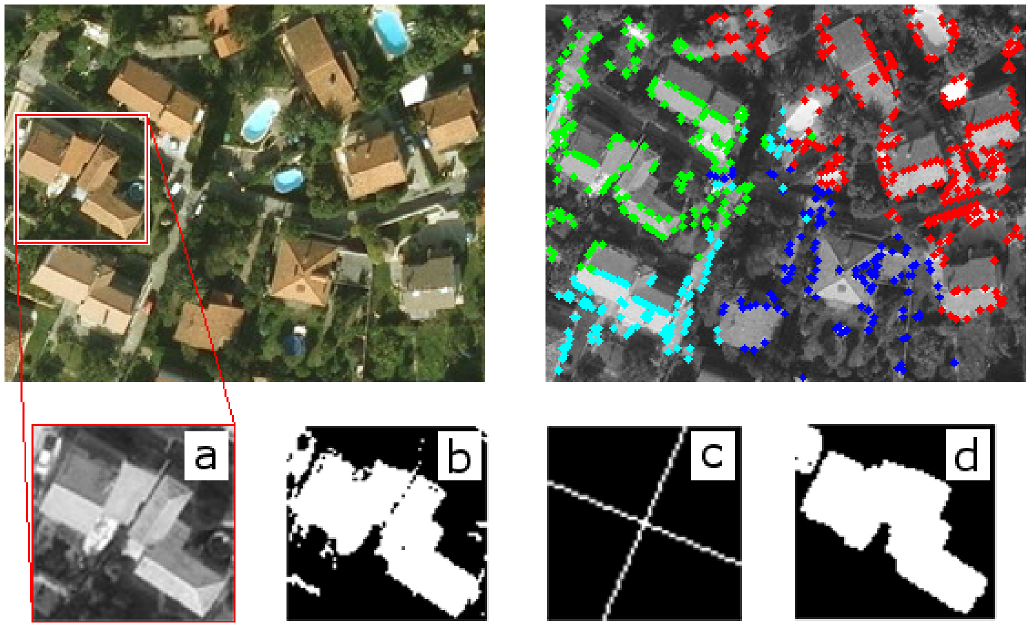


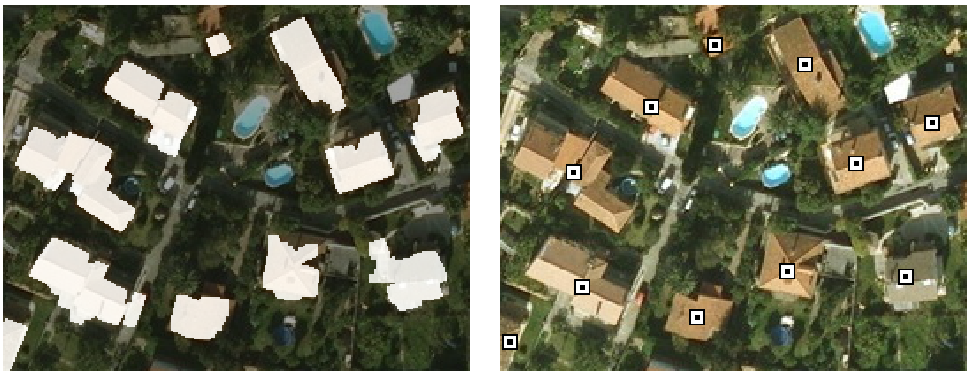
Figure 10: The building detection process: The first row shows the original image with the marked sample area; and on the right the result of the local orientation analysis with different directions marked with different colors. In the second row: (a) shows the sample building candidate area; (b) is the result of the Chan-Vese active contour algorithm; (c) shows the main directions $[-24, 66]$ of the area and (d) is the result of the orientation selective refinement process.

main directions as a joint cross structuring element. During the refinement process, the two orthogonal directions (θ and θ_{o}) are handled separately, both of them represented with a linear element as the two orthogonal lines of the cross (S_θ and S_θ^o). The size of the operator is depending on the image resolution, a detailed analysis is given in Sec. 3.2.

Let AC denote a sample binary area for a building candidate, with pixels inside the contour detected with the Chan-Vese method having values of 1, and 0 outside (see Figure 10(b)). First, the holes of the inner area are filled: for the outer pixels having all the 4-connected neighbors in the inner area (value of 1), the pixel itself is also transferred to the inner area (changed to 1):

$$AC(x \pm 1, y) = 1 \;\cap\; AC(x, y \pm 1) = 1 \;\rightarrow\; AC(x, y) = 1. \tag{27}$$

Then, the hole-filled AC is refined with S_θ and S_θ^o linear structuring



(a) (b)

Figure 11: Result of the building detection: (a) shows the detected contours; (b) defines the estimated locations of the detected buildings.

elements in the following way:

$$AC_{\mathrm{ref}} = \gamma_{S_\theta}\gamma_{S_\theta^o}(AC) \cap \gamma_{S_\theta^o}\gamma_{S_\theta}(AC), \tag{28}$$

where γ is a morphological opening using either S_θ or S_θ^o linear structuring element.

The AC_{ref} refined building area is clear, with smaller noisy blobs eliminated (see Figure 10(d)). The main advantage of the orientation selective refinement process ensures that important edges are preserved, while the contour becomes smoother and more accurate. Moreover, the calculated local orientation information is applied for creating the building-specific structuring element, which means that valuable orientation information is exploited in multiple ways throughout the method. If two different candidates have joint pixels in the calculated final contour, then they are merged into one building object: see the 2nd example in Figure 8 and the related blobs of the same building in Figure 9(b) compared to the result in Figure 11(a). The locations of the buildings are estimated as the centers of the detected blobs. The result of the OSBD process for the sample image is presented in Figure 11.

3. Experimental validation

This section starts with the discussion of interest point detectors performance and parameter estimation, investigating groups of parameters and their behavior. Then, we will continue with the evaluations, with the details of the processed image data sets provided in Table 4. First, an object-level evaluation is performed on the SZTAKI-INRIA Building Detection Benchmark and a comparison is given with five state-of-the-art methods. This data set was designed initially to test the performance of bMBD method and was introduced in Benedek et al. (2012). QuickBird satellite images were provided by *Ali Ozgun Ok* together with pixel-wise ground truth data, used previously to test *GrabCut* in Ok et al. (2013). A pixel-level quantitative evaluation has been performed in the second part of the evaluation and bMBD and *GrabCut* methods were compared with the proposed method. As bMBD performed similarly in the first part of the evaluation as the proposed method, it was also included in the second part along with *GrabCut*. *GrabCut* was designed for multispectral images and required the NIR band, which was not available for the first data set, so we could not include it in the first part of the evaluation.

To show the method's performance on higher resolution, publicly available data set, some test patches (#1, #3, #7, #11, #13, #17, #34) of the Vaihingen data set (Cramer, 2010) were evaluated. As the ground truth classification of the images is also provided with the data set, it was possible to perform a pixel-level evaluation. Finally, we also provide computational time data for the proposed method to show its efficiency.

3.1. Detector performance analysis

To separate the influence of the MHEC feature point detector from the orientation sensitive classification, the first, local orientation analysis step was applied for different, standard interest point detectors on Fig. 5(a). Beside the MHEC, the SIFT, Gabor points used in Sirmacek and Unsalan (2011), the SUSAN, the MSER of Donoser and Bischof (2006) and the SURF of Bay et al. (2008) were compared. The locations of the interest points were extracted by the different detector approaches and the local orientation analysis (Sec. 2.1.2) was carried out on the extracted point sets. After calculating the φ_i main orientation for all feature points the main peaks of the $\vartheta(\varphi)$ orientation histogram (Eq. 5) were investigated. Figure 12 shows the detected θ_i main orientations for the different point detectors. The main directions, generated by the applied MHEC detector, are marked with black horizontal lines to help the comparison. As it is shown, the main directions are similar for each point detector, in some cases extra directions are present, but the main characteristics are constant. Using this information, the improved,

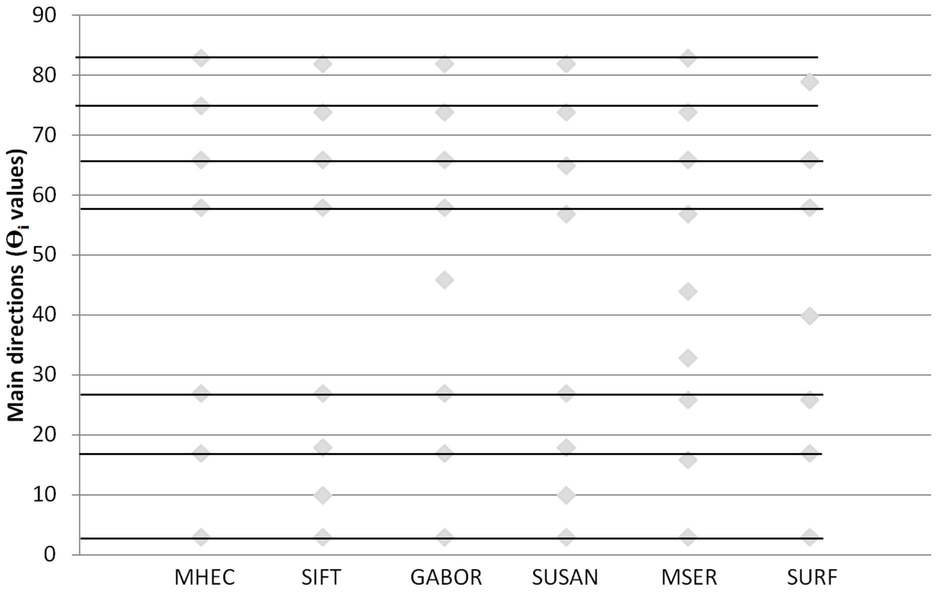


Figure 12: Main directions of the Fig. 5(a) image: main peaks of $\vartheta(\varphi)$ orientation histogram for different interest point detectors. The main directions are almost insensitive to the detector, resulting in a constant improved edge map.

MFC-based edge map (Sec. 2.1.3) is not sensitive to the applied detector. The motivation for using MHEC is its efficiency for representing the urban areas in Kovacs and Sziranyi (2013).

3.2. Parameter settings

The parameters of the introduced OSBD method can be divided into three main groups. The first group includes the threshold parameters for binarization, typically when creating binary feature maps as B_c, B_{sh} and they are calculated with the adaptive Otsu method (Otsu, 1979), which is proved to be robust and reliable in many state-of-the-art works. These thresholds are connected to the actual image, and are calculated separately for every sample. Selection of optimal values for other parameters are described in this section; Table 1 summarizes the parameters, their settings and influence in the different processing steps.

The second group includes parameters depending on the image resolution and accordingly on the expected size of the building objects. Such parameters are the sizes of orientation-selective morphological structuring elements (S_{χ^-},

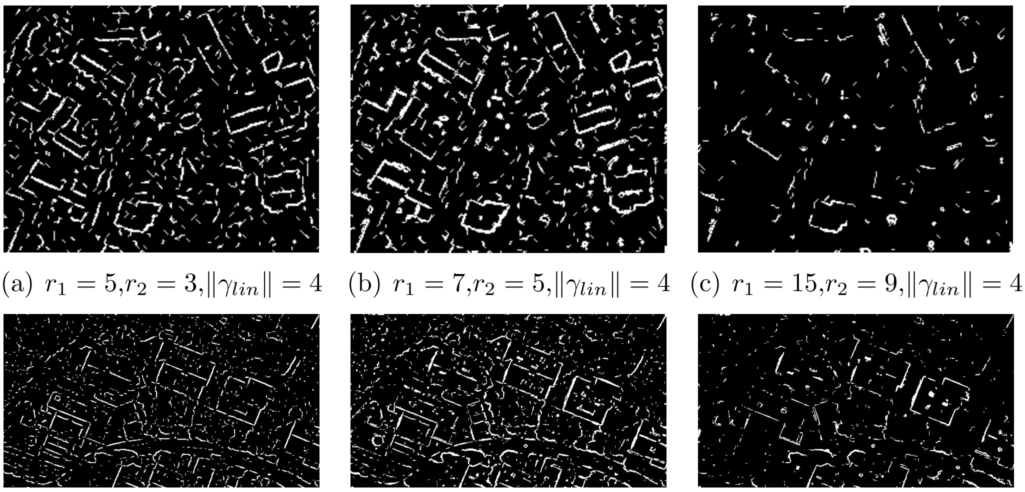
Processing step	Parameter	Setting	Influence
Local orientation analysis (Sec. 2.1.2)	overall correlation (α_{th})	0.04	IBMGM algorithm main directions
	CPR threshold (ϵ)	0.9	
Edge map (Sec. 2.1.3)	MFC SE (r_1)	7	improved, MFC-based orientation selective edge map
	MFC SE (r_2)	5	
	linear SE (γ_{lin})	4	
Fusion of structure & shadow (Sec. 2.2.2)	fusion SE (S_{χ^-})	13	illumination direction selective fusion
	min. /max. shadow size	adaptively	filter shadow blobs
Orthogonality check (Sec. 2.2.3)	orthogonality threshold (Q_{BC})	0.5	filtering false building candidates
Building contour detection and shape refinement (Sec. 2.2.4)	orthogonal SE (S_θ)	adaptively	orientation selective morphological shape refinement

Table 1: Overview on parameters, their setting and influence in the processing steps.

S_θ) and the thresholds for blob sizes in the fusion and detection steps. The used values for different parameters are marked in the related sections of the paper.

The MFC operator is responsible for creating the improved, orientation selective edge feature (Sec. 2.1.3). It is based on a morphological opening and closing, where the structuring elements (SE) should be chosen according to the sizes of the background structure and the feature to be extracted. Following the recommendations of Zingman et al. (2014), results of a few different parameter settings are shown in Figure 13 for the original image in Figure 5(a) and a part of Area17 from the Vaihingen data set. The analysis shows that if the SE of opening and closing are chosen too small, many background structures remain in the edge map (Fig.13(a) and Fig.13(d)); on the contrary, too large values cause the loss of important information (Fig.13(c) and Fig.13(f)). Thus, for both data sets, $r_1 = 7$ and $r_2 = 5$ values are applied in the evaluation. The length of the linear SE in γ_{lin} denoted by $\|\gamma_{lin}\|$ is also tested for $\|\gamma_{lin}\| = 4$ and $\|\gamma_{lin}\| = 10$ values. However, in the latter case, very limited number of structures is extracted for both images, which may cause inaccurate detection results. Therefore, only the results for $\|\gamma_{lin}\| = 4$ are shown in Fig. 13.

The S_{χ^-} is the morphological operator (Sec. 2.2.1), representing the illumination direction in the fusion of the structure and the shadow features. The size of the linear element in the kernel is analyzed for the optimal setting, testing multiple values. For each case, the complete algorithm was executed



(a) $r_1 = 5$,$r_2 = 3$,$\|\gamma_{lin}\| = 4$ (b) $r_1 = 7$,$r_2 = 5$,$\|\gamma_{lin}\| = 4$ (c) $r_1 = 15$,$r_2 = 9$,$\|\gamma_{lin}\| = 4$

(d) $r_1 = 5$,$r_2 = 3$,$\|\gamma_{lin}\| = 4$ (e) $r_1 = 7$,$r_2 = 5$,$\|\gamma_{lin}\| = 4$ (f) $r_1 = 15$,$r_2 = 9$,$\|\gamma_{lin}\| = 4$

Figure 13: Analysis of the MFC-based edge feature operator. Different parameters are tested for two sample images from different data sets: First row is the sample image shown in Figure 5(a); Second row: A part of Area17 from the Vaihingen data set. Results show that in both cases $r_1 = 7$ and $r_2 = 5$ values are the most efficient.

and the overall F-score performance was measured on pixel-level.

$$P = \frac{\mathrm{TD}}{\mathrm{TD} + \mathrm{FD}}, \quad R = \frac{\mathrm{TD}}{\mathrm{TD} + \mathrm{MD}}, \quad F = 2 \cdot \frac{P \cdot R}{P + R}, \tag{29}$$

where TD, FD and MD denote the number of true detections (true positives), false detections (false positives) and missed detections (false negatives) respectively; P stands for precision, R for recall.

To measure the performance on pixel-level, two images with ground truth data were chosen: image2 from *GrabCut* data set (Ok et al., 2013) and Area17 from the Vaihingen data set. The performance remains constant for different parameter values, included in Table 2, which means that the algorithm is not sensible to the size of the S_χ- morphological operator. In the overall evaluation the size of the linear element in the operator was set to 13.

To justify the selected maximum and minimum blob sizes for structure and shadow, the building sizes in the different databases are analyzed in Table 3. This shows that the average building sizes in SZTAKI-INRIA and *GrabCut* databases are very similar, therefore the same limits are applied: minimum shadow blob is 20 pixels, maximum shadow blob is 2000 pixels.

Linear element of S_{χ^-} (pixel)	F-score	
	GrabCut image2	Vaihingen Area17
13	**93.5%**	**85.1%**
23	93.5%	85.1%
33	93.5%	85.1%
43	93.5%	84.9%

Table 2: Overall performance of the method for different S_{χ^-} sizes. The analysis is performed for two samples from different data sets. Results show that the method is not sensible to this parameter.

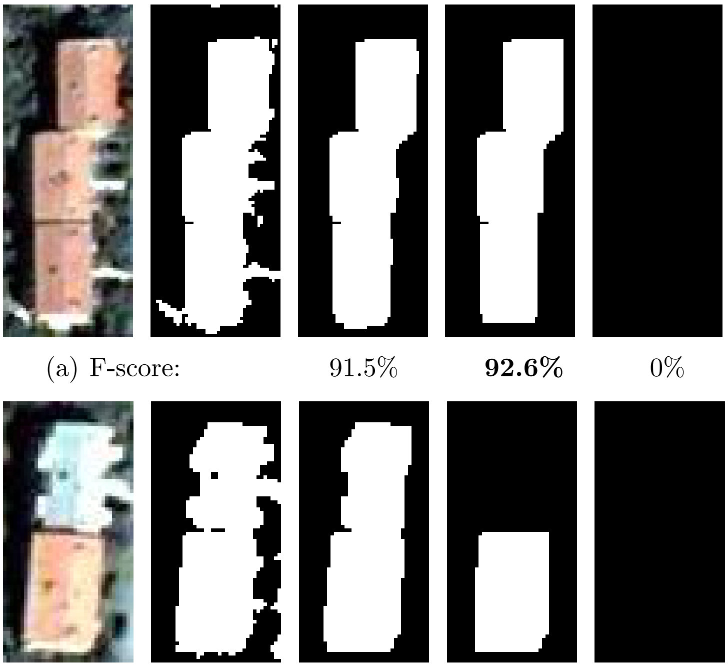
When selecting the maximum threshold for shadow blob, it is also taken into consideration, that neither SZTAKI-INRIA, nor *GrabCut* database contains images with elongated, large shadow blobs. In case of Vaihingen database, the resolution is higher and the sizes of buildings are larger and more varied, so 100 and 10000 minimum and maximum values are used in the evaluation.

To refine the final shape of building objects, an orientation selective morphological refinement process is proposed in Sec. 2.2.4 with S_θ (and orthogonal S^o_θ) structuring element. Different buildings were selected from *GrabCut* database image2 (Fig. 14) and Vaihingen Area17 (Fig. 15) images to test the length of the linear element in the SE (denoted by $\|S_\theta\|$). Different values are tested for $\|S_\theta\|$, 7, 17 and 27 for *GrabCut* database image2; 7, 17, 27 and 37 for Vaihingen Area17. In each case, the overall F-score is given for the building blob. Results show that the $\|S_\theta\|$ is depending on the image resolution: too small values may not cause effective refinement, too large values may clear real building parts. Moreover, proper selection of $\|S_\theta\|$ may also eliminate falsely detected building-like objects (typically cars). In the evaluation process $\|S_\theta\| = 7$ is used for SZTAKI-INRIA and *GrabCut* databases, $\|S_\theta\| = 27$ is applied for Vaihingen test patches.

The third group, including some local orientation analysis parameters (see Section 2.1.2), is defined in a training process. The main point of the

Data Set	Building sizes (pixel)	Mean size (pixel)
SZTAKI-INRIA	348–4186	912
Grabcut	31–6165	1123
Vaihingen	93–52098	7131

Table 3: Occurring building sizes in the databases for the selection of optimal blob sizes.



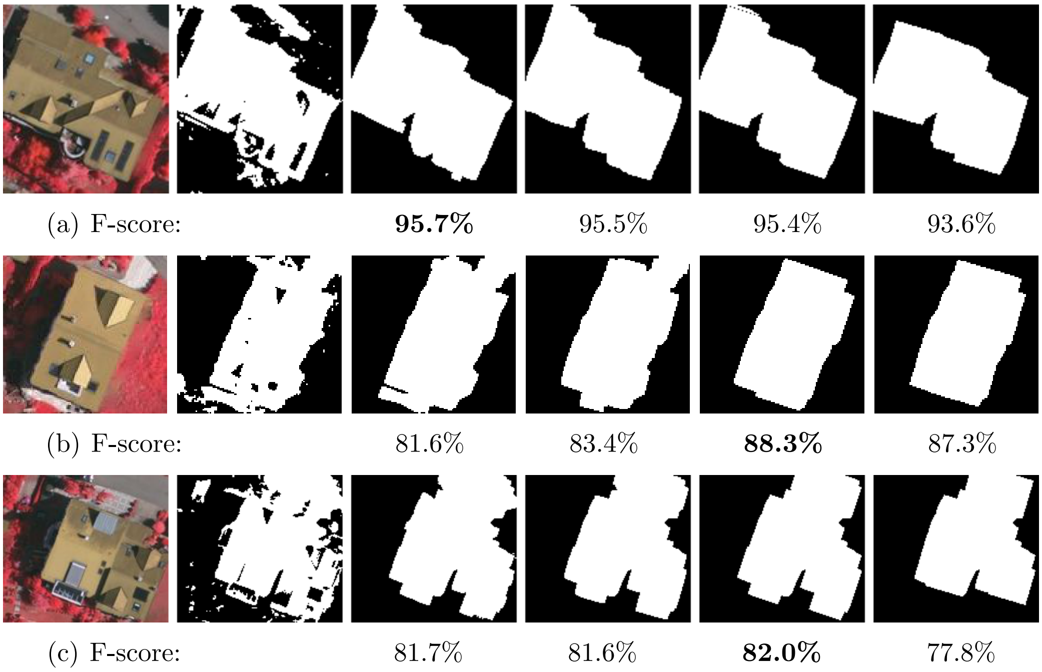
(a) F-score: 91.5% **92.6%** 0%

(b) F-score: **95.1%** 68.6% 0%

Figure 14: Shape refinement structuring element analysis for selected buildings: first column: original images; second column: preliminary detection result with Chan-Vese method; Refinement process with different $\|S_\theta\|$ sizes (in pixel): third column $\|S_\theta\| = 7$; fourth column $\|S_\theta\| = 17$; fifth column $\|S_\theta\| = 27$.

parameter tuning in these cases is to find such values which balance between accuracy and efficiency. The first parameter in the analysis is the n window size indicating the neighborhood size around a feature point, where local orientation is defined. Tests with different n sizes have been performed for the test set in Figure 2(a), which is shown in Figure 16. While the main characteristics of the $\vartheta(\varphi)$ function remains similar using the larger $n = 15$ value (compared to the smaller $n = 5$), the $\vartheta(\varphi)$ orientation histogram becomes less noisy. Therefore, $n = 15$ parameter was applied for extracting local gradient orientation.

The parameters of the IBMGM process are also analyzed: we have to define the value of α_{th} and ϵ in Algorithm 1. The α_{th} is the lowest correlation rate, ϵ is the lowest CPR which we accept to stop the iterative process. To tune these parameters, Figure 17 shows the results for 2 images used for



(a) F-score: **95.7%** 95.5% 95.4% 93.6%

(b) F-score: 81.6% 83.4% **88.3%** 87.3%

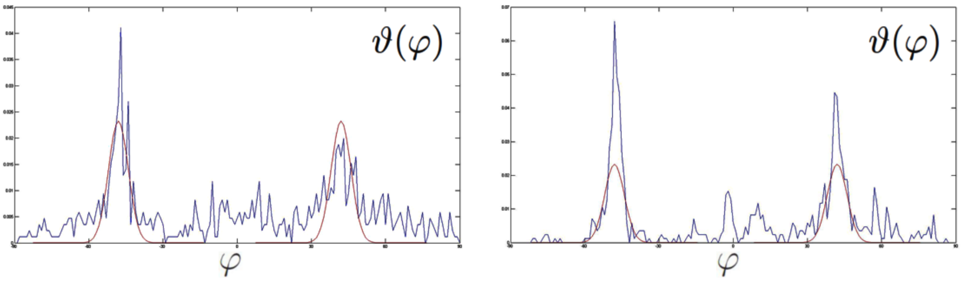
(c) F-score: 81.7% 81.6% **82.0%** 77.8%

Figure 15: Shape refinement structuring element analysis for selected buildings: first column: original images; second column: preliminary detection result with Chan-Vese method; Refinement process with different $\|S_\theta\|$ sizes (in pixel): third column $\|S_\theta\| = 7$; fourth column $\|S_\theta\| = 17$; fifth column $\|S_\theta\| = 27$; sixth column $\|S_\theta\| = 37$.

training. CPR_j is shown in blue and $10 \cdot \alpha_j$ in red, for every iteration. While CPR_j is constantly growing, α is typically having some local maxima. After analyzing the behavior of both parameters, $\epsilon = 0.9$ has been selected for CPR threshold and $\alpha_{\text{th}} = 0.04$ (in Figure 17 0.4 is marked for $10 \cdot \alpha_{\text{th}}$ in red) for the lowest acceptable rate of correlation to stop the IBMGM. The selected parameters for α_j and CPR_j are applied for every test set in the evaluation.

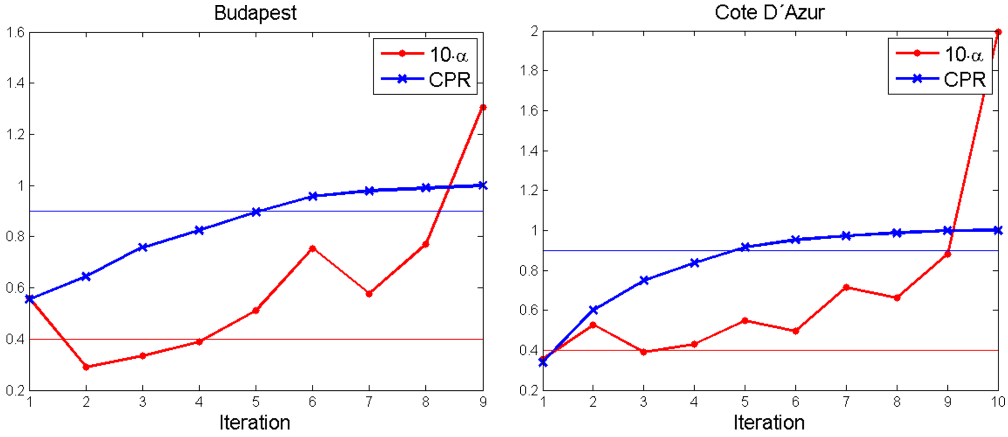
3.3. Object-level evaluation

Multiple quantitative experiments have been carried out on different image databases. In the first case, an object level evaluation was performed on a large data set, including images from different test sites: Budapest, Szada (both in Hungary); Cote d'Azur, Bodensee and Normandy, altogether with 453 building objects. The sources for the images are various: Budapest



(a) $n = 5$ (b) $n = 15$

Figure 16: Local neighborhood size analysis for Image 2(a): the calculated $\vartheta(\varphi)$ functions for different n neighborhood sizes. Results show that while the main characteristics of the function remain similar, the bigger the chosen n, the less noisy the resulting function.



(a) (b)

Figure 17: Iterational behavior of IBMGM parameters for different images together with the selected thresholds: blue indicates CPR, red marks $10 \cdot \alpha$.

aerial images were provided by the City Council, Szada aerial images are provided by Hungarian Institute of Geodesy, Cartography and Remote Sensing (FÖMI); Cote d'Azur, Bodensee and Normandy are satellite images acquired from the Google Earth. Database details are summarized in Table 4. These databases were also used in Benedek et al. (2012) for evaluation. Figure 18 shows the object level quantitative comparison for the data set. Each column shows the ratio of false negative/missed (lower, darker) and false

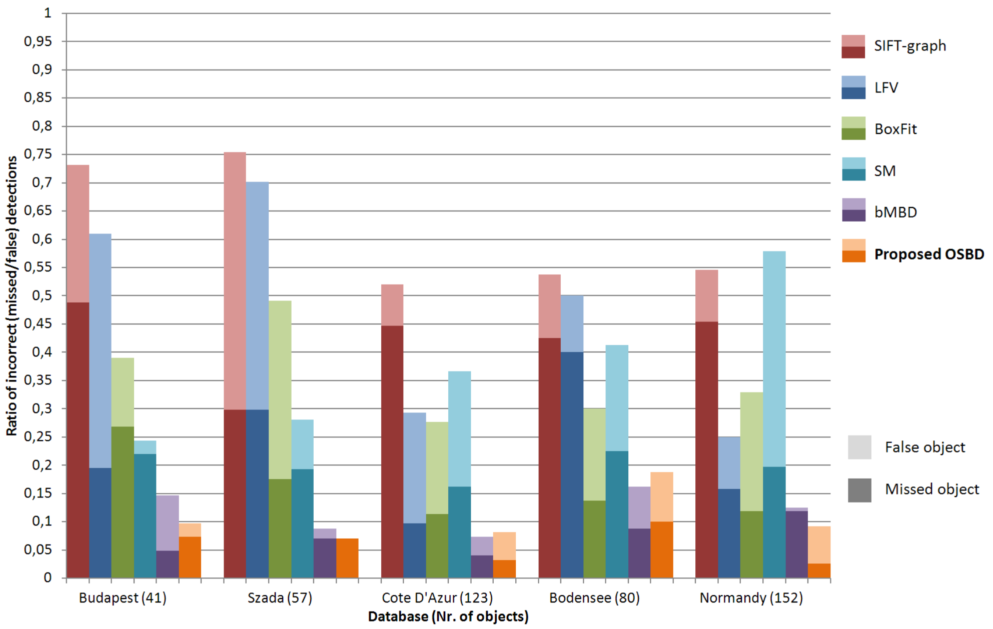
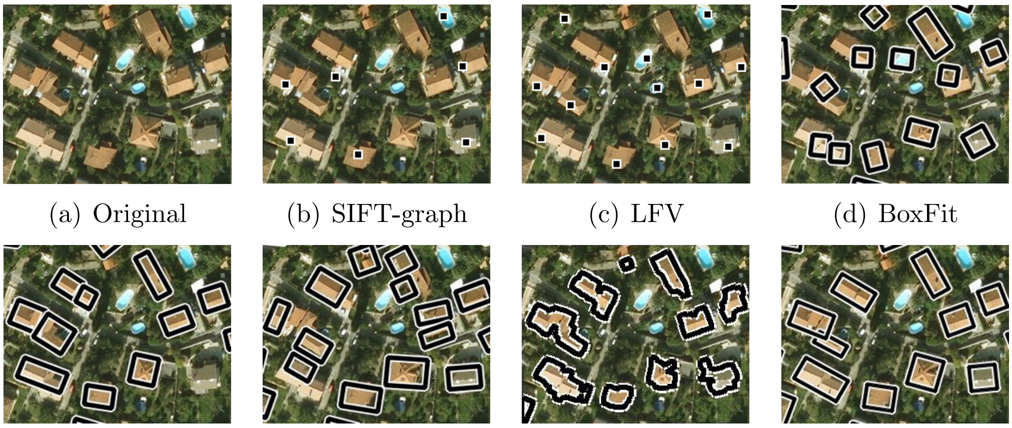


Figure 18: Object level quantitative comparison with state-of-the-art methods. Different patterns denote different approaches (see Sec. 1). Results are evaluated on 5 data sets. The ratio of false and missed objects is shown.

positive/false (upper, lighter) detected objects (including false multiple detections) for each respective database. Image sizes are changing from 0.3 to 1.5 megapixel.

Evaluation results show that the proposed OSBD method is able to outperform the compared approaches producing the lowest number of mistaken objects in nearly all test cases. To give an extended discussion about the performance of the compared methods, a sample result from the *Cote d'Azur* data set is shown in Figure 19 for all methods. The selected site includes some challenging objects: buildings with gray rooftops and varying shapes. The detection results show that some of the compared methods are not able to cope with the difficult objects. Moreover, the remaining methods, while handling the variations in rooftop colors well, cannot accurately detect the scalloped, elongated building in the left part of the image. To aid the visual inspection of the results, the ground truth is also shown in Figure 19(h). It should be mentioned, that the ground truth was created for the bMBD method, therefore the elongated building is covered with multiple rectangle-



(a) Original (b) SIFT-graph (c) LFV (d) BoxFit

(e) SM (f) bMBD (g) Proposed (h) Ground Truth

Figure 19: Sample result for the sample part of Cote d'Azur data set. The detection result for different methods are shown separately: (a) is the original image part; (b) SIFT-graph, Sirmacek and Unsalan (2009); (c) LFV, Sirmacek and Unsalan (2011); (d) BoxFit, Sirmacek and Unsalan (2008); (e) SM, Song et al. (2006); (f) bMBD, Benedek et al. (2012); (g) proposed OSBD; (h) is the ground truth created for bMBD with rectangle templates.

shaped objects.

The SIFT-graph method is sensitive to the selected building template. When the limited number of templates are not enough to represent all building variations in the images, the detection accuracy decreases.

The LFV is not sensitive to altering building characteristics, but features in the background can cause false positive detections. Both SIFT-graph and LVF perform better for the satellite images. These methods are optimized for less densely situated buildings, which is a disadvantage in city regions.

The BoxFit method is hardly able to find accurate building outlines, despite localizing objects correctly. As the method strongly depends on color and shadow information, the lack of these features results in missed detections. Background blobs that are similar to buildings (like pools) can also generate false positive hits. The shape detection step is not able to find complex building contours either.

Similarly, SM is also sensitive to insufficient color and texture information. It is also unable to detect inhomogeneous objects, due to the method's main hypotheses about homogeneity. The *Cote d'Azur* test site has gray colored buildings, surrounded by dark green vegetation (see the building in the lower-

Data set	#obj.	Source	Type	Resolution	Image size(s)
Budapest	41	City Council	Aerial	0.5 m/pixel	600 × 490
Szada	57	FÖMI	Aerial	0.5 m/pixel	800 × 600
					1076 × 444
					996 × 588
Cote d'Azur	123	Google Earth	HR optical satellite	2.5 m/pixel	800 × 473
					800 × 455
Bodensee	80	Google Earth	HR optical satellite	2.5 m/pixel	910 × 618
Normandy	152	Google Earth	HR optical satellite	2.5 m/pixel	1437 × 814
GrabCut	230	QuickBird	VHR optical satellite	0.61 m/pixel	367 × 325
					382 × 393
					524 × 539
					506 × 490
					827 × 624
					550 × 416
Vaihingen	306	RWE Power	Aerial	0.09 m/pixel	1919 × 2569
					2006 × 3007
					1887 × 2557
					1893 × 2566
					2818 × 2558
					2336 × 1281
					1388 × 2555

Table 4: Details of image datasets used for object-level validation.

right part of Fig. 19(a)), which proved to be a great challenge for this method. Additionally, complex building outlines are often detected as multiple hits.

The bMBD method also prefers homogeneous building objects, therefore it is difficult for it to detect partially shadowed rooftops. As it handles a building candidate as a rectangular region, the features inside the region are calculated in the energy term. This means that the method also has problems with diverse building shapes, like the scalloped one in the left part of Figure 19(a).

The proposed OBSD method localizes the buildings correctly, therefore the object-level result is of high quality. Results show that the OSBD method is able to outperform the compared approaches and produces the lowest number of mistaken objects in nearly all test cases. However, two issues might occur: sometimes objects might be reduced and only a smaller building part will be detected; occasionally, parts of the surrounding vegetation might be covered by the object outline. These issues are mainly caused by the variation reducing behavior of the Chan-Vese method.

3.4. Pixel-level evaluation

The second part of the evaluation was a pixel-level evaluation. First, a comparison was performed with 2 state-of-the-art methods, bMBD and *GrabCut* from Ok et al. (2013). The data set was kindly provided by *Ali Ozgun Ok* and it was also evaluated in Ok et al. (2013) for the *GrabCut* method. Therefore, the detection results for *GrabCut* were taken from Ok et al. (2013) directly. The data set contains QuickBird satellite images, including four multispectral bands (R, G, B, and NIR) with a radiometric resolution of 11 bits per band, the images are selected to represent diverse urban area and building characteristics with varying illumination conditions. The data set was provided along with ground truth data, which was produced manually by a qualified human operator.

The overall performance of different techniques was measured by the F-score. Pixel-level quantitative evaluation results can be seen in Table 5 and a corresponding result for the first image of the data set (image1) is shown in Figure 20, where true positives are green, false negatives are blue and false positives are red.

bMBD shows slightly decreased performance compared to the other two approaches which is due to complex building shapes and the lack of color and shadow components in some cases. Figure 20(b) shows that the method

Database	Performance								
	bMBD			GrabCut			Proposed OSBD		
Image	F	R	P	F	R	P	F	R	P
image1	86.6%	87.7%	85.5%	88.1%	89.4%	86.9%	**94.0%**	91.4%	96.7%
image2	80.7%	78.6%	83.0%	89.1%	93.6%	85.0%	**93.5%**	91.2%	95.9%
image3	82.6%	81.0%	84.3%	90.4%	93.5%	87.5%	**90.6%**	87.6%	93.9%
image4	72.5%	90.7%	60.3%	**92.4%**	95.8%	89.3%	91.4%	87.1%	96.0%
image5	62.9%	72.6%	55.5%	81.1%	89.2%	74.4%	**81.8%**	74.6%	90.5%
image6	67.3%	78.9%	58.6%	75.9%	95.9%	62.8%	**77.4%**	66.7%	92.3%
Average	75.4%	81.6%	71.2%	86.2%	**92.9%**	81.0%	**88.1%**	83.1%	**94.2%**

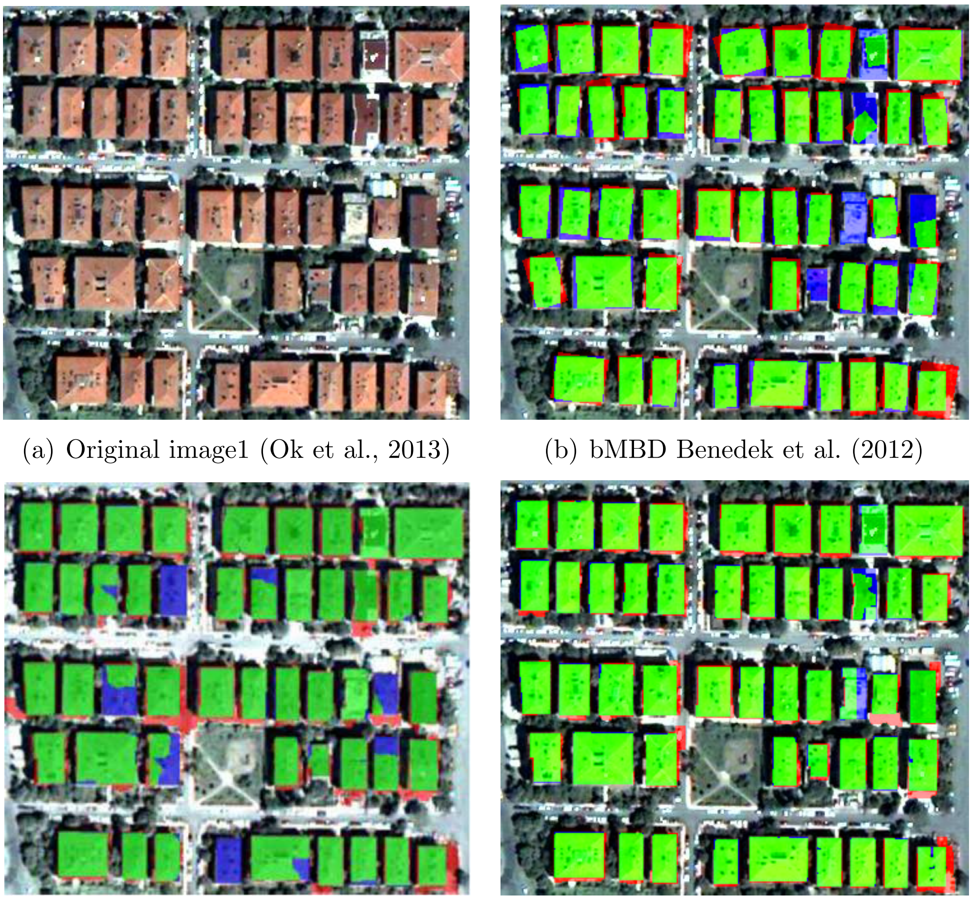
Table 5: Quantitative results of the pixel-level evaluation step for bMBD, Benedek et al. (2012); GrabCut, Ok et al. (2013) and the OSBD method.

has problems with the detection of the white building (third row) and the inhomogeneous rooftop (second row).

GrabCut performs with high accuracy; since it is based on shadow detection, buildings without such features are generally missed. Moreover, the *GrabCut* process misses building parts more often than the proposed OBSD method that uses a refinement step based on the combination of Chan-Vese contours and morphological shape refinement. The advantage of *GrabCut* is the high recall value, which means a lower number of missed detections. It is also interesting to mention, that *GrabCut* and OSBD often detect the same false positives, due to the active contour based final outline detection (see the building in the lower right corner in Figure 20).

OSBD exploits the advantages of orientation as a novel feature, therefore irrelevant features can be eliminated more efficiently. It is also able to detect building objects without shadow information, using only orientation selective edge features, which is the main reason for the performance improvement. However, the active contour based building outline detection still has its disadvantages: sometimes inhomogeneous buildings are only partially detected and similar surroundings can result in false positive detections. Overall, we can say that due to the novel contributions of the proposed OSBD algorithm, our method is able to outperform the other two state-of-the-art approaches.

To measure the method's performance objectively, a few test patches of the Vaihingen data set (Cramer, 2010) have been evaluated. It is important to mention, that only the true orthophoto mosaics of the test set were used



(a) Original image1 (Ok et al., 2013)

(b) bMBD Benedek et al. (2012)

(c) GrabCut (from Ok et al. (2013))

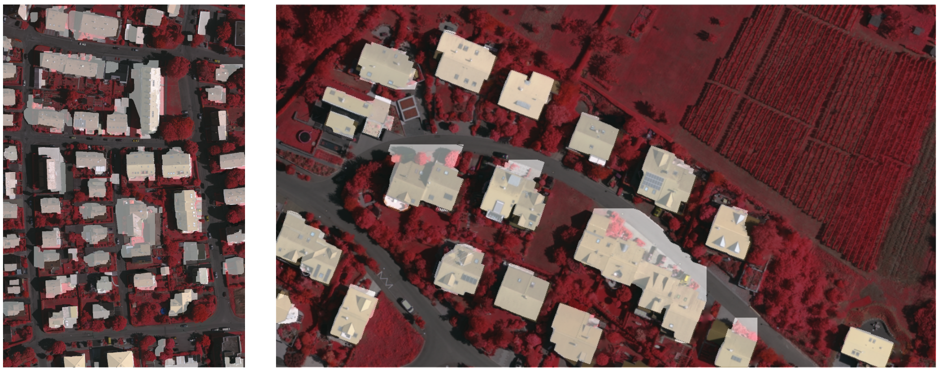
(d) Proposed

Figure 20: Sample result of the building detection for image1: (a) is the original image1; (b)–(d) are the detection results for bMBD Benedek et al. (2012), GrabCut Ok et al. (2013) and the proposed OSBD. True positives are green, false negatives are blue and false positives are red.

for the detection, neither the provided ALS, nor the DSM data. Moreover, as this data set has IR, R and G color channels, the color conversion and application of CIE Luv colorspace cannot be performed effectively. For this reason, only 7 mosaics of the test set was used for evaluation. The results in Table 6 show that the algorithm is still able to achieve a fine detection rate, and its advantage is the ability of detecting various building shapes efficiently (Fig. 21).

Vaihingen database	Performance		
	F	R	P
Area1	59.5%	73.2%	50.0%
Area3	70.4%	67.8%	73.3%
Area7	68.5%	63.7%	74.2%
Area11	58.5%	54.2%	63.5%
Area13	73.1%	71.8%	74.4%
Area17	85.1%	78.9%	92.4%
Area34	71.7%	72.9%	70.5%
Average	69.5%	68.9%	71.2%

Table 6: Detection results for selected true orthophoto mosaics of the Vaihingen dataset.



(a) Area3

(b) Area17

Figure 21: Sample test results for the Vaihingen dataset, original and detection result images are merged, the results are shown as lighter blobs.

To analyze and compare the method's performance, Table 7 shows the computational times for each data set included in the object-level evaluation step (Figure 18). As the final, active contour based detection step is the most computationally intensive step of the method, the candidate localization and contour detection part is also indicated separately beside the overall execution time. Our tests were performed on a PC with an Intel(R) CoreTM i7 2.67GHz CPU with 4 GB RAM using Matlab R2014a. By comparing the execution times to other state-of-the-art methods from Benedek et al. (2012), the speed of our algorithm falls in the middle, meaning comparable efficiency.

Data Set	Computational time (seconds)		
	Localization	Detection	Overall
Budapest	11.0	7.4	18.4
Szada	38.9	14.6	53.5
Cote D'Azur	37.5	28.8	66.3
Bodensee	18.7	10.4	29.1
Normandy	109.6	45.1	154.7

Table 7: Computational time for OSBD method (candidate localization + contour detection) for the database in Figure 18.

Based on Benedek et al. (2012), LFV is proclaimed to be the most efficient from a computational complexity point-of-view with an average run time of $42s$ per image for the 5 used data sets, followed by the SM method with $52s$. Calculating the average computational time for the proposed method, it takes $64s$ for one image, showing that the proposed method is competitive with the compared techniques. Considering the data sets separately, for the *Normandy* data set our method is a bit slower, as the large number of building candidates cause an increased detection time; nevertheless for the other 4, the computational time is in the mid-range. Further considerable speedup could be achieved by an optimized multi-threaded C++ implementation.

As Ok et al. (2013) is only providing average computation times, in the pixel-level evaluation phase we have also calculated an average processing time for one image of the test data, which was 16.2 seconds. This value was the result of a lower number of building candidates in the image and richer feature information, which meant that the final building contour was extracted by a lower number of iterations in the active contour method. Due to the strong low level features, the bMBD method was also performing at a high speed, resulting in an average of 8 seconds per image. However, as it is shown in Table 5, the accuracy is a bit lower, than for the two other methods. Practically, this means that ordinary buildings (red roofs with strong gradient and shadow features) are detected very fast, but extraordinary ones are missed. Both bMBD and OSBD perform fairly well compared to the average time of 23 seconds of the *GrabCut* algorithm, even considering the fact that *GrabCut*'s performance was tested on a Core i5 2.6 GHz CPU.

4. Conclusion

We have proposed an orientation selective building detection framework for aerial images, introducing orientation as a novel feature for object extraction purposes. The algorithm starts with feature point detection, used as a directional sampling set to compute orientation statistics and to define the dominant directions of the urban area. The orientation information is then applied to create a novel improved edge map, emphasizing edges only in the main directions. By integrating color, shadow and the improved edge features, and using the illumination information, building candidates are localized. To find the remaining candidates with limited feature evidence, an orthogonality check is introduced. The contours of the localized candidates are extracted by the Chan-Vese active contour algorithm, which might result in diverse, yet less accurate contours. To compensate for this, a novel orientation-selective morphological operator is introduced to refine the final outlines. The extensive object- and pixel-level quantitative evaluation and comparison with six state-of-the-art methods confirm and support the superiority of the introduced approach.

The present work can be improved in the future by a C++ implementation, considering optimization and multi-threaded design, concentrating especially on the active contour calculation, which is the most computationally intensive part. The exploitation of orientation information looks very promising and it can be extended for different remote sensing applications: it can provide a new perspective for feature based building detection algorithms, for application in building verification besides shadow, color and texture information. Moreover, the orientation based feature concept can be a novel trend in object detection, especially in such applications where the object has a specified shape (like traffic sign detection).

Acknowledgments

This work was partially supported by the Hungarian Scientific Research Fund under Grant No. OTKA #106374 and the European Space Agency DUSIREF project. Test images were provided by the Hungarian Institute of Geodesy, Cartography and Remote Sensing (FÖMI) (*Szada* images) and by András Görög (*Budapest* images). The authors would like to thank Ali Ozgun Ok and Csaba Benedek for providing the test data sets and for their help in the evaluation of the reference methods. The Vaihingen

data set was provided by the German Society for Photogrammetry, Remote Sensing and Geoinformation (DGPF) (Cramer, 2010): http://www.ifp.uni-stuttgart.de/dgpf/DKEP-Allg.html.

References

Aksoy, S. and Cinbis, R. G. (2010). Image mining using directional spatial constraints. *IEEE Geoscience and Remote Sensing Letters*, 7(1):33–37.

Bay, H., Ess, A., Tuytelaars, T., and Van Gool, L. (2008). Speeded-up robust features (SURF). *Computer vision and Image Understanding*, 110(3):346–359.

Benedek, C., Descombes, X., and Zerubia, J. (2012). Building development monitoring in multitemporal remotely sensed image pairs with stochastic birth-death dynamics. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(1):33–50.

Bigun, J., Granlund, G. H., and Wiklund, J. (1991). Multidimensional orientation estimation with applications to texture analysis and optical flow. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(8):775–790.

Canny, J. (1986). A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 8(6):679–698.

Chan, T. F. and Vese, L. A. (2001). Active contours without edges. *IEEE Transactions on Image Processing*, 10(2):266–277.

Cote, M. and Saeedi, P. (2013). Automatic rooftop extraction in nadir aerial imagery of suburban regions using corners and variational level set evolution. *IEEE Transactions on Geoscience and Remote Sensing*, 51(1):313–328.

Cramer, M. (2010). The DGPF-test on digital airborne camera evaluation–overview and test design. *Photogrammetrie-Fernerkundung-Geoinformation*, 2010(2):73–82.

Cui, S., Yan, Q., and Liu, Z. (2008). Right-angle building extraction based on graph-search algorithm. In *International Workshop on Earth Observation and Remote Sensing Applications*, pages 1–7.

Cui, S., Yan, Q., and Reinartz, P. (2012). Complex building description and extraction based on Hough transformation and cycle detection. *Remote Sensing Letters*, 3(2):151–159.

Donoser, M. and Bischof, H. (2006). Efficient maximally stable extremal region (MSER) tracking. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, pages 553–560.

Harris, C. and Stephens, M. (1988). A combined corner and edge detector. In *Alvey Vision Conference*, pages 147–151.

Huertas, A. and Nevatia, R. (1988). Detecting buildings in aerial images. *Computer Vision, Graphics, and Image Processing*, 41(2):131–152.

Izadi, M. and Saeedi, P. (2012). Three-dimensional polygonal building model estimation from single satellite images. *IEEE Transactions on Geoscience and Remote Sensing*, 50(6):2254–2272.

Katartzis, A. and Sahli, H. (2008). A stochastic framework for the identification of building rooftops using a single remote sensing image. *IEEE Transactions on Geoscience and Remote Sensing*, 46(1):259–271.

Kovacs, A. and Sziranyi, T. (2012a). Harris function based active contour external force for image segmentation. *Pattern Recognition Letters*, 33(9):1180–1187.

Kovacs, A. and Sziranyi, T. (2012b). Orientation based building outline extraction in aerial images. In *ISPRS Annals of Photogrammetry, Remote Sensing and the Spatial Information Sciences, Vol. I-7*, volume I-7, pages 141–146, Melbourne, Australia.

Kovacs, A. and Sziranyi, T. (2013). Improved Harris feature point set for orientation sensitive urban area detection in aerial images. *IEEE Geoscience and Remote Sensing Letters*, 10(4):796–800.

Kumar, S. and Hebert, M. (2003). Man-made structure detection in natural images using a causal multiscale random field. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 119–126.

Lin, C. and Nevatia, R. (1998). Building detection and description from a single intensity image. *Computer Vision and Image Understanding*, 72(2):101–121.

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110.

Manno-Kovacs, A. and Sziranyi, T. (2013). Multidirectional Building Detection in Aerial Images Without Shape Templates. In *ISPRS Workshop on High-Resolution Earth Imaging for Geospatial Information*, pages 227–232.

Martinez-Fonte, L., Gautama, S., Philips, W., and Goeman, W. (2005). Evaluating corner detectors for the extraction of man-made structures in urban areas. In *IEEE International Geoscience and Remote Sensing Symposium*, pages 237–240.

Mester, R. (2000). Orientation estimation: Conventional techniques and a new non-differential approach. In *Proc. 10th European Signal Processing Conference*, pages 921–924.

Ok, A. O., Senaras, C., and Yuksel, B. (2013). Automated detection of arbitrarily shaped buildings in complex environments from monocular VHR optical satellite imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 51(3):1701–1717.

Ortner, M., Descombes, X., and Zerubia, J. (2008). A marked point process of rectangles and segments for automatic analysis of digital elevation models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(1):105–119.

Otsu, N. (1979). A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man and Cybernetics*, 9(1):62–66.

Peng, J., Zhang, D., and Liu, Y. (2005). An improved snake model for building detection from urban aerial images. *Pattern Recognition Letters*, 26(5):587–595.

Perona, P. (1998). Orientation diffusions. *IEEE Transactions on Image Processing*, 7(3):457–467.

Rosten, E., Porter, R., and Drummond, T. (2010). FASTER and better: A machine learning approach to corner detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(1):105–119.

Sirmacek, B. and Unsalan, C. (2008). Building detection from aerial images using invariant color features and shadow information. In *23rd International Symposium on Computer and Information Sciences*, pages 1–5.

Sirmacek, B. and Unsalan, C. (2009). Urban-area and building detection using SIFT keypoints and graph theory. *IEEE Transactions on Geoscience and Remote Sensing*, 47(4):1156–1167.

Sirmacek, B. and Unsalan, C. (2011). A probabilistic framework to detect buildings in aerial and satellite images. *IEEE Transactions on Geoscience and Remote Sensing*, 49(1):211–221.

Smith, S. M. and Brady, J. M. (1997). SUSAN - A new approach to low level image processing. *International Journal of Computer Vision*, 23(1):45–78.

Song, Z., Pan, C., and Yang, Q. (2006). A region-based approach to building detection in densely build-up high resolution satellite image. In *IEEE International Conference on Image Processing*, pages 3225–3228.

Tsai, V. (2006). A comparative study on shadow compensation of color aerial images in invariant color models. *IEEE Transactions on Geoscience and Remote Sensing*, 44(6):1661–1671.

Unsalan, C. and Boyer, K. L. (2004). Classifying land development in high-resolution satellite imagery using hybrid structural-multispectral features. *IEEE Transactions on Geoscience and Remote Sensing*, 42(12):2840–2850.

Unsalan, C. and Boyer, K. L. (2005). A system to detect houses and residential street networks in multispectral satellite images. *Computer Vision and Image Understanding*, 98(3):423–461.

Yi, S., Labate, D., Easley, G. R., and Krim, H. (2009). A shearlet approach to edge analysis and detection. *IEEE Transactions on Image Processing*, 18(5):929–941.

Zingman, I., Saupe, D., and Lambers, K. (2014). A morphological approach for distinguishing texture and individual features in images. *Pattern Recognition Letters*, 47:129–138.