Analysis of large-scale UAV images using a multi-scale hierarchical representation

Huai Yu¹,², Jinwang Wang³, Yu Bai³, Wen Yang¹,²,⁵ and Gui-Song Xia¹,²,⁵

¹School of Electronic Information, Wuhan University, Wuhan, China; ²Collaborative Innovation Center of Geospatial Technology, Wuhan University, Wuhan, China; ³State Key Laboratory for Information Engineering in Surveying, Mapping and Remote Sensing (LIESMARS), Wuhan University, Wuhan, China

ABSTRACT
Unmanned aerial vehicle (UAV)-based imaging systems have many superiorities compared with other platforms, such as high flexibility and low cost in collecting images, providing wide application prospects. However, the acquisition of the UAV-based image commonly results in very high resolution and very large-scale images, which poses great challenges for subsequent applications. Therefore, an efficient representation of large-scale UAV images is necessary for the extraction of the required information in a reasonable time. In this work, we proposed a multi-scale hierarchical representation, i.e. binary partition tree, for analyzing large-scale UAV images. More precisely, we first obtained an initial partition of images by an oversegmentation algorithm, i.e. the simple linear iterative clustering. Next, we merged the similar superpixels to build an object-based hierarchical structure by fully considering the spectral and spatial information of the superpixels and their topological relationships. Moreover, objects of interest and optimal segmentation were obtained using object-based analysis methods with the hierarchical structure. Experimental results on processing the post-seismic UAV images of the 2013 Ya'an earthquake and the mosaic of images in the South-west of Munich demonstrate the effectiveness and efficiency of our proposed method.

1. Introduction
1.1. Motivation and objective
Nowadays, unmanned aerial vehicle (UAV)-based imaging systems have been used in many remote sensing applications, such as DEM generation (Yang and Chen 2015) and object detection (Moranduzzo and Melgani 2014). UAVs have several advantages over traditional remote sensing platforms, such as higher flexibility and lower cost in collecting images, higher speed and better security. More importantly, the images are acquired with very high resolution (VHR), providing sufficient details for identification and extraction of objects. In addition, the UAV pre-processing procedures, i.e. image stitching (Li, Hui et al. 2015; Xu et al. 2016; Yu and Yang 2016) and 3D reconstruction (Turner, Lucieer, and Christopher 2012; Schönberger, Fraundorfer, and Frahm 2014), output extremely large-scale Digital Orthophoto Map (panorama or DOM) with VHR. The panorama or DOM can provide a global and detailed inspection of the investigative area. Thus the applications, such as segmentation and object of interest extraction, are very important tasks for processing the large-scale VHR panorama or orthoimage of UAV image. However, the interpretation of very large-scale VHR images remains a great challenge due to the big data volume and semantic complexity. In this paper, we mainly focus on the extraction of objects of interest and optimal segmentation of large-scale UAV images, which include several fundamental and essential problems:

1. The coarse-to-fine representation of large-scale VHR UAV images. A panorama or orthoimage of UAV images is a unification of multi-scale objects, with large-scale objects at coarse levels and small objects at fine levels. Thus, a multi-scale representation is essential to simultaneously detect large-scale objects and small-scale objects in VHR images. To construct a precise multi-scale representation, both color and texture information can make contributions. Thus, the extraction of image features and the design of hierarchical structure are still the key problems in VHR image interpretation.

2. The semantic information mining in VHR images. Based on the hierarchical image representation, several images of the same scene at different resolutions (such as low, medium, high
and VHR) are available. However, the extraction of regions of interest and optimal segmentation still face some challenges related to image size, extraction accuracy and computational complexity. In addition, how to describe these semantic regions in the hierarchical structure is also a problem for the information mining task.

1.2. Related work

The afore-mentioned key issues about UAV image processing can be generally summed up as multi-scale representation and information mining. In the past two decades, there have been many valuable works in the literatures. For the large-scale VHR image representation, multi-scale hierarchical structure representation is a general practice. As a consequence, three main methods are utilized, i.e. image pyramid, wavelet transform and hierarchical image partitions. Binaghi et al. analyzed high-resolution scenes through a set of concentric windows and a Gaussian pyramidal resampling approach (Binaghi, Gallo, and Pepe 2003). Yang and Newsam (2011) proposed a spatial pyramid co-occurrence to characterize the photometric and geometric aspects of images. The pyramid structures capture both the absolute and the relative spatial arrangements of objects, but the fixed regular shape and size choice of analysis window lack semantic difference. Baraldi and Bruzzzone (2004) used an almost complete basis for the Gabor wavelet transform of images at selected spatial frequencies, which appeared to be superior to the dyadic multi-scale Gaussian pyramid image decomposition. Actually, the wavelet decomposition is a low-pass filter of VHR images that represents the multi-scale property by the coefficients in different bands. However, the wavelet decomposition is a decimation of the original image which lacks consideration of the relationship between objects. By fully considering the semantic gap between different objects, some studies use object-based analysis methods to produce hierarchical image partitions. The topographic representation (Monasse and Guichard 2000; Luo, Aujol, and Gousseau 2009; Xia, Delon, and Gousseau 2010) is typically built on the gray-level image, which rarely concerns color difference. In the Refs (Salembier and Garrido 2000; Bai et al. 2015), various types of images, e.g. natural images, hyperspectral images, PolSAR images, were represented by a hierarchical structure, namely, binary partition tree (BPT), which was constructed based on particular region models and merging criteria. In these methods, BPT is an effective hierarchical structure for processing and analyzing images. It is theoretically established, easy to implement, and provides effective results. Starting from initial pixels or partitions of the image support, BPT merges the pair of the most similar regions at each step, until a single region is obtained. This structure represents homogeneous regions in the UAV data at different detail levels. Among these three hierarchical representation methods, the BPT is a relatively better structure with both local intrinsic properties and topological relationships being well reserved.

Segmentation is an important task for processing and analyzing UAV data, where an image is usually partitioned into several distinct and meaningful regions that fulfill a certain homogeneity criteria. These segments can provide rich information of the regions (color, texture, and geometric structure of the targets and scenes). However, the BPT representation often contains more regions than required, and the segmentation based on BPT is usually obtained by setting the number of segments or a threshold. In the particular case of UAV data, optimal segmentation from the hierarchies is still a difficult task. Several approaches about this problem in other domains have been proposed (Akcay and Aksoy 2008; Kiran and Serra 2013). The general practice is using dynamic programming (DP) for a set of energy functions. The original idea of minimizing an energy function over a hierarchy was provided by Breiman et al. (1984). Salembier (2015) proposed to use DP for pruning the BPT. These methods get the optimal partition by cutting a set of edges in the BPT structure. As a greedy algorithm, DP traverses the tree structure in a bottom-up way to find the globally optimal segmentation. The direct minimizing methods in DP treat the semantic gap between a node and its two descendants as a constant value. However, the semantic gap is rising from fine level to coarse level. To overcome the limitation of undersegmentation at coarse levels and oversegmentation at fine levels, the uniform entropy slice (UES) (Xu, Whitt, and Corso 2013) was proposed to flatten the hierarchy into a single segmentation and to seek a selection of objects that balances the objects’ energy function and the relative level.

For the semantic object of interest extraction from the hierarchical structure of VHR images, the general idea is using feature-based classification and deep learning methods. The feature-based classification methods (Felzenszwalb et al. 2010; Yu, Yang et al. 2016) use a list of standard features, e.g. color, texture and structure information, attribute pooling models, such as part-based models and bag of words to describe images. Then the features are put into some classifiers, such as support vector machine, decision tree and random forest, to get the final object labels of interest. To overcome the limitation of handcrafted or shallow learning-based features, deep learning algorithms, especially convolutional neural networks, are used to extract features for object detection (Cheng et al. 2016). However, regardless of traditional machine learning or deep learning methods, there should be a lot of labeled samples to train detection models. Besides, special models should be established to keep the scale invariant. For the object of
interest extraction in large-scale UAV images, various land cover categories and complex semantic information make the model training methods inefficient and inapplicable. The general practice is to select one or several certain samples and then search the similar objects all over the multi-scale representation of images, such as Ecognition.

1.3. Our contributions

Inspired by the excellent work in the afore-mentioned aspects and our preliminary work in UAV image analysis (Yu, Yan et al. 2016), an object-based multi-scale hierarchical structure, the BPT, is utilized in this paper to represent large-scale UAV images to address the two important topics, i.e. optimal segmentation and extraction of the object of interest. The BPT structure derives from the definition of an initial partition obtained by an oversegmentation algorithm, i.e. simple linear iterative clustering (SLIC). During the BPT construction, we fully consider the spectral and spatial information of the superpixels and their topological relationships. Besides, an optimal segmentation is achieved by dynamic programming, which is different from our former work using the uniform homogeneity slice. Moreover, objects of interest are extracted using structural nearest neighbor search method. Figure 1 shows a brief schematic diagram for analysis of large-scale UAV images.

The current paper differs from our conference paper (Yu, Yan et al. 2016) in the following three aspects: (1) the optimal segmentation strategy is modified using dynamic programming to find the globally optimal solution; (2) the objects of interest extraction using structural nearest neighbor search based on BPT is proposed; and (3) the extensive new experiments using the modified optimal segmentation and objects of interest extraction strategies are conducted and re-evaluated.

The remainder of this paper is organized as follows. Section 2 presents the object-based BPT construction of large-scale UAV images. Section 3 introduces the optimal segmentation algorithm and the objects of interest extraction method in detail. Some representative experimental results are exhibited in Section 4. Finally, we draw the conclusion of this work in Section 5.

2. Hierarchical image representation

The big data volume and very high-resolution properties make the interpretation of UAV panorama a great challenge. The object-based image analysis (OBIA) methods not only preserve the useful information in UAV images but also decrease the data volume to analyze. In addition, a hierarchical representation based on OBIA is an essential step for VHR image interpretation. This section mainly introduces the BPT representation of UAV panorama image.

2.1. Superpixels partition

Superpixel segmentation algorithms can be roughly divided into graph-based and gradient-based methods. Considering the computation speed and partition performance, we utilize the SLIC algorithm (Achanta et al. 2012) to obtain the initial superpixels. This method can produce consistent superpixels with similar sizes and shapes, and keep image boundary information at the same time. Furthermore, the segmentation results of large-scale UAV images are convenient for subsequent processing.

In this step, the large-scale VHR image with complex boundaries is segmented into many superpixels. Each superpixel is relatively homogeneous, thus it is unnecessary to consider the information details of the internal superpixels. In addition, the superpixel-based description can speed up the subsequent processing procedure and preserve the useful information. According to the principle of SLIC, the region size and regularity of superpixels can be set experientially.

2.2. Region model and similarity criterion

After image partition, the description of superpixels is an important task, which directly relates to the measurement of similarity between superpixels. Because the major difference of superpixels lies in color, we leave out size and shape information. The region model is characterized by color names (van de Weijer et al. 2009), which are linguistic color labels based on the assignment of colors in the real world. The color labels including 11

Figure 1. A brief schematic diagram for analysis of large-scale UAV images. Source: Li, Tang et al. (2015).
basic terms: black, blue, brown, gray, green, orange, pink,
purple, red, white, and yellow, are learned from Google images. The learning result is a partition of the color space into 11 regions. To use this color feature, the RGB area of superpixel is mapped to the color attribute space. The color names of superpixel are defined as follows:

\[
CN_k = \{p_k(cn_1), p_k(cn_2), \ldots, p_k(cn_{11})\}
\]

\[
p_k(cn_i) = \frac{1}{N} \sum_{x \in i} p(cn_i | f(x))
\]

where \(cn_i(i = 1, \ldots, 11)\) is the \(i\)th color name, \(N\) denotes the number of pixels in region \(R_i\), \(f(x) = (L, a, b)\) is the three channels' value of pixel \(x\) in Lab space, and \(p(cn_i | f(x))\) denotes the probability of a color name of pixel \(x\). Color names are more photometric invariant than other color features because the different shades of color are mapped to the same color names.

The two important concepts in OBIA are region model and similarity measurement. The region model is characterized by color names. In other words, the region model of superpixels is equal to the color names of itself. Furthermore, the union of two superpixels is modeled by the average of their color names. While the similarity of region \(R_i\) and \(R_j\) is measured by the weighted Euclidean distance (WED)

\[
d(R_1, R_2) = N_R_1 \left\| CN_{R_1} - CN_{R_1 \cup R_2} \right\| + N_R_2 \left\| CN_{R_2} - CN_{R_1 \cup R_2} \right\|
\]

where \(CN_{R_1}\) and \(CN_{R_2}\) denote the region models for \(R_1\) and \(R_2\), respectively; \(N_{R_1}\) and \(N_{R_2}\) denote the data volumes of \(R_1\) and \(R_2\), respectively.

### 2.3. BPT construction

Based on superpixels segmentation, the bottom level of hierarchical structure is composed by the original superpixels. BPT structure is constructed by merging superpixels. Every node and every level of the hierarchical structure BPT contain semantic information. The leaves represent the original superpixels and the root represents the entire image. We can reconstruct the tree on the condition that the parent, siblings, and sons of every node are available. To enhance the efficiency of BPT construction, a priority queue is established using all pairs of neighboring regions. If a new pair of regions enters the queue, the position or the order will be determined by the WED of the two regions. The top of the queue, which is the most similar pair of neighboring regions, is popped out for merging. Note that one region has many neighborhoods. Therefore, if a region has been used to generate a new region, all pairs of regions that contain this region will no longer be used. The pseudo-code of the fast BPT construction process is depicted in Algorithm 1.

### Algorithm 1. Fast CBPT construction

| Input: \(x_1, x_2, \ldots, x_k \in I / / x\) is superpixels |
| Output: the root node of BPT. |
|\(//I\) – large scale UAV image, \(N\) – the number of superpixels |
|//class node: value, left, right, parent, adjacents; |
|Initialize: rag ← \(\langle\text{node}_1, \text{node}_2\rangle\) |
|\(\text{RAGS} \leftarrow \text{priority queue} \langle\text{vector} < \text{rag} > \rangle, \text{cmprag} >\) |
|\(i = 1\) |
|Repeat: while \(i < N\) |
|for each \(x_1 < 1\) |
|\(\text{node}_1 ← \{\text{value} ← \{cn_1, \ldots, cn_{11}\}, \text{left}, \text{right}, \text{parent} ← \text{NULL}\}\) |
|Update \(\text{node}_1\) adjacents |
|for \(\text{node}_2 ← \text{node}_1\) adjacents |
|\(\text{rag} ← \langle\text{node}_1, \text{node}_2\rangle\) |
|\(\text{RAGS}.\text{push}(\text{rag})\) |
|end for |
|end for |
|end while |
|end function |

An example of BPT hierarchical representation of UAV image is shown in Figure 2.

### 3. Object-based analysis of the hierarchical representation

#### 3.1. Optimal segmentation

The hierarchical structure, i.e. the BPT, represents the UAV panorama in multiple spatial scales. Based on this structure, the hierarchical segmentation algorithms can analyze images at different scales simultaneously, whose output is a set of regions that capture different partitions of different scales. The optimal segmentation based on the hierarchical structure can overcome the limitation of oversegmentation at fine levels and undersegmentation at coarser levels, which means that the partition of multi-scale meaningful regions can be exactly achieved. Under this conception, we design an optimal segmentation method based on the hierarchical structure.

#### 3.1.1. Setting leaf-to-root path and objective function

Based on BPT construction, we denote the maximum hierarchical level by \(m\), the node set of each level \(T^i\) is denoted by \(V^i\), the entire tree is denoted by
where \( \mathbf{l} \) is a \( p \times 1 \) column vector, \( \mathbf{x} \) is a \( 1 \times n \) vector. If a node is selected as a partition, the value of the corresponding location in \( \mathbf{x} \) is set as 1, otherwise set as 0. Thus any \( \mathbf{x} \) satisfying Equation (4) provides a possible partition of BPT, which corresponds to a plausible segmentation of the UAV panorama.

According to the constraint above, there are still many feasible tree slices which are proper segmentations of the image. However, our purpose is to find the optimal partition that is most meaningful. In this paper, we propose a meaningful criterion named minimal heterogeneity, which is defined as follows:

\[
\phi_R = \sum_{i \in R} \left\| \mathbf{C}_i - \mathbf{C}_R \right\|
\]

where \( \mathbf{l} \) is a \( p \times 1 \) column vector, \( \mathbf{x} \) is a \( 1 \times n \) vector. If a node is selected as a partition, the value of the corresponding location in \( \mathbf{x} \) is set as 1, otherwise set as 0. Thus any \( \mathbf{x} \) satisfying Equation (4) provides a possible partition of BPT, which corresponds to a plausible segmentation of the UAV panorama.

According to the constraint above, there are still many feasible tree slices which are proper segmentations of the image. However, our purpose is to find the optimal partition that is most meaningful. In this paper, we propose a meaningful criterion named minimal heterogeneity, which is defined as follows:

\[
\phi_R = \sum_{i \in R} \left\| \mathbf{C}_i - \mathbf{C}_R \right\|
\]

where \( R \) is a node in the BPT which consists of several adjacent superpixels, \( \mathbf{C}_R \) is the region model, and \( \mathbf{C}_i \) is the model of \( i \)th superpixel. Using this criterion, we can obtain the entire heterogeneity of region \( R \).

Figure 2. The hierarchical representation of a UAV image in different levels. (a) A UAV image; (b) Several levels representation of the image.

Figure 3. All valid tree slices of a particular BPT.

\[ T = \{ T^1, T^2, \ldots, T^m \}, \] and the individual node \( s \) at level \( i \) is denoted by \( V^i_s \). The only node \( V^0_1 \) at \( T^1 \) is the root of \( T \). Because not all original superpixels are at the bottom level, we copy the nodes at the upper level to the bottom level. Thus, each level of BPT corresponds to a partition of the image. During the process of BPT construction, once two nodes are merged, their parent node is at a new level.

A segmentation is a non-overlapped division of an image with the union restoring the image entirely. Thus, a partition in the hierarchy is a set of nodes satisfying the principle that there is one and only one node selected at each leaf-to-root path in the hierarchy. For example, Figure 3 shows valid tree slices of a particular BPT. Each slice is highlighted by a black curve, and the nodes on the slice are darkened.

In the following, we formulate the above constraints. Let \( \mathbf{P} \) denotes a \( p \times n \) binary matrix, where \( p \) is the number of leaf nodes in \( T \) and \( n \) is the total number of nodes in \( T \). Each row of \( \mathbf{P} \) denotes a leaf-to-root path as shown in Figure 4(a). If a node is in the path, the value of the corresponding location at \( \mathbf{P} \) takes 1, otherwise, takes 0. The corresponding path matrix of Figure 3 is shown in Figure 4(b). There are three rows in \( \mathbf{P} \), which represent the three leaf-to-root paths in the BPT. For instance, node sequence \( (V_1, V_2, V_5) \) is the path \( P_2 \). Therefore, the value of the second row of \( \mathbf{P} \) is \( [1, 0, 0, 1] \).

Because a valid tree slice \( \mathbf{x} \) consists of one and only one node in each path, the valid tree slice satisfies the following formula:

\[
\mathbf{P} \mathbf{x} = \mathbf{l}_p
\]
Thus the segmentation objective is to seek a slice that balances the overall heterogeneity of selected nodes,

$$x^* = \arg \min_{i \in T} \sum_{i \in R} \phi_R$$

(6)

which subjects to Equation (4). Here, $T$ is the BPT node set of a UAV image.

### 3.1.2. Solving by dynamic programming

However, it is difficult to directly solve Equation (6), which requires enumerating all tree slices, and the answer is a degenerate minimum which selects all leaf nodes because their heterogeneities are all zero. We add a penalty term which tends to select nodes at coarse levels

$$\phi_R = \frac{\sum_{i \in R} \|CN_i - CN_k\|}{\|CN_k\|} + \lambda$$

(7)

where $\lambda$ is a constant value as data regularization term which encourages the optimization to find partitions with a reduced number of regions. Although nodes in the coarser levels have the relatively higher heterogeneity than nodes in the finer levels, the number of coarser level nodes is less than those at the finer levels. For the purpose of optimal segmentation, dynamic programming (Salembier and Foucher 2016), which is a greedy algorithm starting from the initial superpixels to extract the optimal partition by minimizing the criterion Equation (6). In the experiment, we set $\lambda = 2$ according to the author’s recommendation.

Noticing that this strategy is different from our former work which used the uniform homogeneity slice (Yu, Yan et al. 2016), the reason can be summed up in two aspects. The first is the optimal solution. Quadratic programming problem may obtain local optima while dynamic programming can get global optima. The second is the time cost. Quadratic programming problem is very slow for the optimal segmentation of large-scale UAV images, whereas the dynamic programming is faster in this situation.

### 3.2. Object of interest extraction

The optimal segmentation gives a global inspection of large-scale UAV images by gathering the similar pixels into a region. However, without any annotation, we cannot obtain more details about semantic regions. The extraction of objects of interest deals with this problem with detecting specific objects in UAV image. In the multi-scale hierarchical structure, every single node is a partition of UAV images, and its semantics can be measured by specific features. Owing to the same region model in different levels of BPT, the similarity measurement of multi-scale objects can be unified to the same criteria. Providing one or several labeled samples, the extraction of objects of interest based on BPT can be turned into the search of similar nodes on BPT.

![Figure 5. The same kind of node searching strategy.](image)

#### 3.2.1. Searching strategy

We found that the nodes at low levels are included by the nodes at high levels. If a node is classified as a category, all the descendants are classified as the same category. The relationship of nodes is simply described in Figure 5. Therefore, the search process can be started from the root node in a top-down searching manner, and ended when we find the objects or there is no child node. If a node is classified as the same category with the giving sample, the search path from root node to this node is ended. Otherwise, we continue to compare the similarity between the children and the giving sample. From the aspect of merging operation, if the two child nodes of one node belong to the same category, the node must belong to the same category. Meanwhile, if one node is different from the parent node, the sibling of this node is also different from itself.

#### 3.2.2. Spatial nearest neighbor definition

To measure the similarity between different nodes in BPT and the giving sample, we should define the region model and calculate the distance between nodes and sample. Using the region model defined in BPT construction, we calculate the Euclidean distance as follows:

$$D_i = \|CN_i - CN_o\| \leq Th \begin{cases} \text{true} \quad & \text{if } \|CN_i - CN_o\| \leq Th \\ \text{false} \quad & \text{otherwise} \end{cases}$$

(8)

where $CN_i$ is the color names of $i$th node in BPT, and $CN_o$ is the color names of giving object sample. The setting threshold $Th$ is used to judge whether the $i$th node is similar to the sample. If the distance is smaller than the giving threshold, the $i$th node can be preliminary accepted as the same category with the giving sample. In addition, if a node belongs to the category while the parent node is not, the distance of parent node will be much bigger than the distance of current node. This constraint can be formulated as follows:

$$D_{\text{p}} / D_i > \lambda$$

(9)

where $D_{\text{p}}$ is the Euclidean distance between the parent node of $i$th node and the providing sample. The ratio $D_{\text{p}} / D_i$ is used to measure the dissimilarity between two kinds of nodes. The parameter $\lambda$ is the threshold to judge whether we accept the preliminary node as the same category with the giving node.
The similarity between the \( i \)th node and the given sample can be measured by the above criteria. If the current node is recognized as the same category to the given sample, all the descendants of the \( i \)th node will also belong to this category, so the searching of this path (from root to the \( i \)th) can be stopped. However, if the current node does not belong to the same category to the giving sample, the searching process should go to the two child nodes. That is to say, the path searching ends when we find the nodes of the same category or there is no child node anymore.

4. Experimental results and discussion

In this section, we provide experimental results on two large-scale UAV images to demonstrate the effectiveness of the proposed analysis methods. The first image comes from the stitching of post-seismic UAV images captured by Canon 5D Mark II of the 2013 Ya’an earthquake (Li, Hui et al. 2015). The particular location is Yuxi village, Baosheng Town, Lushan County. Figure 6 shows part of the large-scale UAV image of Ya’an. As a stitching of 93 image pieces, the data volume is 10,643 \( \times \) 18,570 pixels. For convenience, we choose a part of the original panorama (3845 \( \times \) 1038 pixels) to conduct the segmentation experiment.

Another data are the mosaic of UAV images from the area located in the South-west of Munich, Germany. The image sequences were acquired in 2015 using the air-borne DLR 3 K sensor system (Koch et al. 2016) with the resolution of approximately 6 cm. The original data size of the panorama is 12,240 \( \times \) 13,540 pixels. We select a part of the images to conduct our experiments. The experimental platform consists of Intel Core i7 4790 CPU and 32GB RAM. We use Microsoft Visual Studio 2010, OpenCV 2.4.10 and MATLAB software. Figure 7 shows the experimental data-set of Munich (4000 \( \times \) 3000 pixels).

4.1. Optimal segmentation results

Observing the details of the data-sets, we can find the two experimental scenes are very complicated and consist of several different semantic areas. For example, buildings around roads show very similar spectral information, and shadows of buildings have a strong disturbance on the detection of the obscured regions.

Figure 8 shows the optimal segmentation result of the large-scale UAV image in the 2013 Ya’an earthquake. The red lines are the segmentation boundaries. All pixels surrounded by a closed red line are gathered as a semantic object. To better describe the segmentation results, the intensity average (RGB channels, respectively) of all pixels in each object is mapped to each pixel inside. The corresponding result is shown in Figure 9. From an overall perspective, the edge information of objects is well preserved and the objects with high contrast are well segmented, such as roads and buildings. Moreover, the similar objects with different scales are segmented as individual regions. These qualitative analyses demonstrate that dynamic programming is applicable in optimal segmentation on hierarchical image representation. However, in some areas with poor photometric contrast, such as farmland and grassland, the segmented edges may go through some objects. The reason is that dynamic programming will pick up the areas with relatively different intensity. Therefore, the trees are distinguished from the grassland owing to the greener color.

Figure 10 shows the optimal segmentation results of large-scale UAV image in Munich. Because the original image is very big, the red edges are not very clear.
The number of superpixels represents the number of leaf nodes in BPT. The time index contains the time of BPT construction and dynamic programming. It costs only 0.650s for the UAV image in Ya’an with 4508 superpixels and 1.915 s for the UAV image in Munich with 13,390 superpixels. Moreover, we can also find that the time is approximately proportional to the number of

time costs in BPT construction and dynamic programming. The number of superpixels represents the number of leaf nodes in BPT. The time index contains the time of BPT construction and dynamic programming. It costs only 0.650s for the UAV image in Ya’an with 4508 superpixels and 1.915 s for the UAV image in Munich with 13,390 superpixels. Moreover, we can also find that the time is approximately proportional to the number of

Table 1. The efficiency analysis for the two large-scale UAV images.

| UAV image | Number of superpixels | DP-Time (s) | UES (min) |
|-----------|------------------------|-------------|-----------|
| Ya’an     | 4508                   | 0.650       | 18.450    |
| Munich    | 13,390                 | 1.915       | 49.356    |

in Figure 10(a), but it is obvious that there is no red line which goes through any object. To further show the effectiveness of our methods, we present the optimal segmentation result with the average intensity in Figure 10(b). The distinguished objects, such as buildings and roads, are well segmented. From the global perspective, we can still recognize the scene with the average intensity. Moreover, the segmented objects with different scales demonstrate the robustness of the BPT representation.

The experimental results show the effectiveness of the multi-scale hierarchical representation of large-scale UAV image. In addition, we give a brief time analysis to show the efficiency of our method. Table 1 shows the
4.2. Extraction results of objects of interest

The optimal segmentation results show a global inspection of different kinds of objects with distinct boundaries. However, without any annotation of regions, the outcome objects have no semantic labels. Thus, we cannot distinguish different objects. In this section, we show the experimental results of specific object extraction. Based on the hierarchical structure, we first manually select a patch and assign it with a semantic label, then find the similar ones in the BPT. Since the multi-scale property has been considered in the BPT construction, the similar regions can be extracted regardless of what size the patch is. Using the above-mentioned Ya’an dataset, we select a road patch with the size of 30 × 30 pixels, the extraction results are shown in Figures 12 and 13.

From the road extraction results, we can observe the good performance of BPT representation. Note that there is only one labeled sample with the size of 30 × 30 pixels in this experiment. We also conduct the object of interest extraction experiment on the UAV image in superpixels. In addition, we compare this method with our former optimal segmentation method (Yu, Yan et al. 2016) using UES. The qualitative results of the optimal segmentation are almost the same. However, the quadratic programming in UES to find the optimal solution is very slow for large-scale UAV images. From Table 1, we can observe the superiority of dynamic programming.

In this study, dynamic programming uses a constant value $\lambda$ as regularization. This parameter actually sets a hard threshold for the semantic gap between a node and its two descendants. However, the semantic gap is rising from fine level to coarse level, which means the value actually selects levels that satisfy the defined semantic gap.

For UAV image interpretation, the optimal segmentation scheme of UAV panorama image can be used in subsequent tasks, such as image classification, object detection and recognition, change detection. Figure 11 shows the segmented area of grassy land and buildings. The building areas in Figure 11 are finely distinguished, which can be used for extracting residential areas.

**Figure 11.** The distinguished area of grassy land and buildings in the segmentation results. (a) The distinguished area one in Ya’an. (Source: Li, Tang et al. 2015) (b) The distinguished area two in Munich. (Source: Koch et al. 2016)

**Figure 12.** The road extraction result with red edges in Ya’an UAV image. Source: Li, Tang et al. (2015).
by addressing two practical applications, i.e. optimal segmentation and object of interest extraction. This framework provides substantial possibilities for large-scale VHR image classification, object detection, change detection, etc. The experimental results demonstrate that BPT representation based on superpixels is an effective hierarchical structure for processing and analyzing large-scale UAV images. Future work will focus on refining the results of object extraction by considering object’s prior knowledge and contextual information.

Funding
This work was supported in part by the National Key Basic Research and Development Program of China [grant number 2013CB733404] and the National Natural Science Foundation of China [grant number 61271401, [grant number 91338113].

Notes on contributors
Huai Yu received BSc degree in Electronic Information Engineering from Wuhan University, Wuhan, China, in 2015 and is currently pursuing PhD degree at the School of Electronic Information, Wuhan University. His research interests include high-resolution image classification and change detection, vision-based SLAM, and obstacle avoidance for unmanned aerial vehicle.

5. Conclusions
In this paper, a multi-scale hierarchical structure has been proposed to represent large-scale UAV images. The object-based BPT representation deals with the high resolution and large-scale property of UAV images by addressing two practical applications, i.e. optimal segmentation and object of interest extraction. This framework provides substantial possibilities for large-scale VHR image classification, object detection, change detection, etc. The experimental results demonstrate that BPT representation based on superpixels is an effective hierarchical structure for processing and analyzing large-scale UAV images. Future work will focus on refining the results of object extraction by considering object’s prior knowledge and contextual information.

Munich. From the road extraction results (Figure 14), we can see that the road structure is well distinguished. Although some buildings are wrongly detected as roads, these small mistakes have very little effects on the outline of roads. In addition, these false positives can be easily eliminated with some prior knowledge, such as line structure and continuity.

During the objects of interest extraction, there are two important parameters need to be carefully fine-tuned. The first one is the similarity threshold $T_h$ between an object and the giving sample. This threshold defines the semantic gap between two objects, which varies with the definition of region model. In our method, we just use color names as the region model, so it is with little variation, ranging from 0.005 to 0.010. The second parameter is the threshold of spatial distance gap, which further removes the outliers. As it defines the distance ratio of the parent of a node and itself, just as the nearest neighbor search, it varies from 1 to 2. In our experiment, we set $\lambda = 1.4$ as a constant value.

Figure 13. The road extraction result with intensity average in Ya’an UAV image. Source: Li, Tang et al. (2015).

Figure 14. Road extraction results of UAV images in Munich. (a) Extraction result with red edge. (b) Extraction result with intensity average. Source: Koch et al. (2016).
Jinwang Wang received BSc degree in Communication Engineering from Lanzhou University, Lanzhou, China, in 2016 and is currently pursuing MSc degree at the School of Electronic Information, Wuhan University. His research interests include superpixel segmentation for unmanned aerial vehicle aerial image, vision-based tracking algorithm.

Yu Bai received BSc degree in Communication Engineering from South central University for Nationalities, Wuhan, China, in 2013 and is currently pursuing PhD degree at the School of Electronic Information, Wuhan University. His research interests include forest vertical structure parameters estimation, and PolSAR/PolInSAR data processing and interpretation.

Wen Yang received BSc degree in Electronic Apparatus and Surveying Technology, the MSc degree in Computer Application Technology, and PhD degree in Communication and Information System, all from Wuhan University, Wuhan, China, in 1998, 2001, and 2004, respectively. From September 2008 to September 2009, he worked as a visiting scholar with the Apprentissage et Interfaces team of the Laboratoire Jean Kuntzmann in Grenoble, France. From November 2010 to October 2013, he worked as a postdoctoral researcher with the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University. He is currently a full professor with the School of Electronic Information, Wuhan University. His research interests include object detection and recognition, image retrieval and semantic segmentation, multi-sensor information fusion, and change detection.

Gui-Song Xia received BSc degree in Electronic Engineering and the MSc degree in Signal Processing from Wuhan University, Wuhan, China, in 2005 and 2007, respectively and PhD degree in Image Processing and Computer Vision from the CNRS LTCI, TELECOM ParisTech, Paris, France, in 2011. Since March 2011, he has been a postdoctoral researcher with the Centre de Recherche en Mathématiques de la Decision (CEREMADE), CNRS, Paris-Dauphine University, Paris, for one year and a half. He is currently a full professor with the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University. His research interests include mathematical image and texture modeling, content-based image retrieval, structures from motions, perceptual grouping, and remote sensing image understanding.

ORCID

Huai Yu  http://orcid.org/0000-0001-6043-3412
Wen Yang  http://orcid.org/0000-0002-3263-8768

References

Achanta, R., A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk. 2012. “SLIC Superpixels Compared to State-of-the-Art Superpixel Methods.” IEEE Transactions on Pattern Analysis and Machine Intelligence 34 (11): 2274–2282.

Akcay, H. G., and S. Aksoy. 2008. “Automatic Detection of Geospatial Objects Using Multiple Hierarchical Segmentations.” IEEE Transactions on Geoscience and Remote Sensing 46 (7): 2097–2111.
Salembier, P., and S. Foucher. 2016. "Optimum Graph Cuts for Pruning Binary Partition Trees of Polarimetric SAR Images." *IEEE Transactions on Geoscience and Remote Sensing* 54 (9): 5493–5502.

Salembier, P., and L. Garrido. 2000. "Binary Partition Tree as an Efficient Representation for Image Processing, Segmentation, and Information Retrieval." *IEEE Transactions on Image Processing* 9 (4): 561–576.

Schönberger, J. L., F. Fraundorfer, and J. M. Frahm. 2014. "Structure-from-Motion for MAV Image Sequence Analysis with Photogrammetric Applications." *ISPRS – International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* XL-3 (3): 305–312.

Turner, D., A. Lucieer, and W. Christopher. 2012. "An Automated Technique for Generating Georectified Mosaics from Ultra-High Resolution Unmanned Aerial Vehicle (UAV) Imagery, Based on Structure from Motion (SfM) Point Clouds." *Remote Sensing* 4 (12): 1392–1410.

van de Weijer, J. V. D., C. Schmid, J. Verbeek, and D. Larlus. 2009. "Learning Color Names for Real-world Applications." *IEEE Transactions on Image Processing* 18 (7): 1512–1523.

Xia, G. S., J. Delon, and Y. Gousseau. 2010. "Shape-based Invariant Texture Indexing." *International Journal of Computer Vision* 88 (3): 382–403.

Xu, C., S. Whitt, and J. Corso. 2013. "Flattening Supervoxel Hierarchies by the Uniform Entropy Slice." Paper presented at Proceedings of the IEEE International Conference on Computer Vision, Sydney, Australia, December 1–8: 2240–2247.

Xu, Y., J. Ou, H. He, X. Zhang, and J. Mills. 2016. "Mosaicking of Unmanned Aerial Vehicle Imagery in the Absence of Camera Poses." *Remote Sensing* 8 (3): 204.

Yang, B., and C. Chen. 2015. "Automatic Registration of UAV-Borne Sequent Images and LiDAR Data." *ISPRS Journal of Photogrammetry and Remote Sensing* 101: 262–274.

Yang, Y., and S. Newsam. 2011. "Spatial Pyramid Co-occurrence for Image Classification." Paper presented at Proceedings of the 13th International Conference on Computer Vision (ICCV), Barcelona, Spain, November 6–13: 1465–1472.

Yu, H., T. Yan, W. Yang, and H. Zheng. 2016. "An Integrative Object-based Image Analysis Workflow for UVA Images." *ISPRS – International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* XLII-B1: 1085–1091.

Yu, H., and W. Yang. 2016. "A Fast Feature Extraction and Matching Algorithm for Unmanned Aerial Vehicle Images." *Journal of Electronics and Information Technology* 38 (3): 509–516.