IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received May 17, 2019, accepted June 8, 2019, date of publication June 19, 2019, date of current version July 9, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2923765

Computer Vision System for Automatic Counting of Planting Microsites Using UAV Imagery

WASSIM BOUACHIR[1], **KOFFI EDDY IHOU**[1,2], **HOUSSEM-EDDINE GUEZIRI**[3], **NIZAR BOUGUILA**[2], **(Senior Member, IEEE), AND NICOLAS BÉLANGER**[1]

[1]Department of Science and Technology, TÉLUQ University, Montreal, QC G1K 9H6 , Canada
[2]Concordia Institute for Information Systems Engineering, Concordia University, Montreal, QC H3G 1M8, Canada
[3]Department of Biomedical Engineering, McGill University, Montreal, QC H3A 0G4, Canada

Corresponding author: Wassim Bouachir (wassim.bouachir@teluq.ca)

This work was supported by the Research Grant from the Natural Sciences and Engineering Research Council of Canada (NSERC).

ABSTRACT Mechanical site preparation by mounding is often used by the forest industry to provide optimal growth conditions for tree seedlings. Prior to planting, an essential step consists in estimating the number of mounds at each planting block, which serves as planting microsites. This task often requires long and costly field surveys, implying several forestry workers to perform manual counting procedure. This paper addresses the problem of automating the counting process using computer vision and UAV imagery. We present a supervised detection-based counting framework for estimating the number of planting microsites on a mechanically prepared block. The system is trained offline to learn feature representations from semi-automatically annotated images. Mound detection and counting are then performed on multispectral UAV images captured at an altitude of 100 m. Our detection framework proceeds by generating region proposals based on local binary patterns (LBP) features extracted from near-infrared (NIR) patches. A convolutional neural network (CNN) is then used for classifying candidate regions by considering multispectral image data. To train and evaluate the proposed method, we constructed a new dataset by capturing aerial images from different planting blocks. The results demonstrate the efficiency and validity of the proposed method under challenging experimental conditions. The methods and results presented in this paper form a promising cornerstone to develop advanced decision support systems for planning planting operations.

INDEX TERMS Precision forestry, computer vision, UAV imagery, artificial intelligence.

I. INTRODUCTION

For optimal survival and early yields of tree plantations, forest managers employ silvicultural methods to promote fast development of the root system, thus favoring water and nutrient uptake [1]. However, certain conditions such as high soil bulk density due to compaction by logging machinery, high water table, and plant competition for resources can severely impede root establishment of the planted trees [2]. Under such conditions, mechanical site preparation is generally useful because the induced disturbance can improve the quality of planting microsites by increasing soil temperature and water retention capacity, and decreasing soil density and plant competition [3]. The redistribution of some nutrient-rich soil layers at depths that coincide to the rooting depth of tree seedlings is also expected to benefit juvenile growth [4]. Mechanical site preparation by mounding (see Figure 1) is thus a recommended type of planting microsite in North America [5].

One problem when planting is the difficulty in accurately estimating the number of mounds created after a site was mechanically prepared and is ready to be planted. The number of mounds varies due to several factors such as site characteristics and preparation quality. It is usually estimated by manual count and thus requires several workers, is time-consuming, costly and subject to error. Uncertainties associated with manual counting also leads to complex and imprecise handling of seedlings in the field, which causes monetary losses and planting delays.

Using Unmanned Aerial Vehicles (UAV) imagery, our work aims to develop novel computer vision methods for fast and accurate estimation of the number of mounds. UAVs have recently led to significant changes in the forestry practices, replacing satellites and aircraft in several data

The associate editor coordinating the review of this manuscript and approving it for publication was el-Hadi M. Aggoune.

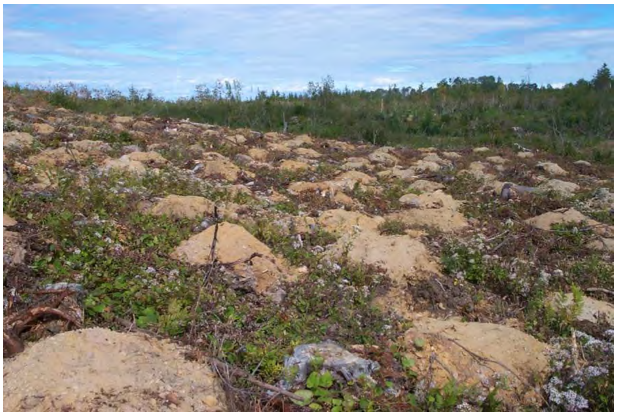

FIGURE 1. Examples of mechanically prepared mounds in the balsam fir-white birch bioclimatic domain in Quebec, Canada (credits: Nicolas Bélanger).

collection tasks, while substantially reducing the risk and time of manual field work. However, existing forestry methods are not taking full advantage of important advances in artificial intelligence and imaging technology.

Based on recent UAV imagery techniques and advanced machine vision methods, we propose technological methods of solving an important forestry problem. To do so, we combine for the first time computer vision and UAV imagery to automate the estimation of the number of planting microsites.

In our conception, mound detection and counting are performed by automating the analysis of high-resolution aerial images captured using a data acquisition UAV. We introduce a novel framework exploiting multi-spectral images for mound detection and counting. The proposed system is first trained in an offline manner to distinguish mounds from surrounding terrain on semi-automatically annotated images. During offline training, detection models are constructed based on a combination of Local Binary Patterns (LBP) and deep features. Online plantation microsite counting is then performed in a two-stage detection approach. A cascade detection algorithm is first used to generate object proposals based on LBP features. Candidate objects are then analyzed by a Convolutional Neural Network (CNN) for final classification as planting microsites or surrounding terrain.

To build prediction models and validate the proposed methods, a new image dataset was created by overflying mechanically prepared sites and capturing high-resolution images. Based on results, a discussion of important challenges and recommendations are provided for further development of the proposed approach.

We believe that the use of UAV imagery in combination with the proposed computer vision methods will contribute to improving fieldwork conditions and saving considerable time and money for forest managers. Our work represents a starting point for promising decision support systems to be used by forest managers when planning planting operations.

This paper is structured as followed. Section II introduces background concepts and related works on automatic object counting from images. Section III presents the methods for detection and counting of planting microsites. The experimental results are presented and discussed in IV. Finally, section V presents a conclusion and suggests directions for future work.

II. AUTOMATIC OBJECT COUNTING ON IMAGES: RELATED WORK

Automatic object counting was one of the most active research areas in computer vision during the last decade. With recent advances in aerial imaging technologies and machine learning, automatic object counting was used in a wide range of applications, including:

- crowd analysis for several purposes, such as public safety, disaster management, urban planning and behavior analysis [6]–[8],
- counting microorganisms (e.g. cells on microscopic images) [7], [9]–[12],
- vehicle counting for traffic control and congestion monitoring [13]–[16],
- wildlife census and environmental surveys for plants or animals [17], [18],

Automatic object counting methods can be categorized into traditional approaches and deep learning approaches, depending on their use of *hand-crafted* or *deep* features, respectively.

A. TRADITIONAL COUNTING APPROACHES

Traditional approaches are mostly based on hand-crafted image features for appearance modeling. We can distinguish three main strategies for using low-level visual features: 1) object detection techniques, 2) regression and density estimation, and 3) clustering and segmentation.

1) COUNTING BY DETECTION

Automatic counting is implemented using a supervised object detector for exhaustive search of object instances in the image. The object detector often consists of a machine learning model that is trained off-line on annotated data to recognize the object of interest. Once the detector is trained, a new image is processed by extracting and classifying visual features for candidate regions. Several representation strategies can be used for feature extraction. For example, a part-based strategy focuses on specific parts of the objects [19], while a holistic representation uses global features, such as the object shape and size [20] and color histograms [21]. One of the limitations of counting by detection methods is the difficulty in handling real-world situations, such as occlusion caused by overlapping objects and scene clutter.

2) COUNTING BY REGRESSION OR DENSITY ESTIMATION

Counting by regression also uses supervised learning from annotated data. Instead of detecting objects individually, counting by regression methods perform a direct mapping from image features to the number of objects in an image [6], [7]. A continuous function can be used for linear mapping between image features and a density map.

In this case, the integral over any region in the density map provides the count of objects. As regression-based methods do not rely on individual object detection, complex situations such as occlusion, scene clutter, and perspective distortions are implicitly handled.

3) COUNTING BY CLUSTERING AND SEGMENTATION

This approach is based on grouping (or clustering) image features in an unsupervised manner. Ahuja and Todorovic [22] used a tree structure to segment the entire image based on a hierarchy of local features. Object counting was then performed by identifying sub-trees having similar structural and geometrical properties. Rabaud and Belongie. [8] proposed to count moving objects by clustering image patches based on motion similarities. Their method is based on the assumption that a pair of points having similar motion are likely to belong to the same object. Thus, the resulting clusters (or groups) correspond to independently moving entities. Since counting by clustering allows to obtain a pixel-level segmentation, this approach is particularly appropriate for applications requiring a precise extraction (or segmentation) of object regions from images.

B. DEEP LEARNING-BASED METHODS

During the last decade, computer vision has Beena increasingly influenced by the significant progress achieved by deep learning in numerous tasks, especially in object detection and recognition. Unlike previously discussed methods that are dependent on hand-crafted features to represent objects, deep learning-based methods rely on learned features obtained by training convolutional neural networks (CNNs).

The AlexNet network [23] was proposed for counting people on crowded images [24]. The original AlexNet architecture was slightly modified to produce the object count for a given image patch. In a similar patch-based approach, Li *et al.* [25] proposed a CNN architecture specifically designed for detecting image patches corresponding to palm trees from satellite images. Once the CNN is trained to recognize the object of interest, satellite images are processed using a sliding window technique for individual detection of trees on small patches.

Instead of extracting local image patches, other methods propose to use the image as a whole to predict object count. For example, Shang *et al.* [26] proposed a CNN model, combining a GoogleNet architecture [27] for feature extraction and long-short time memory decoders (LSTM) to perform multiple local counts. This model uses an entire image as input and provides the global object count through the multiple local counts. Zhang *et al.* [28] also proposed to use the entire image as input for the CNN to produce a crowd density map, with its integral giving the overall crowd count. For a more complete literature review covering CNN-based counting methods, the reader is referred to [29].

Due to their rich hierarchy and design flexibility, CNN models provided significant progress and brought new ideas to the counting problem formulation. However, the network performance for a given problem highly depends on architecture options and hyperparameter optimization, which limits the reusability of the proposed models. Another issue with deep learning relates to the need for a large amount of labeled data for offline-learning. The lack of training data can be addressed by augmentation of the training dataset, such as in [30], where data augmentation is used to provide scale invariant representations.

III. THE PROPOSED APPROACH

A. MOTIVATIONS AND OVERVIEW

To collect aerial images of the mechanically prepared planting blocks, we used a data acquisition UAV system equipped with a multispectral sensor. We aim to automatically estimate the number of mounds in an image without requiring an accurate localization of the objects of interest, nor a precise segmentation. Situations of occlusion and overlapping objects are not part of our application constraints due to the vertical angle of camera view. Therefore, we propose a supervised counting by detection approach to estimate the number of planting microsites. This requires training machine learning algorithms using annotated images. Once the system is trained, mound detection on a new image is performed in two steps:

1) generating object proposals using a cascade detection algorithm,
2) binary classification of candidate objects as planting microsites or background regions using a CNN model.

By using LBP features, the cascade algorithm generates a large number of candidate patches that are considered as object proposals (or candidate objects). The object proposals are then processed by a trained CNN model in order to confirm or reject each proposal. The two models (i.e. cascade and CNN) are trained using the same annotated dataset of multispectral images, including positive and negative training examples. However, the visual information is exploited distinctly. The cascade algorithm uses LBP features from the NIR band, whereas the CNN model is trained using tridimensional patches, including blue, green, and NIR channels. By taking advantage of both LBP and deep features, our conception allows a robust feature representation against object appearance variation.

In the following sections, we present the three main methodological steps of the proposed framework: 1) dataset construction and system training, 2) generation of object proposals, and 3) final classification of candidate regions.

B. DATASET CONSTRUCTION AND SYSTEM TRAINING

The aerial multispectral images are captured by orienting the sensor vertically during stationary flights at an altitude of 100 m. An image of a mechanically prepared block captured at 100 m of altitude included from 1300 to 1600 mounds, depending on site characteristics. Once the images are collected, we performed semi-automatic annotation to extract training examples representing two classes

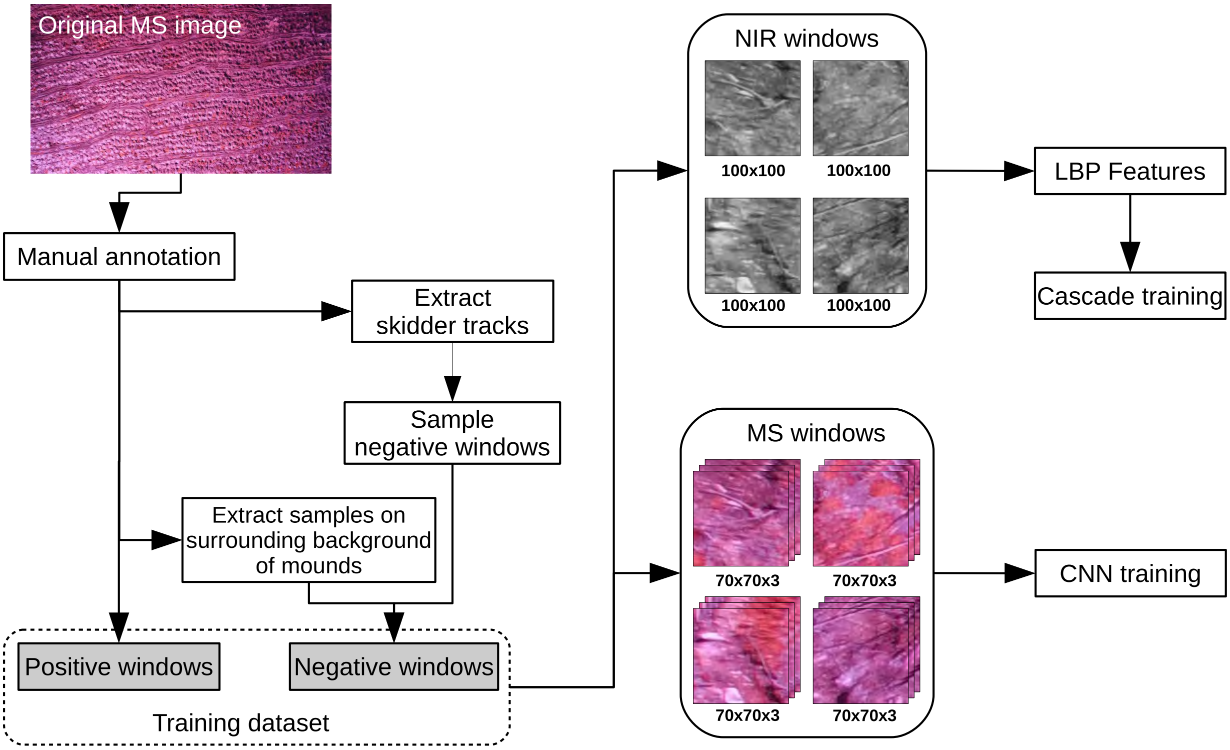

FIGURE 2. **The flowchart of dataset construction and system training.**

of objects: 1) positive examples corresponding to mounds, and 2) negative examples representing background terrain.

First, objects of interest were manually annotated by drawing bounding boxes surrounding each mound. Every annotated image patch was extracted to be used as a positive sample for training machine learning algorithms. Second, using the center locations of the manually annotated mounds, two types of negative samples were automatically generated: 1) samples on skidder tracks and 2) samples on the surrounding background of mounds. Figure 2 illustrates the workflow for training the algorithms.

The negative skidder track samples were extracted by creating a binary image B from the initial annotations, in which bounding boxes regions corresponding to positive examples were considered as foregrounds. Then, a set of patches was selected by sliding an average-size window over the binary image and retaining candidate patches having less than 25 % overlapping ratio with foreground pixels. This selection process can be formulated as

$$\omega \text{ is } \begin{cases} \text{selected} & \text{if } \displaystyle\sum_{\mathbf{x}\in\omega} \frac{B(\mathbf{x})}{|\omega|} < 0.25 \\ \text{ignored} & \text{otherwise,} \end{cases} \tag{1}$$

where $B(\mathbf{x})$ is the binary value for the pixel defined by the coordinate vector $\mathbf{x}$, and $|\omega|$ is the number of pixels in the sliding window ω. Once the selection process is completed, the binary image is updated to include the newly extracted bounding boxes. This method extracts negative patches in areas where manual annotations (i. e. positive examples) have a low density. These regions generally correspond to the terrain affected by the tracks of the excavator. (Figure 3a). However, this method results in an under-representation of negative samples due to the high overlapping ratio of the terrain surrounding the mounds. The second step of negative sampling aims thus to extract negative examples on the terrain surrounding the mounds (i.e. with a high density of positive examples).

This is achieved by computing a set C containing the centers of the annotated bounding boxes. Then, a distance map D is calculated on background pixels, such as

$$D(\mathbf{x}) = \min_{\mathbf{x}_c \in C}(\|\mathbf{x} - \mathbf{x}_c\|), \tag{2}$$

where $\|.\|$ denotes the Euclidean distance and $\mathbf{x}_c$ the coordinate vector of each annotated bounding box center. Therefore, each background pixel on the distance map represents the minimum distance to a bounding box center. Consequently, the local maxima on the distance map correspond to pixels that tend to be equidistant of neighboring mounds. These local maxima are finally selected as centers for the negative patches in between the mounds (Figure 3b).

C. OBJECT PROPOSALS

The first stage of the proposed detection framework performs a rapid extraction of candidate windows that are likely to contain mounds. This is achieved by using a LBP feature-based detector [31] on NIR images. During training, weak classifiers based on LBP features are boosted to yield a linear combination of stronger classifiers [32]. Once

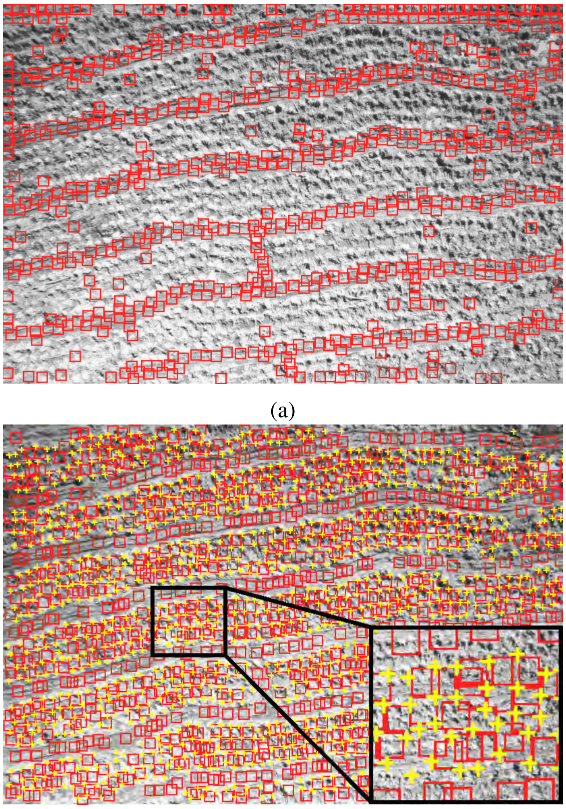

FIGURE 3. Example of negative sample extraction. Centers of positive samples are shown in yellow (+) and bounding boxes of negative samples in red: (a) skidder and excavator trails/tracks and (b) surrounding background of mounds. For visual clarity, the MS image is displayed in grayscale.

the model is trained, a new image is processed through a cascade classification procedure, by assessing the strongest (i.e. most discriminant) classifiers to rapidly discard negative (or background) regions.

Using the cascade classification generates a large number of detections. However, we observed that texture characteristics provided by LBP do not ensure sufficiently discriminative representation against background clutter and appearance variability of planting microsites. This resulted in the generation of a large number of false positive detections. Therefore, we consider image patches provided by the cascade algorithm as candidate regions, which has to be analyzed by a CNN for final classification.

D. CNN-BASED CLASSIFICATION

The object detection and counting is formulated as a binary classification problem of candidate regions provided by the cascade algorithm. We used the AlexNet CNN model [23] pre-trained on the ImageNet database, and transferred the initially learned feature representations to our recognition task [33]. More specifically, the network was fine-tuned by setting the classification output layer according to the two classes: *mound* and *background terrain*, and by retraining the last two fully connected layers on the annotated MS images. The weights on initial layers were frozen during transfer learning, as they correspond to general features that can be reused in different recognition tasks [34].

Once region proposals are generated by the cascade method, the fine-tuned CNN model is applied to iteratively classify each 70 × 70 image patch centered around the candidate object. For a given candidate region, the softmax layer of the CNN yields the probability distribution over the two classes *mound* and *background terrain*, which results in retaining or rejecting the object. Figure 4 illustrates the online detection and counting on a new aerial image.

IV. EXPERIMENTS

A. DATA ACQUISITION AND SETTINGS

The study area of this research is located as shown in figure 5, near the city of Sherbrooke (45°34’47.6“N 71°49’35.3”W), south of the province of Quebec, Canada. We collected aerial images by hovering private forest sites. Mounds with a height of approximately 50 cm and a diameter of 80 cm were constructed with an excavator equipped with a 45 cm-wide bucket. Each mound was planned to accept one seedling of a very fast growing hybrid poplar clone.

Flights were performed using the DJI Matrice 100 UAV shown in Figure 6. We used two sensors: Zenmuse X3 Multispectral (B, G, NIR) and Zenmuse X3 Visual (R, G, B). We captured aerial images during stationary flights at an altitude ranging from 50 to 125 m. The sensor was set at a vertical angle. The size of each image is 4000 × 2250. The proposed computer vision algorithms were implemented using Matlab on a PC (CPU i7-8700 @ 3.2GHZ, 6 cores) equipped with a GPU Nvidia Geforce GTX 1070.

Approximately 18000 mounds were manually annotated on the collected images. Negative examples representing the background were automatically generated, as described in section III-B. We were thus able to explore the efficiency of several acquisition settings, image types, as well as several visual features. The results reported thereafter were obtained using multispectral images captured at an altitude of 100 m from four different planting blocks.

B. RESULTS

To evaluate the detection performance of the proposed framework, we calculated the precision, recall, and F_1 measures as followed:

- precision (Eq. 3) is the percentage of correct detections among all the detected mounds,
- recall (Eq. 4) is the percentage of correctly detected mounds over the total number of mounds in the ground truth,
- F_1 score (Eq. 5) is the harmonic average of precision and recall.

The three metrics are defined as follows:

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

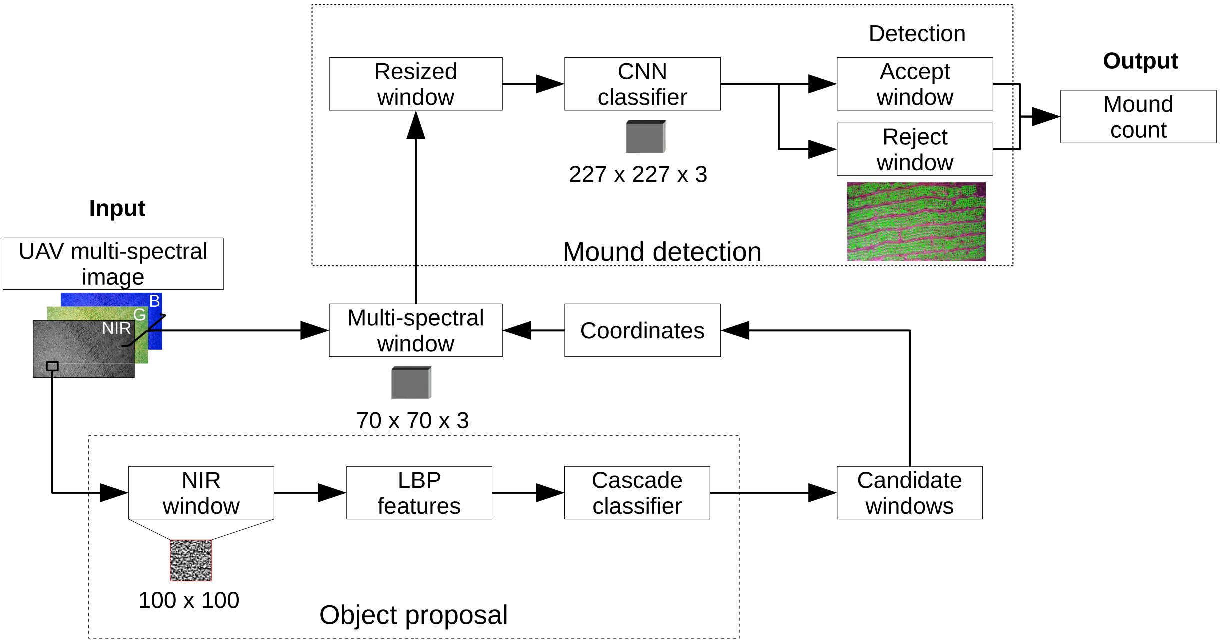

FIGURE 4. Proposed method for mound detection and counting on a new UAV multispectral image.

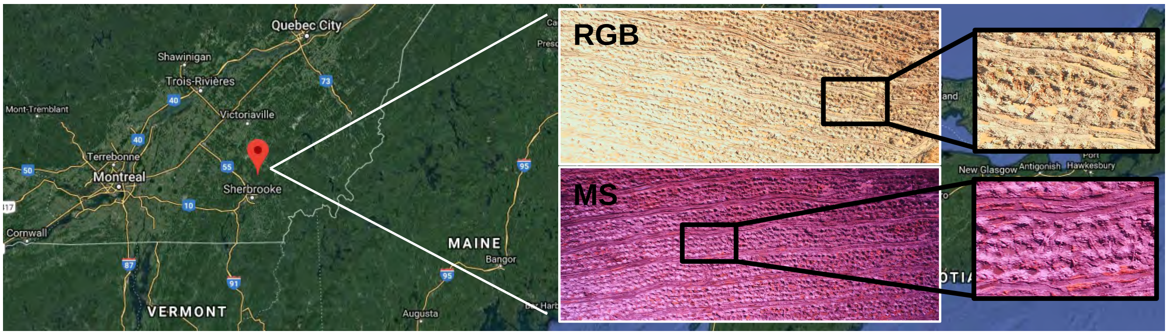

FIGURE 5. Data collection from the study area located in southern Quebec, Canada. We hovered mechanically prepared planting blocks using visible and multispectral sensors mounted on the UAV (see Figure 6).

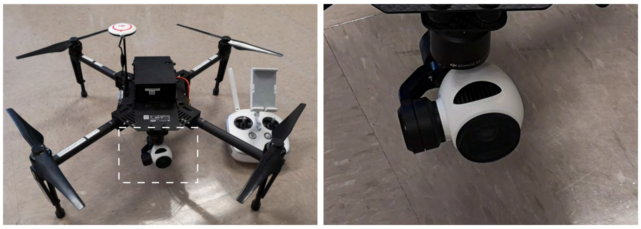

FIGURE 6. Unmanned aerial vehicle used for data acquisition: (a) DJI Matrice 100 and (b) DJI Zenmuse X3 multispectral sensor (B-G-NIR).

$$Recall = \frac{TP}{TP + FN} \tag{4}$$

$$F_1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{5}$$

where TP are the true positives (i.e. number of correctly detected mounds), FP are false positives (i.e. number of objects incorrectly detected as mounds), and FN denote false negatives (i.e. number of missed mounds). On a test image, a detection was considered as correct if the center of the detected mound and the center of a mound in the ground truth was less than or equal to 35 pixels. During method development, we experimentally. explored several image acquisition settings (e.g. flight altitude, sensor type) as well as various algorithm implementation options (e.g. features for region proposal generation, patch sizes, CNN models). The optimal performance of the proposed framework was obtained using MS images captured at an altitude of 100 m.

The system performance was rigorously evaluated using a cross-validation technique. We used four MS images, i.e. one image corresponding to each of the four planting blocks mechanically prepared by mounding. The center location of the four blocks is approximately located as shown in Figure 5 and images did not overlap. Four experiments were performed. Each experiment comprised two steps:

1) training the models on three images of three distinct planting blocks,
2) testing the system on the image of the remaining fourth block.

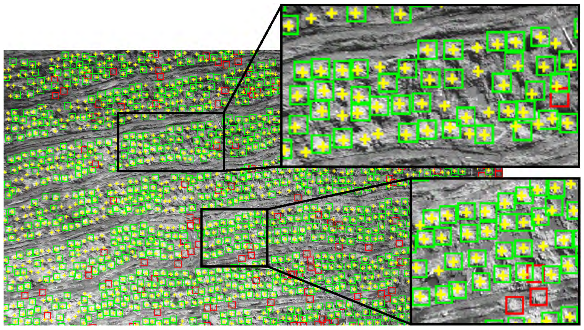

FIGURE 7. **Detection results on NIR image of a mechanically prepared planting block captured at a flight altitude of 100 m. Ground truth centers correspond to yellow (+), true positives to green squares and false positives to red squares. For visual clarity, the MS image is displayed in grayscale.**

TABLE 1. **Quantitative detection results for each testing block and average results. Percentages in bold font correspond to the best achieved result for each measure.**

	B_1	B_2	B_3	B_4	Average
Precision	**77%**	75%	74%	73%	75%
Recall	73%	70%	**90%**	89%	81%
F_1	75%	73%	**81%**	80%	77%

The detection performance was therefore assessed on unseen data collected from an independent planting block. This was used to validate the robustness of the proposed approach to the environmental variability between planting blocks.

Figure 7 shows qualitative detection results for the proposed framework. Detailed quantitative results for the four blocks are presented in Table 1. The highest F_1 score and recall values were obtained for B_3, with respectively 90% and 81%. Based on F_1 score as an overall accuracy indicator, the detection was more accurate for blocks 3 and 4, with F_1 scores exceeding 80%. This result can be explained by the fact that the first two blocks were more affected by visual perturbation factors, including uneven illumination conditions, occlusion by woody debris and coarse rock fragments, water accumulation, and especially mound erosion and collapse by rainwater. Figure 8 illustrates examples of these limiting factors. Despite challenging conditions affecting at various levels the four study blocks, results demonstrate the validity and the robustness of detection approach.

In regard to the counting method traditionally used, it should be noted that due to the vast extent and difficult access of planting blocks, manual counting through direct observation is generally made for a small sample region. The final estimation of the number of mounds for a given block is then obtained by extrapolation, assuming that mound density is constant. However, mound density may vary depending on site characteristics and excavation operations. The traditional method thus implies errors, manual counting, but also because of the constant density assumption. The resulting relative counting error is estimated to approximately 15% by forest managers. By adopting a detection-based approach that analyses entire image regions, our method improves performance compared to the actual method, with an average relative counting error of 13.8%. Our method also offers significant advantages in terms of deployment flexibility and data collection/processing speed.

C. DISCUSSION

This research is the first of its kind. It explores the use of UAV imagery and computer vision for automatic estimation of the number of planting microsites. We believe that this contribution will stimulate the interest of the scientific community. We also anticipate significant progress to be achieved during upcoming years on automatic counting of planting microsites, and more generally on the use of UAV imagery and computer vision for planning planting operations. In this section, we discuss aspects related to design options and important computer vision challenges faced in our project.

1) FLIGHT SETTINGS

The sensor orientation angle and flight altitude are the two main acquisition parameters. Capturing images by orienting the sensor vertically at a sufficiently high altitude allows for a high (and direct) visibility of planting microsites. Therefore, we alleviated the effect of occlusion by other background objects, while situations such as mutual occlusion between mounds almost never occurred, despite that some mounds can be significantly higher than neighboring ones.

Flying at high altitude also simplifies the image acquisition and preprocessing stages prior to applying automatic counting algorithms. In fact, for the sake of practicality, automatic counting at a given block should be preceded by mapping the area of interest to collect overlapping images. An orthomosaic is then created from the overlapping images to represent the entire area of interest. This step can be performed using photogrammetry software (e.g. PIX4D), generally based on Structure From Motion (SFM) algorithms [35].

The higher the flight altitude, the fewer images are needed to map a given area of interest. This not only minimizes UAV battery usage, but also orthomosaic production errors. However, since high-altitude flights reduce the level of details in the images, it is important to find a good trade-off between flight altitude and image quality. In our work, we noticed that the system performance decreases beyond an altitude of 100 m.

2) IMAGE TYPES AND VISUAL FEATURES

Two different sensors were mounted on the drone for data acquisition: a visual sensor providing RGB images, and a multispectral sensor specifically designed for vegetation sensing by capturing BG-NIR images. Figure 5 shows an example of each image type captured in this work (i.e. RGB, BG-NIR). After exploring both image types, we observed that MS images allow better visibility of planting microsites, while RGB images present a high sensitivity to illumination variation.

We also analyzed each channel separately and evaluated the relevance of several visual features, including LBP [36],

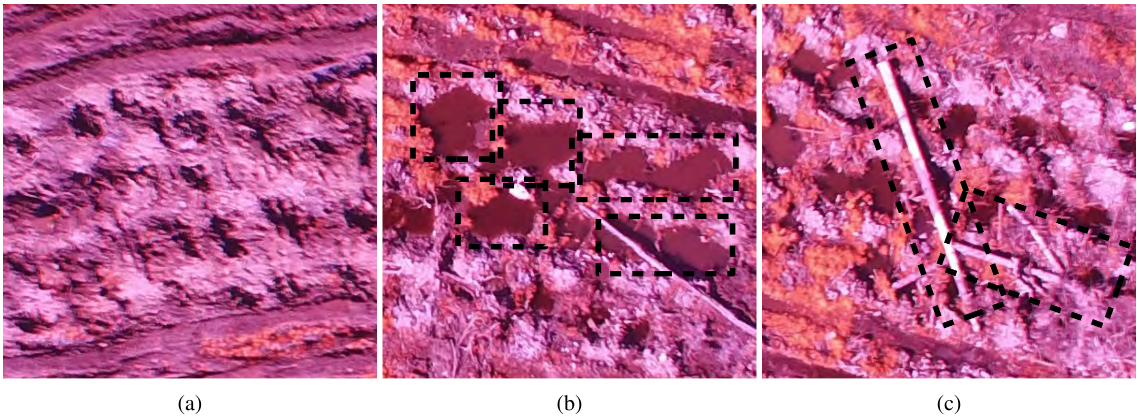

FIGURE 8. Examples of limiting factors on MS images captured at an altitude of 100 m. (a) image region with a good visibility of mounds. (b) water accumulation and mound erosion/deterioration caused by heavy rain events. (c) occlusion due to the presence of woody debris on the forest floor.

HOG [37], and Haar-like features [31]. We suggest that LBP features, when extracted from NIR bands, provide valuable texture information for mound recognition. LBP features were thus used in the first stage of our procedure to generate a large number of object proposals.

In the second stage, we perform CNN classification of candidate objects by exploiting the entire MS information. Our idea is motivated by the observation that region proposals include a large number of false positives, suggesting that LBP texture characteristics are not sufficiently discriminative against background clutter and appearance variability. Due to its rich feature representation, the CNN model allows to significantly improve the system accuracy by retaining or rejecting the candidate object.

3) CNN MODEL

We investigated various options regarding deep learning models by adapting pre-trained CNN models for mound and counting. In addition to the AlexNet architecture, we experimentally evaluated transfer learning with several other models including VGG [38], GoogLeNet [39], and ResNet [40]. In the context of binary classification of region proposals, the best performance was achieved by AlexNet on 70 × 70 BG-NIR patches representing candidate objects.

We also explored the development of a specific architecture for the studied detection problem, which is similar to [25] who designed a specific CNN model for detecting and counting palm trees from satellite images. However, finding the appropriate network architecture was extremely arduous, involving a long task of parameter optimization. It appears that our detection problem with mounds is more complex than for palm trees for two reasons:

1) the relatively large size of our objects of interest on UAV images (ranging approximately between 70 × 70 and 100 × 100 pixel) compared to palm trees on satellite images (17 × 17 pixels), which requires a more sophisticated CNN architecture and a greater number of parameters,
2) the significant variability in the appearance of mounds.

4) LIMITING FACTORS AND FUTURE WORK

Due to the variability in the appearance of the planting microsites, automatic detection and counting were very challenging in certain regions (or terrains). In fact, mounds may have different visual properties depending on several perturbation factors, including site characteristics, operation of excavator, presence of woody debris, and water accumulation in bowls. Moreover, in the context of intensive silviculture, planting blocks are often along hillslopes. Heavy rain events may favor mound erosion, thus deteriorating the planting microsites. Figure 8 shows examples of regions affected by these limiting factors. In order to mitigate the effects of hydro-climatic conditions, we recommend performing overflights and image capture as soon as mechanical site preparation is complete.

Such situations imply challenges not only for the automatic detection of mounds (because of the contamination of appearance models), but also for the manual annotation that becomes difficult and subjective. Indeed, it was noticed that when the visual recognition of a mound is difficult, the human annotator predicts its position and size taking into account the positions and sizes of neighboring mounds. In other words, the annotator takes into account the density of neighboring regions to predict the positions of the mounds that are obscured or deteriorated. Consequently, such annotations may result in contaminating the appearance model by including irrelevant features from the background or occluding objects.

For future work, we suggest addressing these real-world difficulties by integrating the concept of uncertainty to the computer vision framework. A more flexible annotation procedure would include two levels:

- annotating the objects of interest individually (or locally) when they are sufficiently visible,
- indicating image regions where objects are difficult or impossible to annotate.

Starting from these annotations, a promising direction for further development of detection/counting algorithms would be to draw inspiration from the human annotation reasoning described above. From a computer vision perspective, this reasoning could be translated into an automatic counting method that combines both detection and density approaches (see section II). A new image could thus be preprocessed by distinguishing two region categories: 1) regions with sufficiently visible objects, from 2) those that exhibit some complexity. Automatic counting on regions from the first category can then be conducted by detection while the counts for the remaining regions can be globally predicted using a density function that considers the counting results in neighboring regions from the first category. Note that this process is based on the same assumption of the human annotator, suggesting that the proximity between image regions implies similar densities of mounds.

V. CONCLUSION

By using UAV imagery, we proposed novel computer vision solutions to automate the detection and counting of planting microsites (i.e. mounds) that were mechanically prepared using an excavator. Our framework combines classical image features with deep learning methods in a counting-by-detection approach. The experimental results demonstrate the validity of our approach on a challenging dataset created using a multispectral sensor mounted on a UAV. This research provides novel ideas and valuable discussions for further development of this type of application.

Our future work will aim at improving system performance. First, we will explore the combination of detection and density-based approaches in order to handle difficult situations such as rugged terrain and erosion of planting microsites. Future efforts will also be devoted to the exploration of other UAV imaging technologies such as thermal imagery. Our idea is motivated by the temperature difference between a mound and the surrounding terrain, making mounds more distinguishable on thermal infrared images. The surface of the mounds is generally made of mineral soil and is bare of vegetation, whereas the surrounding terrain is covered by a forest floor (organic matter) and vegetation. Therefore, a thermal infrared sensor would allow capturing the contrast in temperature between the mounds (warmer mineral soil) and the surrounding terrain (colder forest floor and vegetation).

ACKNOWLEDGMENT

The authors would like to thank their collaborators from Domtar and Groupe DDM.

REFERENCES

[1] S. C. Grossnickle, “Importance of root growth in overcoming planting stress,” *New Forests*, vol. 30, nos. 2–3, pp. 273–294, 2005.

[2] J. M. Wolken, S. M. Landhäusser, V. J. Lieffers, and M. F. Dyck, “Differences in initial root development and soil conditions affect establishment of trembling aspen and balsam poplar seedlings,” *Botany*, vol. 88, no. 3, pp. 275–285, 2010.

[3] M. Löf, D. C. Dey, R. M. Navarro, and D. F. Jacobs, “Mechanical site preparation for forest restoration,” *New Forests*, vol. 43, nos. 5–6, pp. 825–848, 2012.

[4] S. Bilodeau-Gauthier, D. Paré, C. Messier, and N. Bélanger, “Root production of hybrid poplars and nitrogen mineralization improve following mounding of boreal podzols,” *Can. J. Forest Res.*, vol. 43, no. 12, pp. 1092–1103, 2013.

[5] T. N. Prévost, *La Préparation de Terrain. Dans Le Guide Sylvicole du Québec. Les Concepts et l'Application de la Sylviculture*, vol. 2, J.-P. Larouche, Ed. Les Publications du Québec, 2013.

[6] A. B. Chan, Z.-S. J. Liang, and N. Vasconcelos, “Privacy preserving crowd monitoring: Counting people without people models or tracking,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–7.

[7] V. Lempitsky and A. Zisserman, “Learning to count objects in images,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2010, pp. 1324–1332.

[8] V. Rabaud and S. Belongie, “Counting crowded moving objects,” in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 1, Jun. 2006, pp. 705–711.

[9] K. Chen, C. C. Loy, S. Gong, and T. Xiang, “Feature mining for localised crowd counting,” in *Proc. BMVC*, 2012, vol. 1. no. 2, p. 3.

[10] Y. Wang and Y. Zou, “Fast visual object counting via example-based density estimation,” in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2016, pp. 3653–3657.

[11] E. Walach and L. Wolf, “Learning to count with CNN boosting,” in *Proc. Eur. Conf. Comput. Vis.* Springer, 2016, pp. 660–676.

[12] Q. Geissmann, “OpenCFU: A new free and open-source software to count cell colonies and other circular objects,” *PLoS ONE*, vol. 8, no. 2, p. e54072, 2013.

[13] D. Onoro-Rubio and R. J. López-Sastre, “Towards perspective-free object counting with deep learning,” in *Proc. Eur. Conf. Comput. Vis.* Springer, 2016, pp. 615–629.

[14] M. Ahrnbom, K. Astrom, and M. Nilsson, “Fast classification of empty and occupied parking spaces using integral channel features,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2016, pp. 9–15.

[15] G. Amato, F. Carrara, F. Falchi, C. Gennaro, and C. Vairo, “Car parking occupancy detection using smart camera networks and deep learning,” in *Proc. IEEE Symp. Comput. Commun. (ISCC)*, Jun. 2016, pp. 1212–1217.

[16] P. R. L. de Almeida, L. S. Oliveira, A. S. Britto, Jr., E. J. Silva, Jr., and A. L. Koerich, “PKLot—A robust dataset for parking lot classification,” *Expert Syst. Appl.*, vol. 42, no. 11, pp. 4937–4949, 2015.

[17] B. Zhan, D. N. Monekosso, P. Remagnino, S. A. Velastin, and L.-Q. Xu, “Crowd analysis: A survey,” *Mach. Vis. Appl.*, vol. 19, nos. 5–6, pp. 345–357, 2008.

[18] G. French, M. Fisher, M. Mackiewicz, and C. Needle, “Convolutional neural networks for counting fish in fisheries surveillance video,” in *Proc. Mach. Vis. Animals Their Behav. (MVAB)*, 2015, pp. 1–7.

[19] M. Li, Z. Zhang, K. Huang, and T. Tan, “Estimating the number of people in crowded scenes by mid based foreground segmentation and head-shoulder detection,” in *Proc. 19th Int. Conf. Pattern Recognit.*, 2008, pp. 1–4.

[20] G. Grenzdörffer, “UAS-based automatic bird count of a common gull colony,” *Int. Arch. Photogram., Remote Sens. Spatial Inf. Sci.*, vol. 1, p. W2, 2013.

[21] N. Rey, M. Volpi, S. Joost, and D. Tuia, “Detecting animals in African Savanna with UAVs and the crowds,” *Remote Sens. Environ.*, vol. 200, pp. 341–351, Oct. 2017.

[22] N. Ahuja and S. Todorovic, “Extracting texels in 2.1 D natural textures,” in *Proc. IEEE 11th Int. Conf. Comput. Vis.*, Oct. 2007, pp. 1–8.

[23] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[24] C. Wang, H. Zhang, L. Yang, S. Liu, and X. Cao, “Deep people counting in extremely dense crowds,” in *Proc. 23rd ACM Int. Conf. Multimedia*, 2015, pp. 1299–1302.

[25] W. Li, H. Fu, L. Yu, and A. Cracknell, “Deep learning based oil palm tree detection and counting for high-resolution remote sensing images,” *Remote Sens.*, vol. 9, no. 1, p. 22, 2016.

[26] C. Shang, H. Ai, and B. Bai, “End-to-end crowd counting via joint learning local and global count,” in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, 2016, pp. 1215–1219.

[27] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, “Going deeper with convolutions,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 1–9.

[28] Y. Zhang, D. Zhou, S. Chen, S. Gao, and Y. Ma, “Single-image crowd counting via multi-column convolutional neural network,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 589–597.

[29] V. A. Sindagi and V. M. Patel, "A survey of recent advances in CNN-based single image crowd counting and density estimation," *Pattern Recognit. Lett.*, vol. 107, pp. 3–16, May 2018.

[30] L. Boominathan, S. S. Kruthiventi, and R. V. Babu, "CrowdNet: A deep convolutional network for dense crowd counting," in *Proc. 24th ACM Int. Conf. Multimedia*, 2016, pp. 640–644.

[31] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 1, Dec. 2001, p. 1.

[32] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *J. Comput. Syst. Sci.*, vol. 55, no. 1, pp. 119–139, Aug. 1997.

[33] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, "Learning and transferring mid-level image representations using convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1717–1724.

[34] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?" in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 3320–3328.

[35] J. P. Dandois and E. C. Ellis, "High spatial resolution three-dimensional mapping of vegetation spectral dynamics using computer vision," *Remote Sens. Environ.*, vol. 136, pp. 259–276, Sep. 2013.

[36] T. Ojala, M. Pietikäinen, and T. Mäenpää, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 971–987, Jul. 2002.

[37] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2005, vol. 1, no. 1, pp. 886–893.

[38] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," Sep. 2014, *arXiv:1409.1556*. [Online]. Available: https://arxiv.org/abs/1409.1556

[39] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 2818–2826.

[40] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2016, pp. 630–645.

WASSIM BOUACHIR received the Ph.D. degree in computer engineering from Polytechnique Montréal, QC, Canada, in 2015, and the M.Sc. degree in computer science from the Université de Moncton, New-Brunswick, Canada. He joined l'École de Technologie Supérieure, QC, Canada, as a Postdoctoral Researcher. He is currently a Professor of computer science with TÉLUQ University. His research interests include image analysis, video processing, and machine learning, especially to develop intelligent computer vision systems for several application areas, such as security (human behavior analysis and tracking), environment sciences (precision forestry and phenological event monitoring), and healthcare applications (respiratory monitoring and gait analysis).

KOFFI EDDY IHOU received the B.Sc. and M.Sc. degrees in electrical engineering from the Florida Institute of Technology (Florida Tech), Florida, USA, in 2005 and 2009, respectively. He is currently pursuing the Ph.D. degree with the Concordia Institute for Information Systems Engineering (CIISE), Concordia University, Montreal, QC, Canada. His research interests include computer vision, machine learning, and pattern recognition.

HOUSSEM-EDDINE GUEZIRI received the B.Sc. degree in computer science from the University of Science and Technology Houari Boumediene, Algiers, Algeria, in 2010, the M.Sc. degree in computer vision from Paris Descartes University, Paris, France, in 2011, and the Ph.D. degree in electrical engineering from the École de Technologie Supérieure, Montreal, QC, Canada, in 2017. He is currently a Postdoctoral Fellow with McGill University, Montreal. His research interests include image processing, surgical navigation, the design of softwares for image segmentation and registration, and human–computer interaction.

NIZAR BOUGUILA received the Engineering degree from the University of Tunis, Tunis, Tunisia, in 2000, and the M.Sc. and Ph.D. degrees in computer science from Sherbrooke University, Sherbrooke, QC, Canada, in 2002 and 2006, respectively. He is currently a Professor with the Concordia Institute for Information Systems Engineering (CIISE), Concordia University, Montreal, QC, Canada. He has authored or coauthored more than 200 publications in several prestigious journals and conferences. His research interests include machine learning, data mining, computer vision, and pattern recognition applied to different real-life problems (e.g., quality control, security, smart buildings, and finance). He received the Best Ph.D. Thesis Award in Engineering and Natural Sciences from Sherbrooke University, in 2007. He was a recipient of the prestigious Prix d'excellence de l'association des doyens des etudes superieures au Quebec (Best Ph.D. Thesis Award in Engineering and Natural Sciences in Quebec), and a runner-up for the prestigious NSERC doctoral prize. He is a Regular Reviewer for many international journals and serving as an Associate Editor for several journals, such as *Pattern Recognition* journal. He is a licensed Professional Engineer registered in Ontario.

NICOLAS BÉLANGER received the Ph.D. degree in natural resource science from McGill University, in 2000. He was a Postdoctoral Fellow for four years with Canadian Forest Service before accepting a position as an Assistant Professor of soil science with the University of Saskatchewan. He is currently a Full Professor of environmental science with TÉLUQ University, where he leads several projects on nutrient cycling, foliar nutrition, and forest and plantation yields. He is recognized for his use of isotopes (ex. d15N, d13C, 87Sr/86Sr, and d44Ca) to reconstruct nutrient cycling and comprehend the use of nutrients and water by trees in different environments. He is also the Scientific Co-Director of the Réseau Reboisement Ligniculture Québec, which facilitates the development of plantation forestry in Quebec. He is or has been a Principal Investigator for several NSERC, FRQNT, and CFI grants dealing with fast growing tree species plantations and fertilization and climate change impacts on soil processes (ex. respiration and litter decomposition) and forest growth.