Structure From Motion Technique for Scene Detection Using Autonomous Drone Navigation

Yo-Ping Huang, *Senior Member, IEEE*, Lucky Sithole, and Tsu-Tian Lee, *Fellow, IEEE*

***Abstract*—A method is presented for scene detection and estimation using high-resolution imagery acquired through autonomous drone navigation aided with landmark detection and recognition. The proposed system comprises a drone platform that facilitates efficient autonomous flight; it can capture images and provide real-time video streaming of the ground cover using a camera equipped with a 14-megapixel CMOS sensor and a fish-eye lens. In addition, landmark detection and recognition was performed by applying the histogram of oriented gradients and linear support vector machine methods on each frame of the video stream. The high spatial resolution of the acquired drone images makes the detection and interpretation of environments less complicated. First, through image processing, orthomosaic images and 3-D environment reconstruction (point clouds) of the scene are generated from a set of drone images by using an automatic photogrammetric technique called "structure from motion." Subsequently, an unsupervised classification method is used to detect and differentiate environmental classes (scene interpretation) in the target or investigated area by using the high-resolution images. Finally, the results of the proposed method are evaluated by comparing them against ground-truth points.**

***Index Terms*—Autonomous navigation, drone, scene detection, structure from motion (SfM).**

I. Introduction

RECENT technological advancements have resulted in an increase in the popularity of drones, also known as unmanned aerial vehicles (UAVs). The drone technology has become an effective alternative to satellite remote sensing for achieving major research breakthroughs. Because drones are compact and user- and eco-friendly and have the ability to capture images from a very low altitude, they can be used to obtain high-resolution imagery to supplement satellite imagery systems, whose performance may be limited by atmospheric phenomena (e.g., cloud cover) and lack of coverage over a targeted environment because of the orientation of the orbit around the Earth [1]. Furthermore, compared with traditional manned aerial systems, drones are highly effective for targeted remote-sensing operations in areas that are inaccessible and complex.

Efficient exploration and precise monitoring of complex urban environments are critical for applications such as forestry management and planning, flood modeling, pollution modeling, mapping and cartography, urban planning, coastline management, transportation planning, oil and gas exploration, volumetric analysis and exploration of quarries and minerals, archaeology, and cellular network planning. Furthermore, it can prevent problems related to poor disaster management. Buildings, roads, and landscapes deteriorate over time because of factors related to human interactions and environmental conditions. These conditions may pose a threat to both environment and health of human beings. Failing to maintain urban environments is tantamount to an act of disinvestment because it leads to huge loss of investments and may result in disasters if no preventive actions are taken.

The drone technology is an efficient and reliable method to continuously and precisely monitor urban environments through autonomous drone navigation and perform image processing for remote sensing with little human intervention. Drone platforms have become an increasingly popular and valuable source of data acquisition for interpretation, surveillance, environment mapping, and 3-D modeling applications. Because drones are usually less expensive than traditional manned aerial systems, they are being increasingly studied for short and close-range applications [2]–[4]. Rotary- and fixed-wing drones capable of performing photogrammetric data acquisition with amateur or digital SLR cameras can be used in manual, semiautomatic, and autonomous modes. By following a classical photogrammetric workflow, 3-D results such as digital surface or terrain models, contours, textured 3-D models, and raster and vector information can be generated for very large areas in a limited amount of time.

UAV technology was originally developed for military applications; nevertheless, it has gained popularity in recent years because of its great performance in civilian applications. The high-resolution images of ground objects in the investigated areas produced by UAVs make it easy to describe and differentiate these objects in a more detailed and explicit

Manuscript received June 30, 2017; accepted August 18, 2017. Date of publication September 14, 2017; date of current version November 19, 2019. This work was supported in part by the Ministry of Science and Technology, Taiwan, under Grant MOST105-2221-E-027-042 and Grant MOST106-2221-E-027-001, and in part by the Joint Project between the National Taipei University of Technology and Mackay Memorial Hospital under Grant NTUT-MMH-105-04 and Grant NTUT-MMH-106-03. This paper was recommended by Associate Editor M. Celenk. *(Corresponding author: Yo-Ping Huang.)*

Y.-P. Huang is with the Department of Electrical Engineering, National Taipei University of Technology, Taipei 10608, Taiwan, and also with the Department of Computer Science and Information Engineering, National Taipei University, New Taipei City 23741, Taiwan (e-mail: yphuang@ntut.edu.tw).

L. Sithole is with the Department of Electrical Engineering, National Taipei University of Technology, Taipei 10608, Taiwan (e-mail: luckysithole92@gmail.com).

T.-T. Lee is with the Department of Electrical Engineering, Tamkang University, New Taipei City 25137, Taiwan (e-mail: ttlee@ee.tku.edu.tw).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TSMC.2017.2745419

manner. Consequently, objects exhibiting the same features or belonging to the same class (e.g., trees and grass) can be easily identified, making detection of the entire class possible in a considerably small amount of time [5], [6].

In the recent years, there has been a huge improvements concerning the UAV technology but still suffers lack in one crucial aspect which is the capability of autonomy since early UAV are remotely piloted. Recent UAV platforms use global navigation satellite system (GNSS) information to perform limited autonomous flights that addresses many important issues such as take-off and landing without any human intervention. In order to enable autonomous navigation capabilities the UAV platform requires machine vision systems to complement GNSS to estimate the UAV state trajectory. The most common way to estimate the UAV state is to integrate inertial navigation system information with a GNSS, e.g., GPS. Ranft *et al.* [7] developed a very inexpensive framework for autonomous navigation for micro air vehicles which depends on a single camera and some additional on-board sensors such as the inertial measurement units (IMUs) to solve the challenges of flight planning and collision avoidance using sparse 3-D points to evaluate the quadcopter's position relative to the ground plane. The authors used artificial landmarks in areas with an ambiguous flight path, such as corridor crossings or junctions to provide topological localization, which enables the platform perform tasks such as way point following.

Many research work over the years focused on 3-D laser measurement, 3-D laser mapping, etc. Zhuang *et al.* [8] proposed a 3-D-laser-based place recognition system for a mobile robot to autonomously learn complex indoor scenes and avoid obstacle collisions caused by moving objects and people effectively. Zhang *et al.* [9] developed a framework that transform 3-D point clouds from mobile robot equipped with a custom-built 3-D laser scanner to 2-D to reduce dimensionality to obtain a less computational cost to novel multiclass and multiview 3-D object detection system.

In this paper, a drone model is proposed that can navigate over very large areas using GPS information, equipped with a real-time landmark detection and recognition system that improves the reliability of autonomous drone navigation. Landmarks have proved to be a more robust and reliable tool for scene detection and recognition because of their color and shape properties. In the detection phase, the color and shape properties of the landmarks are extracted as features. In the recognition phase, the model evaluates the regions found in the detection stage and identifies the landmarks.

Over the years, well-known features such as Haar-like features introduced by Viola and Jones [6] for face detection [10], histogram of oriented gradients (HOGs), speeded up robust features, and scale-invariant feature transform (SIFT) have been implemented using computer vision. The HOG algorithm, introduced by Dalal and Triggs [11] for pedestrian detection, outperforms existing algorithms such as Haar-like features.

To overcome the challenges associated with long-distance drone flights, we apply the HOG and linear support vector machine (SVM) algorithms to the detection and recognition of landmarks in the investigated region. We propose the use of landmarks to detect safe landing areas and automatic recharging platforms for drones. The limited capacity of batteries has restricted the flight distance of drones. The proposed method triggers an automatic landing protocol to enable the drone find the nearest charging platform in case of low battery charge. In addition, during emergency landings in cases of aborted missions or drone system malfunction, the drone is prone to collision and destruction because of a lack of safe landing places. The proposed method enables the drone to automatically land safely by locating the nearest safe landing place.

Fig. 1. Google Maps aerial view of the NTUT Athletic Field study area.

When the flight mission is complete, the captured images are processed for 3-D reconstruction of the environment to perform scene detection and interpretation through unsupervised learning algorithms. Although existing algorithms such as the K-means algorithm are widely applied for solving clustering problems, in this paper, the iterative self-organizing data analysis technique (ISODATA) algorithm is used for image clustering and classification because of its robustness, efficiency, and superiority over the K-means algorithm.

The remainder of this paper is organized as follows. Section II describes the test site and equipment used to implement the system. Section III explains the techniques used in the proposed system for scene detection. Section IV presents the experimental results and discussions. Finally, Section V presents the conclusion.

II. Study Area and Data Acquisition

We used a drone with an onboard camera equipped with sensors spanning the visible light range to capture a sequential set of images of our study area, namely the National Taipei University of Technology (NTUT) Athletic Field, Taipei, Taiwan. This study area, which measures approximately 6200 m^2, is a field wherein some grass areas have deteriorated due to poor management (Fig. 1). The images were acquired using the quadcopter Parrot BEBOP 2 drone with a camera equipped with a 14-megapixel CMOS sensor and a fish-eye lens; the flying altitude of the rotor is approximately 100 m. The images are characterized by three channels (RGB) and a spatial resolution of approximately 7.5 cm. The high-spatial-resolution drone images were used for 3-D reconstruction of the scene for further interpretation. Images and video streams were recorded by a ground control station (GCS) for image processing and real-time video processing.

Fig. 2. Parrot BEBOP 2 drone.

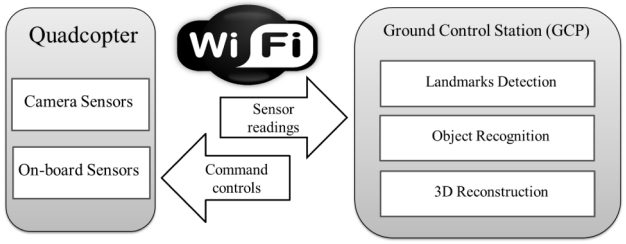


Fig. 3. Architecture of the proposed autonomous navigation.

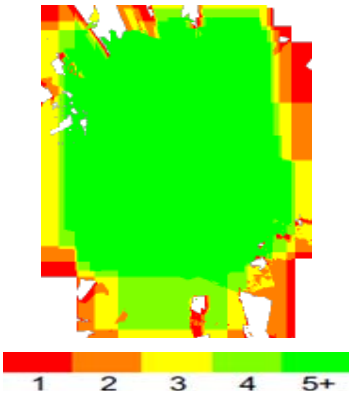


Fig. 4. Number of correctly matched keypoints for each pixel of the orthomosaic image.

A. Equipment

We used the Parrot BEBOP 2 drone for data acquisition and a remote personal computer (PC) for planning the drone flight missions and image processing.

1) Parrot BEBOP 2 Drone: The Parrot BEBOP 2 quadcopter drone has a front camera with a 14-megapixel CMOS sensor and a fish-eye lens for image acquisition (Fig. 2). The resolution of the acquired images and videos is 4096 × 3072 and 1920 × 1080 pixels (30 frames/s), respectively. To overcome the challenges of flight control and to avoid collision, the drone was equipped with the following onboard sensors. A pressure sensor that measures air pressure around the drone and analyzes the flight altitude above 4.8 m (16 ft) and an ultrasound sensor that analyzes the flight altitude up to 4.8 m. The GNSS chipset (GPS + Glonass + Galileo) geolocalizes the drone and maintains its flight path and measures the speed to stabilize the drone at very high altitudes. A triaxial gyroscope detects and maintains changes in the drone's direction. An accelerometer determines the position and orientation of the drone in flight and measures its linear speed. In addition, the drone was mounted with an additional vertical camera sensor on its bottom to determine the altitude and to capture images of the ground every 16 ms; moreover, the sensor compares consecutive images to determine the speed of the drone.

2) Ground Control Station: Our system uses a standard PC for autonomous flight planning and data acquisition. An advanced image processing technique is used for scene interpretation. Autonomous navigation of the drone requires the extraction of landmark features [7] to detect safe predefined landing locations and automatic recharging platforms. In addition, the onboard sensors enable the drone to detect potential collisions. The GCS communicates with the drone through a WiFi connection. Subsequently, the GCS uses Parrot SDK3, a widely used drone open-source middleware, and the Open-CV library executed in a Python environment to stream and process the real-time video; furthermore, the GCS uses sensor measurements from the drone and sends appropriate control commands for each of the four degrees of freedom in the drone (Fig. 3). The SIFT filter tools are used to enhance images and to extract low-level features, following which the images are subject to sharpening, brightening, noise removal, and texture feature extraction.

B. Data Acquisition

The images were captured using the Parrot BEBOP 2 quadcopter with a 14-megapixel CMOS sensor and a fish-eye lens during flight at an altitude of approximately 100 m. Images with a spatial resolution of 7.5 cm were used for 3-D reconstruction of the scene for further analysis. Approximately 50 images were captured over the archeological area of NTUT Athletic Field, with an overlap of more than 50% between the images (Fig. 4). Red, orange, and yellow areas indicate low overlap, with poor 3-D points generated for three or fewer matched keypoints. The green areas indicate good overlap, with more than five matched keypoints for every pixel. High-quality 3-D reconstruction of the scene can be realized when the number of correctly matched keypoints is sufficiently high for the target areas. Thus, the main objective of this experiment was to maintain a high overlap between the images in order to obtain sufficient matched keypoints to generate a high quality 3-D representation of the scene.

The sequential images captured from the drone were used for the 3-D reconstruction of objects using photogrammetry software such as the structure from motion (SfM) software, Pix4DMapper, and Photoscan Professional. The GPS information usually cannot be exported for other use while the drone is navigated by the inertial GPS unit. Therefore, an additional GPS unit sensor is attached to the drone for tagging the GPS

location on the images and 3-D objects for creating images with geographic coordinates.

The drone framework, aided by the landmark detection and recognition system, is capable of autonomously flying the drone over very large complex environments. The drone is also equipped with onboard sensors, collision detectors, to overcome flight challenges. The drone transmits real-time video to a GCS, which performs landmark detection and recognition on each video frame. The drone scans the environment to identify objects of interest and triggers appropriate actions when necessary.

C. Flight Planning

The flight and data acquisition is planned using a remote PC with dedicated geographic information system software. All information on the area to be investigated, the required ground sample distance (GSD), and the intrinsic parameters of the onboard drone camera is gathered. The required image scale and camera focal length are fixed for estimating the precise mission flying height. The camera perspective centers ("waypoints") are computed by fixing the longitudinal and transversal overlap of the strips. Depending on the objective of the particular flight, the parameters are varied using a detailed 3-D reconstruction structure that requires high overlaps between sequential images and a low flying height to obtain small GSDs. In addition, landmarks are defined in the investigated area to ensure safe landing and precise positioning of the drone. Sufficient drone battery capacity and safe landing platforms covering the entire investigated area are required to perform safe autonomous flights over very wide areas.

D. Autonomous Navigation

For safe and efficient navigation, drones must be able to localize themselves autonomously using their onboard sensors and interpretation of the unknown environment features. To this end, we propose a vision-based target detection and localization system.

The Parrot platform was used to obtain a real-time video stream to enable the drone to navigate autonomously over very large areas using GPS information and perform the real-time landmark detection and recognition system without relying on GPS. The proposed system identifies the regions of interest (ROIs) in each video frame during the flight and triggers autonomous responses based on the detected landmark and the current drone status. For example, on detecting low battery charge, the drone searches for the nearest charging platform in the video frame and executes a landing at the target location. Moreover, if the drone detects system malfunctioning during the flight or if the mission is aborted, the drone safely executes an emergency landing by searching and locating the nearest safe landing platforms, thus ensuring the smooth and safe return of the drone to the ground for easy recovery. For landmark detection, we applied the HOG feature extraction algorithm to extract features from our landmark datasets. For the landmark recognition system, we trained the model using a linear SVM algorithm (Fig. 5).

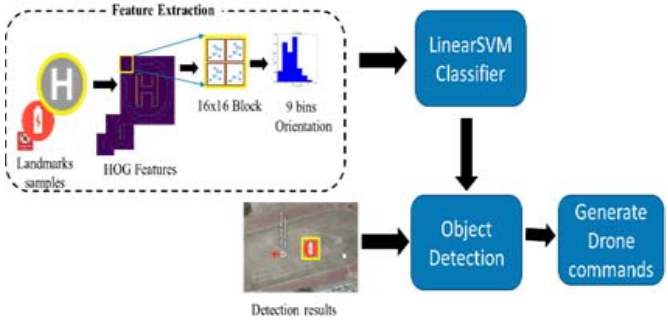


Fig. 5. Flow diagram for landmark detection and recognition using HOG and linear SVM algorithms for autonomous drone navigation.

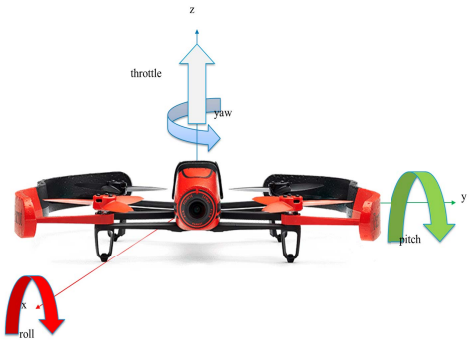


Fig. 6. Parrot BEBOP 2 quadcopter with four degrees of freedom during flight.

We calculated an HOG feature descriptor used for landmark detection based on a 128 × 128 patch of input images during the feature extraction phase. Then, we computed each image gradient in both x and y directions in the 8 × 8 cells. To make our descriptor robust and independent of illumination variations, we normalized the contrast of the image gradient using 16 × 16 blocks of 50% overlap and calculated the magnitude and direction of the image gradient. Finally, we trained the landmark recognition model by using the extracted features of the linear SVM model.

The GCS can send control commands to the drone, allowing it to autonomously take appropriate actions. These control commands adjust the acceleration and direction of the drone by controlling the speed of the four rotors to perform the yaw, roll, pitch, and throttle movements (Fig. 6).

III. Methodology

In this section, the proposed method for detecting and classifying the damaged grass areas of an athletic field is described. Fig. 7 delineates the proposed approach.

A. Data Preprocessing

The sequential drone images must be preprocessed to produce high-resolution orthomosaic images and the point clouds used as input data in our classification algorithms. The Pix4DMapper software [12] was used during the preprocessing. The high-resolution orthomosaic images with a resolution of 1137 × 1430 pixels and 3-D point clouds with an average

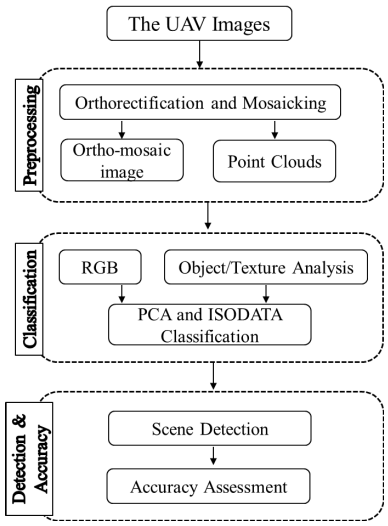


Fig. 7. Flow diagram of the proposed method for environment detection using drone imagery.

density of 6.32 points/m^3 were then used as the two forms of data for the investigated area.

1) Rectification: Raw digital images cannot be used as maps because of the presence of geometric distortions resulting from the image acquisition phase, which varies with the camera lens. Therefore, the original raw images must undergo geometric correction, and distortions such as variations in altitude and earth curvature must be corrected to achieve the same geometric integrity so that the images can be used as maps. First, radial distortion can be adequately corrected by, for example, applying a third-degree polynomial approach [13] to correct straight lines that appear curved in all the images. Second, every raw image must be rectified using a nonparametric rectification approach.

To eliminate the noise and correct distortion in all the images, we employed the Gaussian filter by using the 2-D function $f = f(I)$, where $I = (x,\ y) \in R^2$, to represent an image [14]. The index (x, y) of the center element of a 5×5 mask was set to (0, 0), and the entire mask was defined to span from $(-2, -2)$ to $(2, 2)$. The value of each element in the 5×5 Gaussian mask is defined as follows:

$$G_0(x, y) = e^{-(x^2+y^2)/2\sigma^2} \tag{1}$$

where $x, y = \{-2, \ldots, 2\}$ and the standard deviation $\sigma = [0.1, 5.0]$.

The drone images are challenging to process because of the lack of camera pose estimation in the direct and real-time measurements. To overcome this problem, we implemented a computational cost by applying a set of algorithms and mathematical operations. These algorithms perform automatic image matching, camera pose estimation for detecting outliers, and 3-D feature point triangulation. We used the RANSAC algorithm and a cost function to classify the outliers on the basis of pose estimation and the additional constraints from the sensor data of the IMU. For controlled flight, the drone is equipped with a magnetometer, an accelerometer, and a gyroscope, all of which collectively forms the IMU for sensor

Fig. 8. High-resolution orthomosaic image obtained by stitching the drone-acquired images.

measurements. Bundle adjustment was applied to detect the remaining outliers from the mismatches. We then optimized the pose estimation by minimizing the cost function by varying the constraints as follows [15]:

$$E_r = \sum_{k=1}^{n} \rho_x\left(\left\|x_k - q(P_i, X_j)\right\|_2^2\right) + \lambda \sum_{l=1}^{m} \rho_r\left(\left\|\hat{R}_l - R_l\right\|_F\right) \tag{2}$$

where $\|\cdot\|_F$ denotes the Frobenius norm and P_i is the camera pose. The 3-D points X_j, with x_k image measurements, the image projection function q, and the Cauchy functions ρ_x and ρ_r are used to optimize the results. The measured rotations from the IMU are denoted by $\hat{R}_l$, and the rotational parts of the camera pose P_i are given by R_l, with a regularization parameter λ being a weight term between the image measurements and IMU measurements.

2) Mosaicking: Mosaicking is the seamless joining or stitching of adjacent imagery [11] or several overlapping images to generate a large uniform image of the scene, as shown in Fig. 8. On completion of image rectification, the images can be merged together to form a high-resolution mosaic image; however, the image may still contain visible borders, which need to be rectified. We used image blending, an effective method that can result in better-quality mosaics, because drone-acquired images often possess radiometric variations of overlapping views. The 2-D sequential images from the drone were used for 3-D reconstruction of the investigated scene through the "SfM" technique. Although the drone is navigated using the inertial GPS unit, GPS information is not always available; consequently, accurate coordinates cannot be registered on the 3-D points. The SfM is a technique used for computing the camera parameters and 3-D coordinates of feature points from 2-D image sequences captured from different angles in computer vision. This process is used for 3-D reconstruction of the target scene and calculation of camera parameters [16] that consider the usefulness of matched points and bundle adjustment on completing feature-point extraction and matching.

The pinhole camera geometry models the projective camera with two subparametrizations, intrinsic and extrinsic parameters, describing the relationship that exists between the points on the image and the ground points. Assuming that the lens axis passes through the center of the image plane and that the pixel of the camera is foursquare, the camera parameters can be defined with seven parameters,

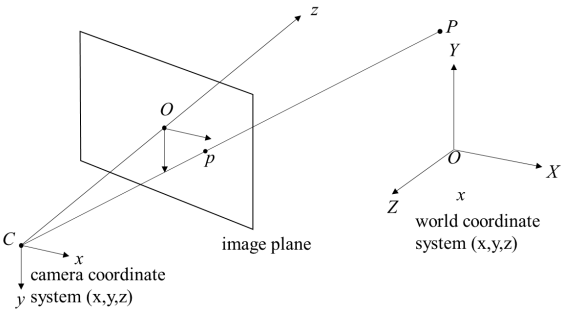


Fig. 9. Camera pinhole.

namely rotation parameters $\theta_i = (\theta_1, \theta_2, \theta_3)$, translation parameters $\boldsymbol{t}_i = (t_1, t_2, t_3)$, and the lens focus f_i, where i represents an image. The calibration matrix K_i is given by

$$\boldsymbol{K}_i = \begin{bmatrix} f_i & 0 & 0 \\ 0 & f_i & 0 \\ 0 & 0 & 1 \end{bmatrix}. \tag{3}$$

Fig. 9 shows point p in the image plane as a projection of the ground point P in the focal plane with coordinate system $\boldsymbol{X}_i = (X, Y, Z)$. The coordinates of image point p in the image space coordinate system are $\boldsymbol{x}_i = (x, y, z)$, with θ as the rotation parameter for each axis. Thus, the relationship is

$$\boldsymbol{x}_{ij} = \boldsymbol{K}_i\left(\boldsymbol{R}_i\boldsymbol{X}_j + \boldsymbol{t}_i\right) \tag{4}$$

where $\boldsymbol{x}_{ij}$ is the projection of point $\boldsymbol{X}_j$ in image i and $\boldsymbol{R}_i$ denotes the rotation matrix [16], which can be described as

$$\boldsymbol{R}_i = \begin{bmatrix} 0 & -\theta_3 & \theta_2 \\ \theta_3 & 0 & -\theta_1 \\ -\theta_2 & \theta_1 & 0 \end{bmatrix}. \tag{5}$$

Considering that the pixel is foursquare, α is the size of the pixel, and the principal point (u_0, v_0) of the image plane is assumed to be in the center of the image. The transformation from the pixel coordinates of one point to its relative image plane coordinates (x, y) can be represented as follows:

$$\begin{cases} x = \alpha(u - u_0) \\ y = \alpha(v - v_0). \end{cases} \tag{6}$$

In general, drone images possess large distortion because they are captured using a nonmetric digital camera; the most common type of distortion is radial distortion [3], which is proportional to the distance square of the image point to the principal point of the image plane. Assuming that the distortions in the directions of u and v are the same, the radial distortion model is as follows:

$$\begin{cases} u' - u_0 = \frac{xk(r)}{\alpha} \\ v' - v_0 = \frac{yk(r)}{\alpha} \end{cases} \tag{7}$$

$$k(r) = 1 + k_1 r^2 + k_2 r^4 \tag{8}$$

where (u', v') are the pixel coordinates with distortion difference, $k(r)$ is the ratio factor from nondistortion coordinates to distortion coordinates, $r = \sqrt{x^2 + y^2}$ is the distance of the image point to the principal point of the image plane, and k_1 and k_2 are the distortion parameters. Thus, the camera has nine parameters ($\theta_1, \theta_2, \theta_3, t_1, t_2, t_3, f, k_1$, and k_2).

B. Unsupervised Classification Using ISODATA Algorithm With PCA

Given the input data described in the previous sections, vegetation mapping is specified as an unsupervised classification of drone orthomosaic images and point clouds (texture features) to serve as our inputs. We first performed principal component analysis (PCA) for dimensionality reduction of the data. Subsequently, we used the ISODATA algorithm for image classification because of its higher robustness and efficiency relative to the K-means algorithm. Image classification is a partitioning procedure in which all the pixels of an image are clustered or grouped together such that pixels with the same features can belong to the same class and are closely related [17], [18].

1) Principal Component Analysis: PCA is an algorithm commonly employed in data analysis to reduce the dimensions. It is used to reduce a dataset with higher dimensional vectors to a dataset with lower dimensional vectors [19], [20]. The PCA is executed by simultaneously applying both matrix method and data method. PCA involves at least the following four general steps.

1) Find the mean vector in x-space.
2) Assemble covariance matrix in x-space.
3) Compute eigenvalues and corresponding eigenvectors.
4) Form the components in y-space.

PCA compresses information on the number of bands present into number of new bands called principal components to reduce redundancy and increase the covariance, resulting in a dataset of a much lower dimensionality.

2) ISODATA Classification Algorithm: Over the years, the ISODATA algorithm has been widely used for unsupervised classification. It assumes that each cluster follows a multivariate normal distribution. Therefore, the cluster means and covariance matrices need to be computed for the individual clusters. The K-means method is one of the simplest unsupervised learning algorithms that overcomes the clustering problem. The main idea is to determine the cluster centers in the data input, following which each pixel belonging to a given data set is associated to a K-means cluster using the nearest center. Basically, the ISODATA and K-means algorithms both approach randomly assigned cluster centers; new cluster means and covariance are subsequently computed. The new cluster means and covariance are computed for all the pixels belonging to that cluster. This process is performed repeatedly until a change in the iterations that satisfies a certain threshold or insignificant is encountered. The change can be determined either by measuring the distances between the cluster means if it differs from one iteration to the next or by using the percentage of pixels that possess a change between iterations. Both these algorithms are iterative procedures, and the main difference between them is that in the K-means algorithm, the number of clusters is known *a priori*, whereas in the ISODATA algorithm, the number of clusters can vary [21].

In this paper, the applied ISODATA algorithm was modified to account for all possible cases of Euclidean distance from maximization to minimization in order to generate the initial cluster centers and realize high performance [22].

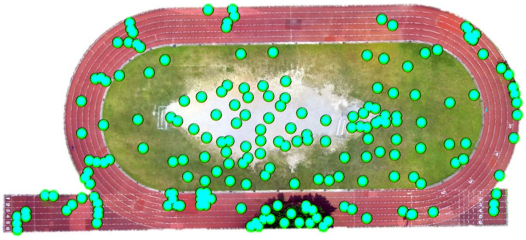


Fig. 10. Reference points were computed from the orthomosaic images as ground truth for use in performance evaluation.

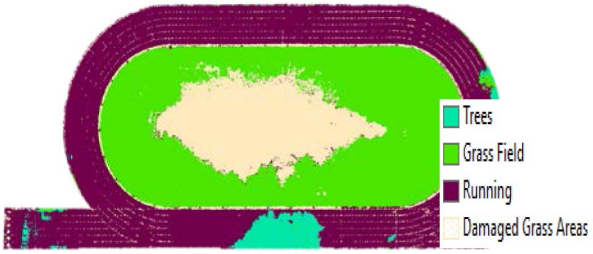


Fig. 11. Classification results from the orthomosaic image.

Fig. 12. Distorted checkerboard calibration pattern.

The Euclidean distance given by D_{ki} was calculated between every pixel and the initial cluster centers that make the classification procedure possible. We computed the Euclidean distance from every pixel to every cluster center in the same band and then summed the distance of all the bands as follows:

$$D_{ki} = \sum_{j=0}^{M-1} \left(M_{kj} - C_{kj}\right)^2 \tag{9}$$

where D_{ki} is the sum of the band's distance between the pixel M_{kj} and the cluster center C_{kj} of the same band.

Therefore, for pixels M_{k*} if $D_{ki} < D_{kj} (i = 0, 1, \ldots, NC - 1 \& i \neq j)$, $M_{kj} \in \Phi_i$, Φ_i is the cluster i with center C_{k*}. By modifying the cluster center value, where N_i is the number of pixels of all the bands in the cluster Φ_i, we have

$$C_{i*} = \frac{1}{N_i} \sum_{M_{k*} \in \Phi_i} M_{k*}, \quad i = 0, 1, \ldots, NC - 1. \tag{10}$$

Next, we calculated the average distance between the pixels of the same cluster and the corresponding cluster center, where W_i is the average distance among pixels of the same cluster

$$W_i = \frac{1}{N_i} \sum_{M_{k*} \in \Phi_i} (M_{k*} - C_{i*}), \quad i = 0, 1, \ldots, NC - 1. \tag{11}$$

Lastly, every pixel of the same cluster was summed to calculate the total average of all the clusters as follows:

$$\overline{W} = \frac{1}{N} \sum_{i=1}^{NC} W_i * N_i, \quad i = 0, 1, \ldots, NC - 1. \tag{12}$$

C. Target Scene Interpretation

The orthomosaic image was visually interpreted; 200 points were extracted as reference points from the ground truth (Fig. 10). We identified four main classes, namely damaged grass areas, grass field, running track, and trees. Before ISODATA clustering, we performed a PCA on the data to reduce the dimensionality of the data. Then, because the number of clusters needs to be known in advance for the ISODATA classification, we used numbers ranging from 4 to 6; the results are more satisfactory when the value is 4. Once the clustering process was complete, the clusters were manually labeled (Fig. 11) to the closest match using the reference image. The ISODATA method generates a cluster map, with clusters assigned arbitrary colors for easy identification. Accordingly, during labeling, each cluster was matched to a class from the reference image and given a unique color. Classes that exist in the cluster map can be easily recognized by their spectral properties. In our experiment, we matched as many classes as possible to the clusters produced by the ISODATA clustering; however, results from only four clusters perfectly matched all the targeted clusters.

IV. Experimental Results and Discussion

We calibrated the drone camera using a checkerboard calibration pattern (Fig. 12) in MATLAB to improve the accuracy and performance of the proposed method. Camera calibration involves the estimation of a camera's intrinsic, extrinsic, and lens-distortion parameters. Subsequently, we used the calibrated camera parameters to accurately measure the size of any object in the image.

The recommended calibration procedure to accurately measure the area of the objects is as follows.

1) Prepare calibration images of the model plane taken under different orientations by moving either the plane or the drone camera.
2) Estimate the camera parameters and evaluate the calibration errors (Fig. 13).
3) Detect the feature points in the images (Fig. 14).
4) Estimate the intrinsic and extrinsic parameters (Fig. 15).
5) Correct lens distortion (e.g., radial lens distortion) parameters.
6) Detect objects in the image by using segmentation and measure their areas by using the calibrated camera parameters (Fig. 16). It took the GCS 0.2 s to detect landmarks on each video frame.

The detected object in our calibration image was a book of surface area 173 cm^2. Next, we evaluated the performance of the proposed method when used for scene detection. On the basis of our datasets (orthomosaic images and point clouds),

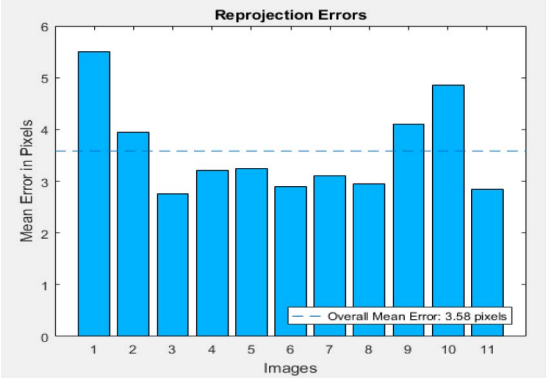


Fig. 13. Visualization of reprojection errors.

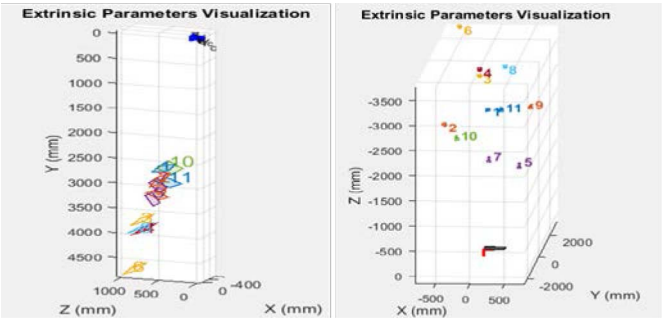


Fig. 14. Visualization of extrinsic parameters. Left: pattern locations. Right: camera locations.

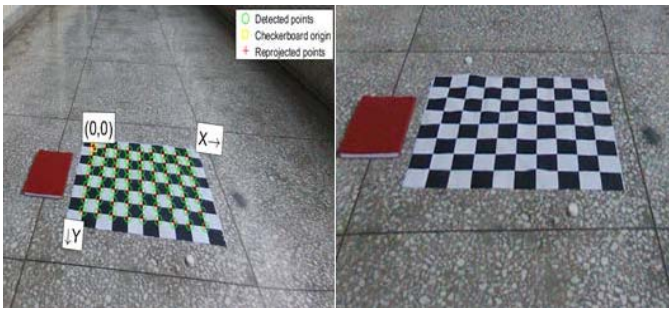


Fig. 15. Left: detected points on checkerboard with distortion. Right: detected points on checkerboard without distortion.

Fig. 16. Left: object segmentation. Right: object detection of undistorted image.

we specified the number of clusters for our experiment. Furthermore, we extracted statistical data from the datasets to evaluate the performance of the proposed method.

A. Model Parameter

In this part of the experiment, the number of clusters for our datasets needs to be provided as prior information. Using the ISODATA classification algorithm, the minimum description length criterion [23], [24] was implemented on the datasets to select the optimum number of clusters as 4.

B. Evaluation of the Results

We further investigated the performance of the proposed method and validated its accuracy against the K-means approach as the reference algorithm. For the K-means algorithm, the number of clusters was set to 4 when processing both orthomosaic images and cloud points.

The main objective of environmental classification is to label each point or pixel of the image to a specific group or cluster. We labeled pixels in the image to belong to one of the four class labels: 1) "damaged grass areas"; 2) "grass field"; 3) "running track"; and 4) "trees." Accordingly, the classification performance was evaluated using the following statistics: true positive (TP), false positive (FP), true negative (TN), and false negative (FN).

We calculated the TP, FP, FN, and NP for the proposed method, and the overall accuracy was compared with that of the K-means algorithms.

To measure the accuracy and performance of our algorithm, we calculated four evaluation indices, namely accuracy, sensitivity, FN ratio (FNR), and FP ratio (FPR) [25]

$$\text{Accurancy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{13}$$

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{14}$$

$$\text{FNR} = \frac{\text{FN}}{\text{FN} + \text{TP}} \tag{15}$$

$$\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}} \tag{16}$$

where accuracy refers to the number of correctly detected points in the entire dataset, sensitivity indicates the number of correctly detected points as either damaged grass areas, grass field, running track, or trees in the ground truth. In addition, FNR and FPR were computed for each land cover in terms of commission–omission errors in the reference point of the ground truth [26].

The image classification results of the ISODATA algorithm and the corresponding execution time increased exponentially as the number of iterations increased. A comparison of the classified image against the ground truth pixels reveals that the ISODATA algorithm is more precise and accurate because each pixel in the image was correctly classified. The overall accuracy of the classification process using the unsupervised ISODATA algorithm with PCA was 84.0%.

The grass field pixels were classified with 100% accuracy (Tables I and II). The tree pixels were misclassified as grass field pixels, thereby lowering the sensitivity of the proposed method (Tables III and IV) for the entire experiment in the study area. In addition, the evaluation of the three indices was high (Table V). The proposed method shows that sensitivity is

TABLE I
CONFUSION MATRIX FOR THE PROPOSED METHOD USING ISODATA WITHOUT PCA

	Grass Field	Damaged Area	Running Track	Trees
Grass Field	50	0	0	0
Damaged Area	5	39	6	0
Running Track	0	0	41	9
Trees	23	0	0	27

TABLE II
CONFUSION MATRIX FOR THE PROPOSED METHOD USING ISODATA WITH PCA

	Grass Field	Damaged Area	Running Track	Trees
Grass Field	50	0	0	0
Damaged Area	3	47	0	0
Running Track	0	0	42	8
Trees	19	0	2	29

TABLE III
EVALUATION OF THREE INDICES USING ISODATA WITHOUT PCA

Land cover	Sensitivity	FNR	FPR
Grass Field	100%	0%	56%
Damaged Area	78%	22%	0%
Running Track	82%	18%	12%
Trees	54%	46%	18%

TABLE IV
EVALUATION OF THREE INDICES USING ISODATA WITH PCA

Land cover	Sensitivity	FNR	FPR
Grass Field	100%	0%	44%
Damaged Area	94 %	6%	0%
Running Track	84%	16%	4%
Trees	58%	42%	16%

TABLE V
OVERALL ACCURACY OF THE PROPOSED AND THE *K*-MEANS METHODS

Method	Overall Accuracy
Proposed Method	84.0%
ISODATA without PCA	78.5%
K-means	76.0%

much higher when using ISODATA algorithm with PCA than when using the ISODATA without PCA, and this result is very encouraging. Moreover, FNR and FPR of the proposed method are lower when the ISODATA and PCA algorithms are used. Furthermore, the proposed method has an overall accuracy higher than that of the *K*-means algorithm, which implies that the proposed method has higher accuracy in detecting and classifying pixels in an image than does the *K*-means algorithm.

C. Area Estimation of the Damaged Grass Field

We proposed an automatic method to calculate the area of the damaged field after classification. The ROI pixels were selected and grouped together to form contours and were used to estimate the area of the damaged grass field. To estimate the area of the ROI, we calculated some statistical values for pixels of the input raster inside each target contour. The area is calculated as

$$\text{Area} = \text{Pixel count} * (\text{Cellsize})^2 \tag{17}$$

where Cellsize is a constant value of 7.84 m, which is the cell resolution of the raster image, and the pixel count is the sum of all pixels that belong to the ROI. The area of the damaged field was calculated as shown in Fig. 17.

TABLE VI
SIZE OF THE CLASSIFIED AREAS

Land cover	Area (m^2)
Damaged Area	602.34
Grass Field	1441.40
Estimated Grass Area	2043.74
Actual Grass Area	2292.00

Fig. 17. Contours representing pixels in the damaged grass field.

Approximately 40% of the soccer field was damaged (Table VI), indicating that the soccer field is poorly managed and needs to be fixed urgently.

Subsequently, after determining the area of the damaged field, we estimated the amount of grass that must be planted to restore the soccer field. To validate the accuracy of the results, we used the actual known grass area measured using the GPS information against the summation of the damaged area and the grass area from the classification, and the error rate was calculated as follows:

$$\text{Error} = \frac{|\text{Actual Grass Area} - \text{Estimated Grass Area}|}{\text{Actual Grass Area}} 100\%. \tag{18}$$

The error rate, which reveals the number of pixels that were incorrectly classified as either damaged or grass pixels, was 10.83% (Table VI), indicating that most of the pixels in the field were correctly classified by the proposed algorithm.

D. Car and Vacant Parking Space Detection

The proposed method was also used as an image-based system to automatically detect both cars and vacant parking spaces in a parking lot to facilitate smart parking. It is aimed at providing car drivers with reliable intelligent parking information and guiding them to the nearest vacant parking space to have easy and efficient parking. We consider the similar architecture and algorithms used for the landmark detection and recognition system for this application. HOG features are computed by dividing the input images, each resized to 124×64 pixels from dataset containing both images of cars and vacant parking spaces, into a set of overlapping cells.

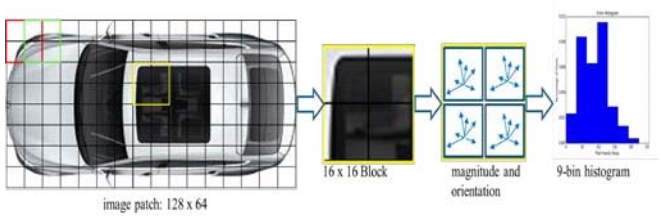


Fig. 18. Example of car HOG feature extraction process.

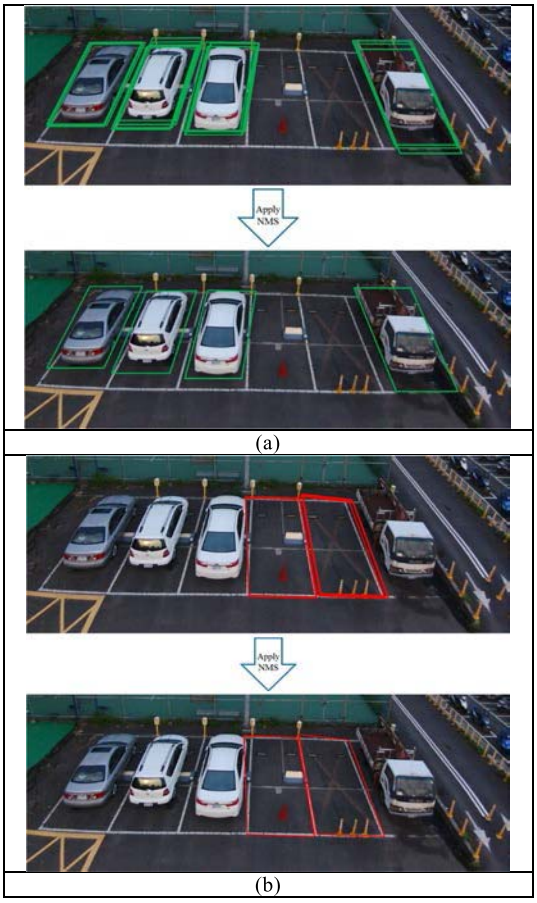


Fig. 19. Result from parking lot UAV images. (a) Car detection. (b) Vacant parking spaces.

In each cell the histograms of gradient directions are computed as shown in Fig. 18. The histograms are further grouped together to form the descriptors and then normalized to make them more invariant to illumination and shadowing changes over blocks of $N \times N$ cells. For this experiment, the HOG parameters used were nine bins on cells of 8×8 pixels, with a block size of 2×2 cells overlapping by 50%. A dataset of 2000 images were collected using the UAV over an open parking space during daylight at an altitude of approximately 40 m. One thousand six hundred images from the dataset were used for training and the rest 400 for testing using their HOG descriptors and linearSVM with C parameter set to 100. After applying the linearSVM classifier to the HOG features extracted from the sliding windows over the image pyramid which is the multiscale representation of the UAV image, a number of bounding boxes are derived by a threshold to the prediction score as demonstrated in Fig. 19. A nonmaximum suppression algorithm is used to fuse the overlapped detection with an overlapping threshold of 0.1 over all scales into one final result. In Fig. 19(a), the result shows that all the four cars in the parking lot were correctly detected and in Fig. 19(b), both two vacant parking spaces were also correctly detected. The overall accuracy achieved for the entire dataset is 98%.

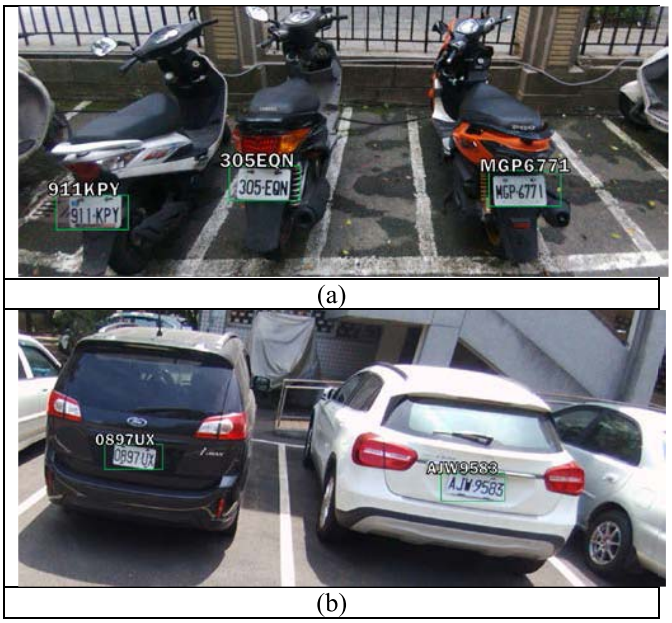


Fig. 20. Detection results from license plate images taken by UAV for (a) motorcycles and (b) automobiles.

E. Automatic License Plate Recognition

The proposed method was further used for an automatic license plate recognition (ALPR) system designed to detect and recognize license plates for vehicles in the target areas. We propose a unified approach that integrates the three general ALPR steps (license plate detection, character segmentation, and optical character recognition) via the use of a deep convolutional neural network (CNN) that operates directly on the image pixels [27]. In this method, the CNN is used as a feature extractor and classifier to recognize alphabets (A–Z) and digits (0–9) that constitute the license plate character combinations. Our model was trained on a set of training dataset labeled with expected outputs generated from synthetic images to represent the vehicle license plates in the different styles.

A dataset of 25 000 synthetic license plate images each of size 128×64 pixels were generated from 30 000 randomly selected background images sourced from ImageNet dataset [28]. After training the model over 20 000 images, the test accuracy from 5000 images achieved 99% from detection of license plates with either six or seven characters. Fig. 20(a) and (b) shows that all license plate characters for the motorcycles and automobiles, respectively, were correctly recognized.

V. Conclusion

We presented an unsupervised classification method that uses high-resolution images to detect and differentiate the

geo-object classes (scene interpretation) by using spectral information in the target or investigated area; this system was supported by a landmark detection and recognition system for autonomous drone navigation. Because the classification method is unsupervised, no prior knowledge about the investigated environment is required. The proposed method outperformed the reference model in the experiments. Furthermore, we proposed a method to autonomously navigate a drone in very large environments. The proposed classification method, which uses high-resolution orthomosaic images and 3-D point clouds, yielded satisfactory experimental results; however, the overall accuracy of the method can still be improved.

In the future, supervised or semisupervised training of the models as well as their application to different target environments can be investigated. 3-D collision detection and obstacle avoidance can be considered to enhance the safe landing of drone. Deep learning algorithms can be implemented in landmark detection and recognition systems to improve the overall accuracy of the method. Furthermore, using the drone's onboard CPU rather than the GCS may improve the connection and efficiency of the algorithm for use in real-time landmark detection and recognition systems.

REFERENCES

[1] A. Kato *et al.*, "Fusion between UAV-SFM and terrestrial laser scanner for field validation of satellite remote sensing," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Milan, Italy, Jul. 2015, pp. 2642–2645.

[2] F. Nex and F. Remondino, "UAV for 3D mapping applications: A review," *Appl. Geomatics*, vol. 6, no. 1, pp. 1–15, Mar. 2014.

[3] K. Dorling, J. Heinrichs, G. G. Messier, and S. Magierowski, "Vehicle routing problems for drone delivery," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 47, no. 1, pp. 70–85, Jan. 2017.

[4] Y. Fu, M. Ding, C. Zhou, and H. Hu, "Route planning for unmanned aerial vehicle (UAV) on the sea using hybrid differential evolution and quantum-behaved particle swarm optimization," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 43, no. 6, pp. 1451–1465, Nov. 2013.

[5] T. Moranduzzo and F. Melgani, "Detecting cars in UAV images with a catalog-based approach," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 10, pp. 6356–6367, Oct. 2014.

[6] 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.*, vol. 1. Kauai, HI, USA, Dec. 2001, pp. 511–518.

[7] B. Ranft, J.-L. Dugelay, and L. Apvrille, "3D perception for autonomous navigation of a low-cost MAV using minimal landmarks," in *Proc. Int. Micro Air Veh. Conf. Flight Competition*, Sep. 2013, pp. 17–20.

[8] Y. Zhuang, N. Jiang, H. Hu, and F. Yan, "3-D-laser-based scene measurement and place recognition for mobile robots in dynamic indoor environments," *IEEE Trans. Instrum. Meas.*, vol. 62, no. 2, pp. 438–450, Feb. 2013.

[9] X. Zhang, Y. Zhuang, H. Hu, and W. Wang, "3-D laser-based multiclass and multiview object detection in cluttered indoor scenes," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 1, pp. 177–190, Jan. 2017.

[10] H. Sima, P. Guo, Y. Zou, Z, Wang, and M. Xu, "Bottom-up merging segmentation for color images with complex areas," *IEEE Trans. Syst., Man, Cybern., Syst.*, to be published.

[11] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 1. San Diego, CA, USA, Jun. 2005, pp. 886–893.

[12] C. Strecha and O. Küng. (2011). [Online]. Available: http://www.pix4d.com

[13] U. Niethammer, M. R. James, S. Rothmund, J. Travelletti, and M. Joswig, "UAV-based remote sensing of the Super-Sauze landslide: Evaluation and results," *Eng. Geol.*, vol. 128, pp. 2–11, Mar. 2012.

[14] M. Brown and D. G. Lowe, "Unsupervised 3D object recognition and reconstruction in unordered datasets," in *Proc. 5th Int. Conf. 3-D Digit. Imag. Model.*, Ottawa, ON, Canada, Jun. 2005, pp. 56–63.

[15] F. Fraundorfer, "Building and site reconstruction from small scale unmanned aerial vehicles (UAV's)," in *Proc. Joint Urban Remote Sens. Event (JURSE)*, Lausanne, Switzerland, Mar./Apr. 2015, pp. 1–4.

[16] H. Wang, J. Li, L. Wang, H. Guan, and Z. Geng, "Automated mosaicking of UAV images based on SFM method," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Quebec City, QC, Canada, Jul. 2014, pp. 2633–2636.

[17] M. Cai-Hong, D. Qin, and L. Shi-Bin, "A hybrid PSO-ISODATA algorithm for remote sensing image segmentation," in *Proc. Int. Conf. Ind. Control Electron. Eng.*, Xi'an, China, Aug. 2012, pp. 1371–1375.

[18] B. Li, H. Zhao, and Z. H. Lv, "Parallel ISODATA clustering of remote sensing images based on MapReduce," in *Proc. Int. Conf. Cyber Enabled Distrib. Comput. Knowl. Disc.*, Huangshan, China, Oct. 2010, pp. 380–383.

[19] R. Kadmon, *Remote Sensing and Image Processing*. Academic Press, 2001, pp. 121–143.

[20] S. A. El Rahman, "Hyperspectral imaging classification using ISODATA algorithm: Big data challenge," in *Proc. of 5th Int. Conf. e-Learn.*, Manama, Bahrain, Oct. 2015, pp. 247–250.

[21] A. Ahmad and S. F. Sufahani, "Analysis of landsat 5 TM data of Malaysian land covers using ISODATA clustering technique," in *Proc. IEEE Asia–Pac. Conf. Appl. Electromagn.*, Malacca, Malaysia, Dec. 2012, pp. 92–97.

[22] Q. Wang, Q. Li, H. Liu, Y. Wang, and J. Zhu, "An improved ISODATA algorithm for hyperspectral image classification," in *Proc. 7th Int. Congr. Image Signal Process.*, Dalian, China, Oct. 2014, pp. 660–664.

[23] I. O. Kyrgyzov, O. O. Kyrgyzov, H. Maître, and M. Campedel, "Kernel MDL to determine the number of clusters," in *Proc. Int. Conf. Mach. Learn. Data Min. Pattern Recognit.*, Leipzig, Germany, Jul. 2007, pp. 203–217.

[24] W. Yi, H. Tang, and Y. H. Chen, "An object-oriented semantic clustering algorithm for high-resolution remote sensing images using the aspect model," *IEEE Geosci. Remote Sens. Lett.*, vol. 8, no. 3, pp. 522–526, May 2011.

[25] A. Madabhushi and D. N. Metaxas, "Combining low-, high-level and empirical domain knowledge for automated segmentation of ultrasonic breast lesions," *IEEE Trans. Med. Imag.*, vol. 22, no. 2, pp. 155–169, Feb. 2003.

[26] S. Li *et al.*, "Unsupervised detection of earthquake-triggered roof-holes from UAV images using joint color and shape features," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 9, pp. 1823–1827, Sep. 2015.

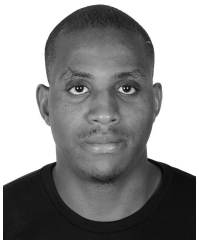
[27] I. J. Goodfellow, Y. Bulatov, J. Ibarz, S. Arnoud, and V. Shet, *Multi-Digit Number Recognition From Street View Imagery Using Deep Convolutional Neural Networks*, Google Inc., Mountain View, CA, USA, Dec. 2013.

[28] J. Deng *et al.*, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Miami, FL, USA, Jun. 2009, pp. 248–255.

Yo-Ping Huang (S'88–M'92–SM'04) received the Ph.D. degree in electrical engineering from Texas Tech University, Lubbock, TX, USA.

He is currently a Professor with the Department of Electrical Engineering, National Taipei University of Technology, Taipei, Taiwan, where he was the Secretary General from 2008 to 2011. He was a Professor and the Dean of research and development from 2005 to 2007, the Dean of the College of Electrical Engineering and Computer Science from 2002 to 2005, and the Department Chair from 2000 to 2002 with Tatung University, Taipei. His current research interests include intelligent systems and modeling, Internet of Things, medical data mining, deep learning, and rehabilitation systems design.

Prof. Huang serves as the President of the Taiwan Association of Systems Science and Engineering, and the Chair of the IEEE SMC Taipei Chapter. He was the Chair of the IEEE CIS Taipei Chapter, and the CEO of the Joint Commission of Technological and Vocational College Admission Committee, Taiwan, from 2011 to 2015. He is an IET Fellow in 2008, and an International Association of Grey System and Uncertain Analysis Fellow in 2016.



Lucky Sithole was born in Swaziland, in 1988. He received the B.Sc. degree in computer science from the University of Swaziland, Kwaluseni, Swaziland, in 2013, and the M.Sc. degree in electrical engineering and computer science from the National Taipei University of Technology, Taipei, Taiwan, in 2017.

In 2013, he joined the Computer Science Department, Limkokwing University of Creative Technology, Cyberjaya, Malaysia, and the Swaziland College of Technology, Mbabane, Swaziland, respectively, as a Lecturer. In 2015, he joined Sun International, Valley Mbabane, Swaziland, as an IT Engineer. His current research interests include in computer vision and image processing, machine learning and deep learning, unmanned aerial vehicle systems, and robotics programming.

Tsu-Tian Lee (F'97) received the Ph.D. degree in electrical engineering from the University of Oklahoma, Norman, OK, USA, in 1975.

He is currently the National Endow Chair of the Ministry of Education, Tamkang University, New Taipei City, Taiwan, where he has been a Chair Professor with the Department of Electrical Engineering since 2014. He had served as a Professor and the Chairman of the Department of Control Engineering, National Chiao Tung University, Hsinchu, Taiwan, and a Full Professor of electrical engineering with the University of Kentucky, Lexington, KY, USA. He was the President of the National Taipei University of Technology, Taipei, Taiwan.

Dr. Lee was a recipient of the National Endow Chair from the Ministry of Education, Taiwan, in 2003 and 2006, the TECO Science and Technology Award from TECO Technology Foundation in 2003, and the IEEE SMC Society Norbert Wiener Award in 2009. He was elected as an IET Fellow in 2000, and a fellow of the New York Academy of Sciences in 2002, the Chinese Automatic Control Society in 2007, and the Institute of Complex Medical Engineering in 2015.