Intel Serv Robotics (2014) 7:79–92
DOI 10.1007/s11370-014-0146-x

SPECIAL ISSUE

3D reconstruction and classification of natural environments by an autonomous vehicle using multi-baseline stereo

Annalisa Milella · Giulio Reina

Received: 12 August 2013 / Accepted: 27 January 2014 / Published online: 2 March 2014

Abstract In natural outdoor settings, advanced perception systems and learning strategies are major requirement for an autonomous vehicle to sense and understand the surrounding environment, recognizing artificial and natural structures, topology, vegetation and drivable paths. Stereo vision has been used extensively for this purpose. However, conventional single-baseline stereo does not scale well to different depths of perception. In this paper, a multi-baseline stereo frame is introduced to perform accurate 3D scene reconstruction from near range up to several meters away from the vehicle. A classifier that segments the scene into navigable and non-navigable areas based on 3D data is also described. It incorporates geometric features within an online self-learning framework to model and identify traversable ground, without any a priori assumption on the terrain characteristics. The ground model is automatically retrained during the robot motion, thus ensuring adaptation to environmental changes. The proposed strategy is of general applicability for robot's perception and it can be implemented using any range sensor. Here, it is demonstrated for stereo-based data acquired by the multi-baseline device. Experimental tests, carried out in a rural environment with an off-road vehicle, are presented. It is shown that the use of a multi-baseline stereo frame allows for accurate reconstruction and scene segmentation at a wide range of visible distances, thus increasing the overall flexibility and reliability of the perception system.

A. Milella (✉)
Institute of Intelligent Systems for Automation (ISSIA), National Research Council (CNR), Via Amendola 122/D-O, 70126 Bari, Italy
e-mail: milella@ba.issia.cnr.it

G. Reina
Department of Engineering for Innovation, University of Salento, Via Arnesano, 73100 Lecce, Italy
e-mail: giulio.reina@unisalento.it

Keywords Field robotics · Multi-baseline stereovision · 3D environment reconstruction and classification

1 Introduction

In recent years, vehicles that can drive autonomously in natural environments, such as agricultural, forested, and rural settings, have received increasing interest. In these contexts, autonomous navigation presents many challenges, due to the lack of structured elements that makes most of the conventional navigation approaches unfeasible [1]. The high variability of the terrain characteristics and environmental conditions, e.g., different vegetation types, changes in the illumination conditions and weather phenomena further complicates the design of even basic functionalities, including obstacle detection, mapping and path planning. Therefore, advanced perception devices and online environment learning strategies are primarily required for the vehicle to operate safely and reliably.

Stereovision is a widely adopted input for outdoor navigation, as it provides an effective technique to extract range information and perform complex scene understanding tasks [2–6]. Nevertheless, the accuracy of stereo reconstruction is generally affected by various design parameters, such as the baseline, i.e., the distance between the optical centers of two cameras in a stereo head [7,8]. A larger baseline guarantees higher accuracy at each visible distance, but it leads to a loss of information in the near range. In addition, a long baseline requires a larger disparity search range, which implies a greater possibility of false matches. Hence, the choice of the optimal baseline results from the balance of opposing factors, depending on the requirements of the target application.

In this paper, a multi-baseline stereo frame is proposed, which allows an autonomous vehicle that operates in

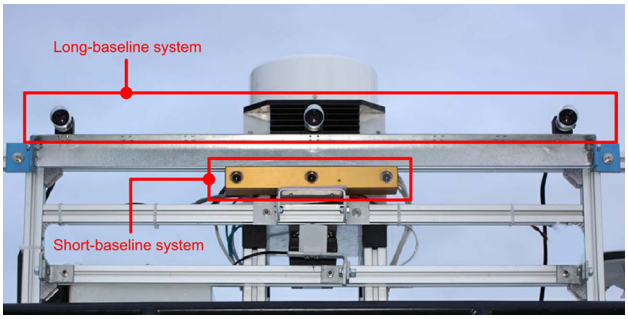

Fig. 1 The multi-baseline stereo vision system

Fig. 2 Experimental test bed provided by IRSTEA and used for field validation in the QUAD-AV project

natural settings to perform accurate 3D scene reconstruction and segmentation in a wide range of distances. The system was implemented within the project Ambient Awareness for Autonomous Agricultural Vehicles (QUAD-AV) funded by the ERA-NET ICT-AGRI action, aimed to enable safe autonomous navigation in high-vegetated, off-road terrain [9]. The developed stereo frame is shown in Fig. 1. It is composed by two trinocular heads, one featuring a short baseline system and the other one featuring a long baseline system. By employing the narrow baseline to reconstruct nearby points and the wide baseline for more distant points, this system takes the advantage of the small minimum range of the narrow baseline, while preserving the higher accuracy and maximum range of the wide baseline configuration. The two trinocular cameras can be either used simultaneously to widen the overall perception range of the vehicle, or alternately depending on the vehicle travel conditions. For instance, the narrow baseline configuration is useful in low-speed operations, where less noisy measurements are needed, while the wide baseline is suitable when the vehicle travels at higher speed, enabling it to perceive far away obstacles [10]. In addition, the wide baseline can improve the quality of the stereo range data for distant terrain mapping [11]. Therefore, the use of a multi-baseline stereo frame allows one to get good results from the near range up to several meters away from the vehicle, and to increase the overall flexibility and reliability of the system.

The 3D point cloud returned by either trinocular camera provides a rich source of information for the vehicle to perform key navigation tasks, such as terrain identification and scene segmentation. In this research, a geometry-based classifier is described that uses geometric features extracted by a 3D point cloud to segment the scene into two broad classes: ground and non-ground. The ground class denotes points from traversable terrain, whereas the non-ground class corresponds to all other data, including points from non-traversable ground, above ground objects (i.e., obstacles) or occluded areas, and poor sensor reconstructions. The performance of the classifier is demonstrated for stereo reconstructed points, provided by both the short-range and the long- range trinocular sensor. In detail, given 3D points, the system, first, maps them to cells and extracts geometric features of the points in each cell. Then, these features are used to label single cells as ground or non-ground patches. The classifier adopts a self-learning framework, whereby the ground model is automatically built through an initial bootstrapping stage and is continuously retrained to incorporate changes in the ground characteristics. During the training stage, the classifier learns to associate the geometric appearance of data with class labels. Then, it makes predictions based on past observations classifying new acquired data.

For the experimental verification of the proposed system, the multi-baseline stereo frame was integrated with an off-road vehicle (see Fig. 2) that was made available by the partner National Research Institute of Science and Technology for Environment and Agriculture (IRSTEA) at the Montoldre farm facility [12]. The vehicle's sensor suite included, as well, a 3D SICK laser rangefinder, a frequency modulated continuous wave (FMCW) radar, and a thermal infrared camera.

The remainder of this paper is organized as follows. Section 2 describes related work. Section 3 provides details about the implementation of the multi-baseline stereo system, including design and calibration issues, and stereo processing algorithms. In Sect. 4, the ground detection approach is introduced. Experimental results showing the performance of the 3D reconstruction and segmentation algorithms on field are presented in Sect. 5. Section 6 concludes this paper.

2 Related work

Stereo cameras are among the most widely adopted sensing devices for vehicle perception in unstructured outdoor environments, since they generally provide dense depth maps at relatively high frequency. Stereo range data are co-registered with intensity information, thus allowing the vehicle to learn

color and texture models of the environment that can be used to perform critical tasks, including obstacle classification and terrain typing [13–15]. However, for typical camera configurations and resolutions, stereo systems are generally accurate only up to 10–12 m. Higher reconstruction distances may be achieved by increasing the baseline. For instance, in [11], a wide-baseline stereo vision system is proposed that uses subsequent images from a monocular camera and a visual odometry algorithm along with stereo-matching methods, for accurate distant terrain mapping on sandy and rocky soil. In [10], a trinocular camera featuring three baselines, with the largest one of 1.5 m, is adopted to perform robust obstacle and lane detection in the long range on urban roads. While it improves accuracy in the far range, a large baseline moves farther the point of view of the stereo pair, thus determining a loss of information in the near range. To deal with this issue, in [10], two additional stereo systems are used in conjunction with the trinocular camera for precise detection of obstacles and lane markings in the vicinity of the vehicle. Solutions based on variable baseline stereo, whereby the baseline can be changed adaptively according to the operation conditions, have also been proposed. One of the first works on variable baseline stereo is the slider stereo by Moravec [16]. It consists of a monocular camera that is shifted along a track to acquire multiple snapshots of a scene; then, each possible image pair is considered as a stereo baseline, and is used for feature depth estimation. All estimates are accumulated in a histogram, whose peak is finally chosen as the best distance estimation. In [7], a stereo-matching approach using multiple baselines obtained by a lateral displacement of a camera is proposed to improve the accuracy of 3D estimation and remove ambiguities based on the SSSD-in-inverse-distance function. In [8], a variable baseline/resolution stereo method is developed. It uses multiple images to vary the baseline and resolution proportionally to depth and obtain a reconstruction with constant depth error. An adaptive variable baseline stereo system is developed in [17], where the baseline is modified depending on local spatial frequency content. A high-speed linear slider to vary the stereo baseline is proposed in [18]. Two cameras are independently moved along the slider, to adaptively change the baseline length according to the distance of the object to track. Active camera positioning is advantageous, as it avoids the use of multiple stereo devices, while guaranteeing, at the same time, good accuracy at different depths of perception. However, active stereo entails the use of linear and rotating actuators that must be precisely controlled via real-time visual servoing algorithm, which is not always feasible in high-speed autonomous vehicle applications.

Another critical parameter for stereo reconstruction is the lens focal length: a short focal length increases the angular field of view, but induces higher distortion. It also increases (i.e., makes it worse) the range resolution. Lenses with larger focal lengths produce images that are zoomed in farther, allowing for the detection of distant objects. Nevertheless, the greater the focal length, the narrower the field of view. In this work, both the baseline and the focal length are taken into account as design parameters. Different baselines and optics are combined in a multi-camera framework featuring two trinocular cameras, one with narrow baseline and short focal length to be used to reconstruct regions close to the vehicle, and the other one with wide baseline and long focal length to reconstruct more distant portions of the environment.

When dealing with vision systems in outdoor settings, a major issue is the diversity of terrain and lighting conditions, which makes it unfeasible to employ predefined templates or features. In this respect, the use of machine learning techniques relying on online training approaches may be helpful. Recent investigations have tried to solve both the short-sightedness problem of stereovision and the need for online environment learning and modeling techniques, by developing self-supervised near-to-far learning approaches. Self-supervised methods have the advantage of reducing or eliminating the need for hand-labeled training data, as the training set is automatically produced by an additional classification module [19–23] or by self-teaching strategies [24], thus gaining flexibility in unknown environments. For instance, in [25], the 3D point cloud produced by a stereo device is processed to extract the ground plane in the vicinity of the vehicle; then, points belonging to the ground plane are projected onto a monocular image to train a color-based classifier for long-range classification. In [26], a stereo algorithm produces a 3D point cloud; then, ground plane and footline estimation methods are applied to classify these points as ground, obstacle, or footline. Visual features are successively extracted by projecting the labeled points onto images, and are used for prediction in the long range.

Research on self-learning ground detection using stereovision has been developed by the authors in previous work. Specifically, in [6], a self-learning geometry-based classifier is proposed to detect the broad class of ground using "conventional" stereo reconstruction in the short range (up to 18 m). In this work, the same self-learning framework is extended to the case of a multi-baseline stereo set up. In contrast to previous research, where near- to long-range extension was applied only to the visual image, here also reliable range information is available at several meters away from the vehicle, thus making the system especially useful for efficient goal-driven planning and driving. In addition, different baselines can be selected according to the operational conditions of the vehicle, thus ensuring higher system adaptability and flexibility. It should also be noted that in contrast to most of the algorithms in the literature [25–27], the proposed approach does not require ground plane reasoning and it aims to detect scene regions that are traversable safe for the vehicle rather than attempting to explicitly identify obstacles [2,15].

Table 1 Specifications of the Bumblebee XB3

Sensor	Baseline	Resolution and FPS	Focal length	Field of view
Three Sony 1/3" CCD color	12 cm/24 cm	1,280 × 960 pixel at 15 FPS	3.8 mm	66°(H) × 50°(V)

Table 2 Specifications of the custom-built trinocular system

Sensor	Baseline	Resolution and FPS	Focal length	Field of view
Three Sony 1/2" CCD color	40 cm/80 cm	1,384 × 1,032 pixel at 16 FPS	12.0 mm	30°(H) × 23°(V)

3 Multi-baseline stereo vision

In this section, first, details concerning the implementation and calibration of the multi-baseline stereo frame are provided, then, the stereo processing algorithms for 3D scene reconstruction are presented.

3.1 Description of the system

The system comprises two trinocular cameras, featuring four baselines, two for each of them, covering the short range and the medium–long range, respectively. The system is shown in Fig. 1. The short-range camera is the Bumblebee XB3 by Point Grey. It consists of a trinocular stereo head with 3.8 mm focal length lenses, featuring two stereo configurations: a narrow stereo pair with a baseline of 0.12 m (XB3-Narrow) using the right and middle cameras, and a wide stereo pair with a baseline of 0.24 m (XB3-Wide) using the left and right cameras. The second trinocular system is custom built. It comprises three Flea3 cameras by Point Grey with 12.0 mm focal length lenses, disposed in line on an aluminum bar to form two baselines: a narrow baseline of 0.40 m (Flea3-Narrow) using the left and middle cameras and a wide baseline of 0.80 m (Flea3-Wide), using the left and right cameras. Additional technical details are provided in Tables 1 and 2 for the XB3 and the Flea3, respectively.

3.2 Reconstruction error

The theoretical percentage error $E_{z_c}(\%)$ in the reconstruction along the direction of the camera optical axis (z_c-axis) can be calculated as [8]:

$$E_{z_c}(\%) = \frac{E_{z_c}}{z_c} \times 100 \tag{1}$$

where

$$E_{z_c} = \frac{z_c^2}{B \cdot \text{StereoF}} \cdot \text{corrAcc} \tag{2}$$

where B is the baseline, StereoF the focal length in pixels, corrAcc the correlation accuracy (i.e., the matching error in pixels). The percentage reconstruction error expressed by Eqs. (1) and (2), assuming an image resolution of 640 × 480 pixels and a correlation accuracy of 0.2 pixels, is shown in the graph of Fig. 3, for each stereo pair of the system. It can be observed that the reconstruction accuracy decreases with the range and improves at higher baseline and focal length. Therefore, by combining different baselines and optics, it is possible to keep low reconstruction error at a wide range of distances. In particular, using the wide baseline device, a theoretical range error less than 2 % can be obtained up to approximately 100 m. On the other hand, for a disparity search range of 128 pixels, the closest distance along the camera optical axis that can be reconstructed by the XB3 camera is of about 0.5 m, while it is of about 5 m for the Flea3 system. Hence, using a wide baseline, information in the near range is lost, while it can be preserved by adopting a short baseline. In addition, a longer focal length moves farther the viewpoint and restricts the angular field of view of the system, thus making even more significant the loss of information in the short range.

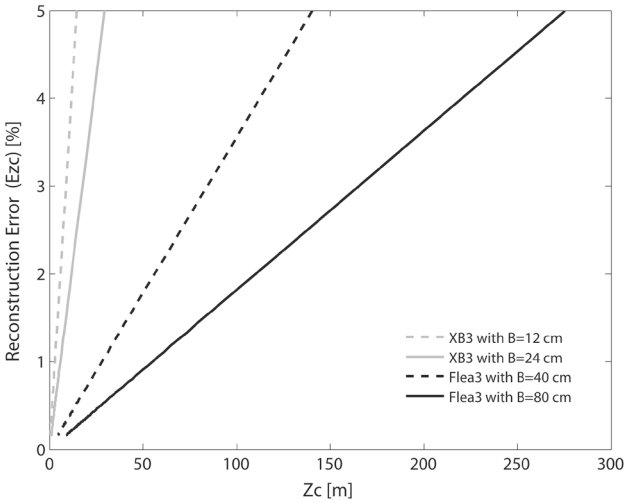

Fig. 3 Theoretical percentage reconstruction error for the multi-baseline system

In this investigation, the two cameras were mounted onboard an off-road vehicle and were used separately for

3D reconstruction and segmentation of a rural environment. Specifically, the XB3 was employed to get information in the short range up to 30 m away from the vehicle, while the Flea3 was used to survey farther regions up to distances of 60 m.

3.3 Calibration

Each stereo pair was calibrated using the OpenCV calibration functions [28]. Both intrinsic and extrinsic parameters were estimated based on a set of images of a planar checkerboard that was appropriately moved across the field of view of each stereo system. The calibration functions also returned the rectification matrices to rectify the images as a preliminary step before applying the stereo-matching algorithm. Since the four stereo pairs were calibrated separately, an additional calibration step was successively performed to align all the systems with respect to a common reference frame attached to the vehicle. To this aim, the calibration pattern was positioned at a known location with respect to the vehicle. Then, the extrinsic parameters relative to the pattern and, consequently, the position and orientation relative to the vehicle were inferred, for each stereo pair, using a least-squares optimization process.

3.4 Stereo processing algorithms

Since the two trinocular systems have very different field of view also due to the use of different lenses (see Fig. 4 as an example), the wide and narrow baseline of each trinocular camera are integrated separately, so that, in the end, two point clouds are obtained: one from the Flea3 system, to be used to get accurate information in the long range, and the other one from the XB3 system to get accurate information in the short range.

For each stereo pair, the stereo processing algorithm includes the following steps:

- Rectification: each image plane is transformed so that pairs of conjugate epipolar lines become collinear and parallel to one of the image axes. Using rectified images, the problem of computing correspondences is reduced from a 2D to a 1D search problem, typically along the horizontal raster lines of the rectified images. Rectification matrices were computed in the calibration step as described in Sect. 3.3.
- Disparity map computation: to compute the disparity map, a stereo block-matching algorithm is used that finds corresponding points by a sliding Sum of Absolute Difference (SAD) window [28].
- 3D point cloud generation in the reference camera frame: being the stereo pair calibrated both intrinsically and extrinsically, disparity values can be converted in depth values and 3D coordinates can be computed in the reference camera frame for all matched points.
- Transformation from the reference camera frame to the vehicle reference frame: 3D points are transformed from the camera frame to the vehicle frame, using the transformation matrix resulting from the calibration process.
- Statistical filtering: a statistical filter is applied to reduce noise and remove outlying points.
- Voxelization: to decrease the computational burden, the number of points is reduced using a voxelized grid approach. A 3D voxel grid is created over the input point cloud space. Then, all the points in each voxel are approximated with their centroid.

The point clouds reconstructed by the narrow pair and by the wide pair of each trinocular sensor are fused in a unique point cloud: if a point of the scene has been reconstructed

Fig. 4 Sample images acquired during experimentation by the Flea3 system (**a**) and the XB3 system (**b**). It can be noted that the angular field of view of the XB3 is wider than the one of the Flea3 due to the use of a shorter focal length. However, the Flea3 image displays more clearly far away objects

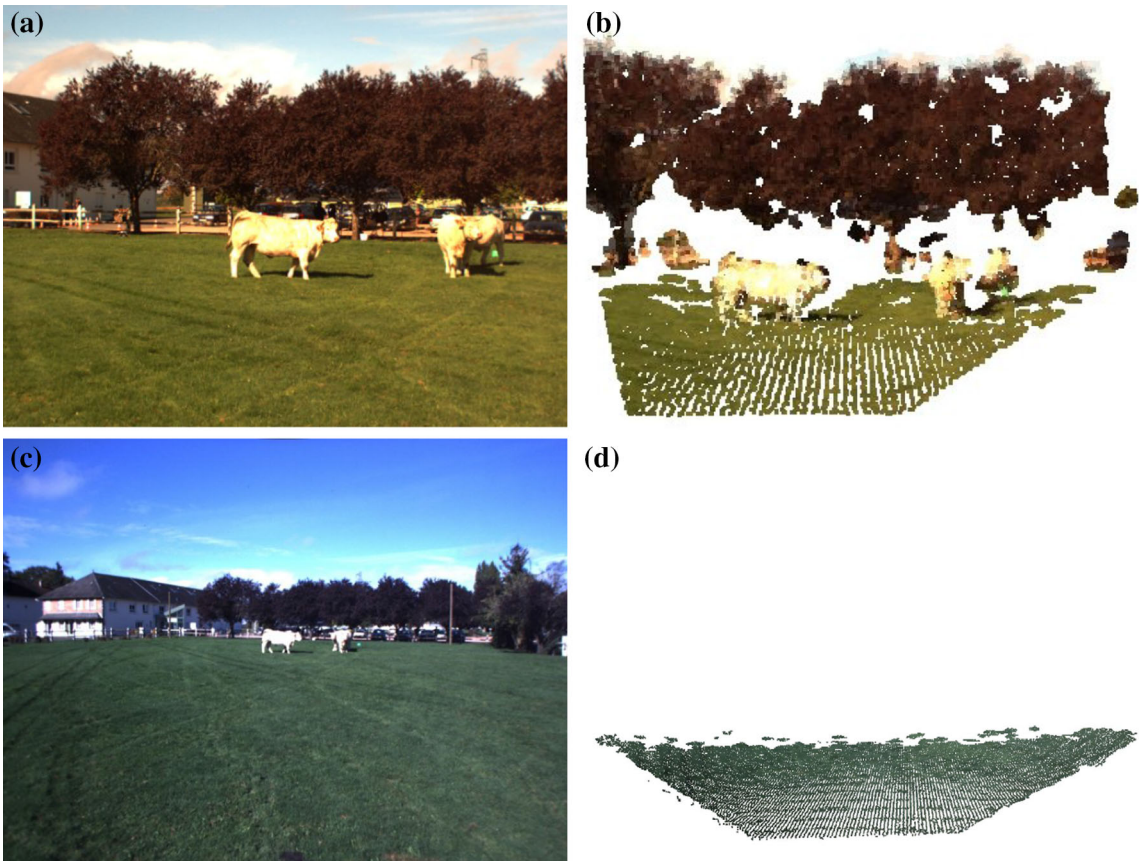

Fig. 5 Reference images acquired by the Flea3 (**a**) and the XB3 (**c**) in a typical test scenario, and corresponding results of the stereo processing algorithm in the long range (**b**) and in the short range (**d**)

by both the wide baseline pair and the narrow baseline pair, only information coming from the wide baseline is retained, since a wider baseline generally assures better accuracy at every distance. Figure 5 shows, for a sample scene, the results obtained in the far range by integrating the wide and narrow pairs of the Flea3 system (a, b), and those obtained in the short range by integrating the wide and narrow pairs of the XB3 (c, d).

4 Geometry-based scene classifier

Stereo vision endows a mobile robot with perception ability in the form of raw 3D point clouds. Online classification of stereo data as pertaining to traversable-safe regions or to obstacles would result in an enabling technology for autonomous navigation systems.

In this research, a geometry-based classifier is proposed to label data from a 3D point cloud as ground or non-ground according to their geometric properties. It adopts a self-learning scheme, whereby training instances to build the ground model are automatically produced using a rolling training set. The latter is initialized at the beginning of the robot's operation via a bootstrapping approach. No a priori assumption about the terrain surface characteristics is needed. The only hypothesis to initialize the training set is that the system starts its operation from an area free of obstacles in proximity of the vehicle, so that the sensor initially "looks" at ground only. Then, geometric features are extracted from the 3D point cloud and are associated with the ground class. When sufficient data is accumulated, the geometry-based ground classifier is trained, and the ground class is related with the point cloud properties. This allows the system to predict the presence of ground in successive scenes, using a Mahalanobis distance-based classifier. To account for variations in ground characteristics during the vehicle travel, the ground model (i.e., the training set) is continuously updated using the most recent acquisitions.

The proposed classification scheme is used for 3D points reconstructed by the XB3 and the Flea3 cameras to achieve reliable classification both in the near range and in the far range, respectively. The single steps of the approach, namely, feature extraction, ground modeling and classification, are described in more detail in the rest of this section.

4.1 Extraction of geometric features

The output of stereo processing consists of two 3D point clouds, one in the long range provided by the Flea3, and one in the short range provided by the XB3, as explained in Sect. 3.4. Both point clouds are processed to get a set of features, representative of their respective geometric properties. Specifically, each point cloud is, first, divided into a grid of terrain patches projected onto a horizontal plane. In this implementation, a regularly spaced grid of 0.4×0.4 m was found to be a good compromise between computational requirements and precision. Approaches using variable size cells, such as [29], may also be adopted to compensate possible issues due to non-uniform density of stereo reconstructed points, without altering the rest of the algorithm. Geometric features are, then, extracted as statistics obtained from the point coordinates associated with each terrain patch. The first element of the geometric feature vector is the average slope of the terrain patch, i.e., the angle θ between the least-squares fit plane and the horizontal plane. The second component is the goodness of fit, E, measured as the mean-squared deviation of the points from the least-squares plane along its normal. The third element is the variance in the height of the range data points with respect to the horizontal plane, σ_h^2. The fourth component is the mean height of the range data points, $\overline{h}$. Thus, the geometric properties of each patch are represented as a four-element vector

$$f = [\theta, E, \sigma_h^2, \overline{h}] \tag{3}$$

A typical test case acquired on field is shown in Fig. 6, as an example. It refers to the bootstrapping process during which the geometric ground model is initialized at the beginning of the operation. In general, few frames are needed to populate the ground model, e.g., six frames were found to be sufficient in our system, corresponding to a boot time of about 3 s at a frame rate of 2 fps. In Fig. 6a and c, the 3D points obtained by the stereovision processing are projected, as white dots, on the reference image of each trinocular camera through perspective transformation. The same 3D point clouds are divided into a grid of 0.16 m^2 cells, as shown in Fig. 6b for the Flea3 and in Fig. 6d for the XB3. Then, feature vectors are extracted from each cell. The normalized histograms of the distribution of the geometric features accumulated at the end of the initialization phase are shown in Fig. 7a and b for the Flea3 and the XB3, respectively. These histograms exhibit

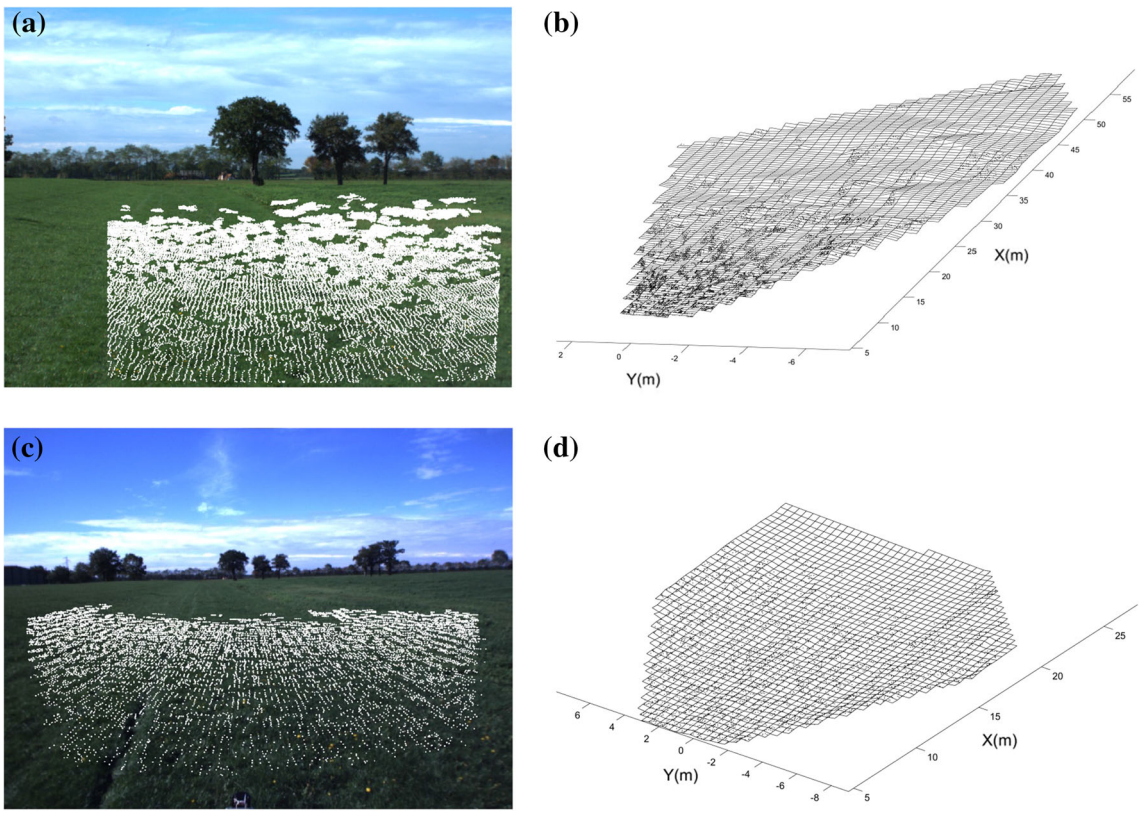

Fig. 6 Sample images acquired from a relatively flat area, during the bootstrapping process to build the initial model of the ground class. *Left* stereo reconstructed points projected, as *white dots*, on the visual image, for the Flea3 system (**a**) and for the XB3 camera (**c**). *Right* the same point clouds divided into a grid of 0.16 m^2, in the long range (**b**) and in the short range (**d**)

an approximately unimodal distribution, which suggests that the ground model for the geometry-based classifier can be reasonably modeled using a multivariate Gaussian. Note that, since the slope, the fit error and the variance appear as having positive skewed distribution, a log-transform was applied, as a pre-processing step, to better approximate a normal distribution.

4.2 Ground modeling and classification

A multivariate Gaussian distribution is fit to the geometric features extracted from the training ground labels. Then, a Mahalanobis distance classifier [30] is implemented to determine whether a new unlabeled patch is an instance of ground or not.

Let us consider N_G training ground patches, as defined in Sect. 4.1. The ground patch i is represented by its m-dimensional row feature vector f_G^i, with m being the number of feature variables (4 in our case). These vectors constitute the training set X, expressed in the form of a $N_G \times m$ matrix. If we compute the sample mean μ and the sample covariance Σ of the data in X, we can denote the ground model as $M(\mu, \Sigma)$. Then, given a new pattern f_{new}, the squared Mahalanobis distance between f_{new} and $M(\mu, \Sigma)$ is defined as:

$$d^2 = (f_{new} - \mu)\Sigma^{-1}(f_{new} - \mu)^T \tag{4}$$

Assuming that the feature vectors are independent and have Gaussian distribution, it can be proved that the squared Mahalanobis distance is distributed asymptotically as the m degrees of freedom chi-square distribution χ^2_m [31]. Then, we can use a quantile of χ^2_m, as the delimiter (cutoff) for outlying observations. Let α denote a constant probability level: $0 < \alpha < 1$. Let $\chi^2_{m;\alpha}$ denote the appropriate quantile of the distribution. Then, it holds

$$p(d^2 \geq \chi^2_{m;\alpha}) = 1 - \alpha \tag{5}$$

which means that values of d^2 greater than or equal to $\chi^2_{m;\alpha}$ appear with a probability equal to $1 - \alpha$. Now we define the cutoff for the Mahalanobis distance as

$$d^2_{crit} = \chi^2_{m;\alpha} \tag{6}$$

The pattern is an outlier, i.e., it is defined as a non-ground sample, if d^2 is greater than the critical value d^2_{crit}. This approach allows one to automatically set the classification threshold once the significance level, i.e. the admitted probability of classifying a patch as non-ground when it is actually a ground, has been fixed.

It is worth noting that, to update the ground class during the vehicle motion, the model $M(\mu, \Sigma)$ is continuously updated. New-labeled ground observations are incorporated in the model discarding, at the same time, the oldest examples. This allows the training window to provide always a fresh "photograph" of the ground, whose properties may change geographically and over time during the vehicle travel.

5 Experimental results

In this section, experimental results are presented to validate the proposed system. The multi-baseline stereo frame was mounted on the off-road vehicle shown in Fig. 2. During the experiments, the vehicle was driven by a human operator in a rural environment, with a travel speed ranging between 10 and 20 km/h, as the onboard sensors acquired data from the surroundings. The proposed classification framework was successively applied offline. Various scenarios were analyzed including positive obstacles (trees, crops, metallic poles, buildings, agricultural equipment), negative obstacles (holes, ditches), moving obstacles (vehicles, people and animals), and difficult terrain (steep slopes, highly-irregular terrain, etc).

In the rest of this section, first, results concerning the stereo reconstruction phase are presented, showing the performance of both trinocular sensors in different scenarios. Then, a quantitative evaluation of the classifier for each system is presented for a subset of salient images acquired on field.

5.1 Stereo reconstruction

To evaluate the reconstruction capabilities of the multi-baseline frame, a subset of salient test cases was analyzed in detail. No cutoff threshold on the range was used in these tests. As an example, the results of scene reconstruction obtained using XB3 and Flea3 data for two different scenarios are shown in Figs. 8 and 9. In Fig. 8, the scenario presents relatively flat ground and a building on the left in the vicinity of the vehicle, and buildings and people in the far range. Specifically, Fig. 8a and b show the reference images acquired by the Flea3 and XB3, respectively. A 3D view of the point clouds returned by each system is shown in Fig. 8c and d. It can be observed that while the Flea3 is able to reconstruct also the farthest building located at approximately 100 m from the vehicle, this building is filtered out in the XB3 reconstruction. This can be better seen in the close up of the far range reported in Fig. 8e and f for the Flea3 and the XB3, respectively. Finally, a close-up of the short range in the upper view with the two point clouds overlapped is shown in Fig. 8g, with green points representing the Flea3 point cloud and RGB points representing the XB3 point cloud. The latter view shows that the Flea3 is not able to detect nearby regions, as its point of view is located farther than the one of the XB3. In addition, it has a narrower angular field of view that causes the loss of important information on the building on the left of the vehicle, which is detected

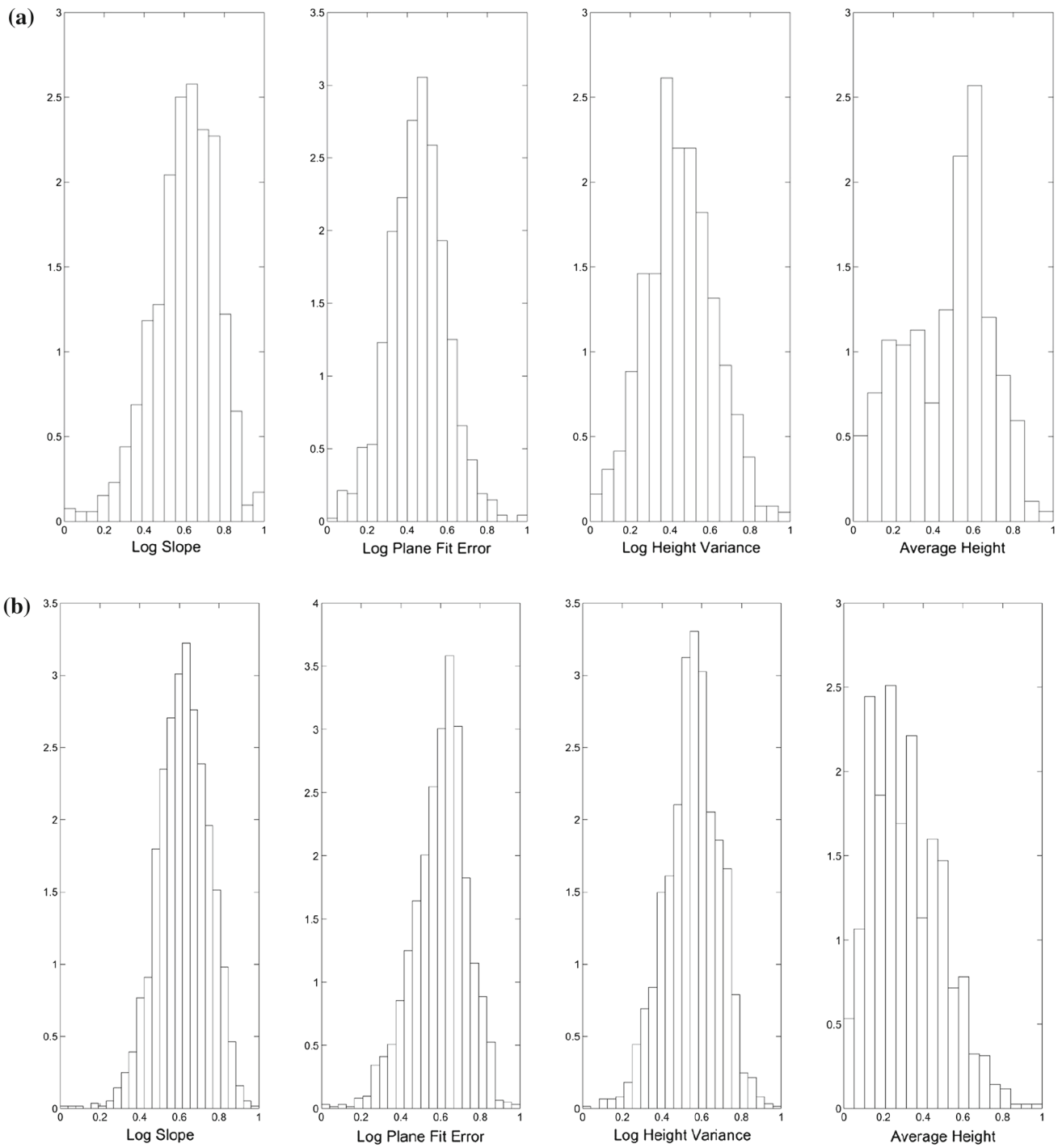

Fig. 7 Normalized histograms of the distribution of the geometric features for a training window referring to relatively even terrain, for the Flea3 system (**a**) and the XB3 system (**b**)

by the XB3 system, instead. These considerations justify the need for combining the two systems.

Similar observations can be done for the sample case reported in Fig. 9. This figure is referred to a scenario with relatively even ground in the vicinity of the vehicle, and cars, trees, and a building in the medium–far range (see Fig. 9a for the Flea3 and Fig. 9b for the XB3). The better accuracy of the Flea3 system in the medium-far range with respect to the XB3 can be seen by looking at the results obtained for the reconstruction of the building, which was located at a distance of approximately 35 m from the vehicle. Specifically, Fig. 9c and d show an upper view of the 3D reconstruction of the

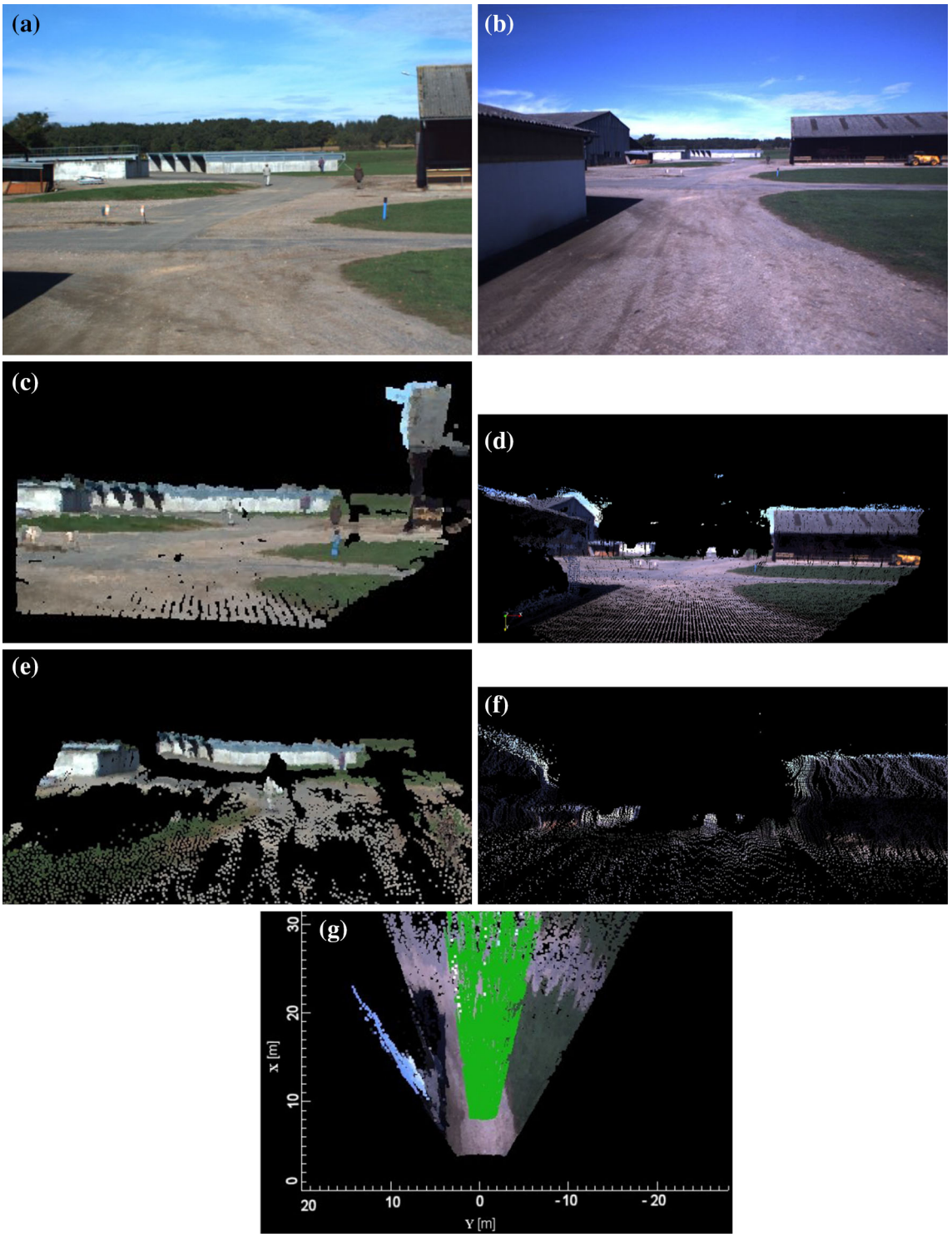

Fig. 8 Sample scenario acquired during field experiments **a** reference image of the Flea3 system, **b** reference image of the XB3 system, **c** point cloud obtained by Flea3 for the whole scene, **d** point cloud obtained by XB3 for the whole scene, **e** close-up of the long range in the Flea3 reconstruction, **f** close-up of the long range in the XB3 reconstruction, **g** upper view of the close range: RGB points are used for XB3 and green points for Flea3 data. It is shown that the Flea3 system provides accurate information in the long range while losing information in the short range compared to the XB3 camera

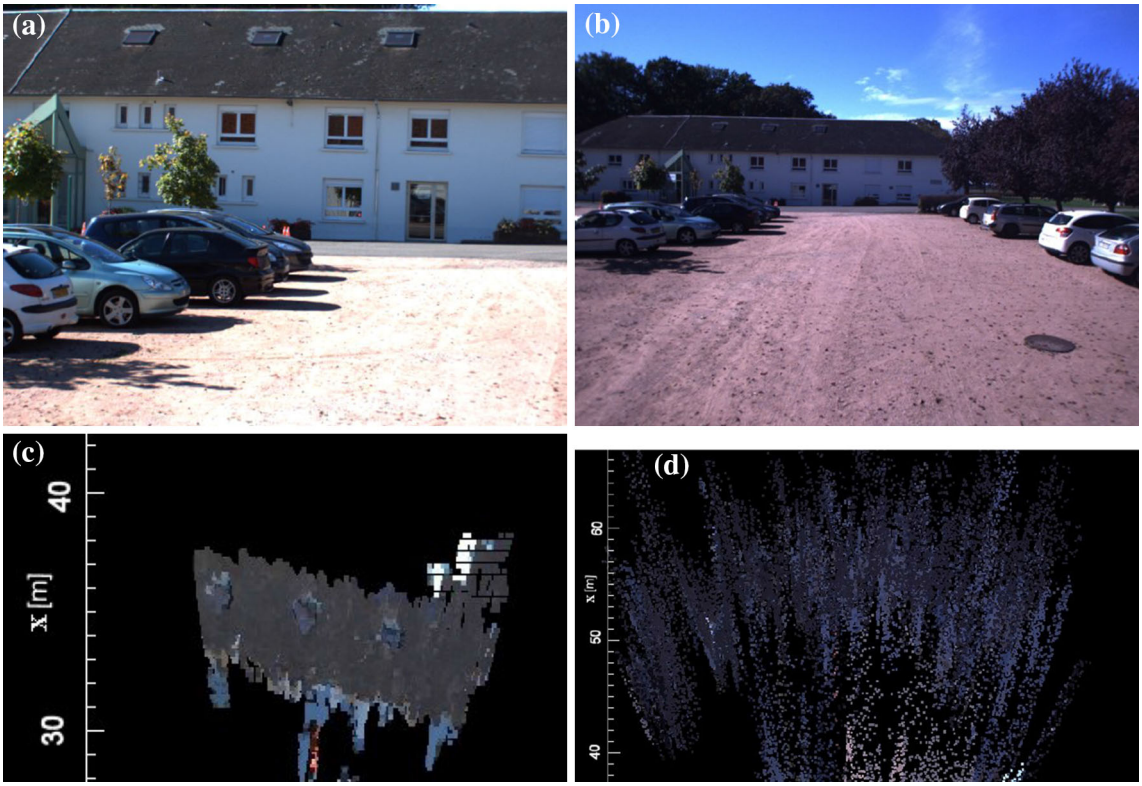

Fig. 9 Sample scenario acquired during experimentation on field **a** reference image of the Flea3 system, **b** reference image of the XB3 system, **c** upper view of the 3D reconstruction of the building obtained using Flea3 data, **d** upper view of the 3D reconstruction of the building obtained using XB3 data. It can observed that in the long range the XB3 produces noisier results with respect to the Flea3 camera

building as obtained by the two stereo systems. As expected, the results provided by the XB3 in the long range are noisier and have lower range accuracy than those produced by the Flea3 system (using the XB3, points belonging to the building are reconstructed with a wide range span between 40 and 60 m).

5.2 Ground detection

To provide a quantitative evaluation of the classification algorithm, precision, recall (i.e., true positive rate), specificity (i.e., true negative rate), accuracy, and F1-score were measured for a subset of salient images ($s_b = 135$ for each camera) taken from different data sets acquired by both trinocular systems. This subset was hand labeled to identify the ground truth corresponding to each pixel. Some sample images are shown in Fig. 10. In these figures, the results obtained from the geometry-based classifier are projected over the image plane of the reference camera of each system, through perspective transformation. Specifically, figures on the left column refer to the Flea3 camera, while figures on the right column refer to the colocated images acquired by the XB3. Points that belong to a cell labeled as ground are denoted by green dots, whereas points falling into cells marked as non-ground are denoted by red dots. Different scenarios are shown, including different types of ground (e.g., low grass and unpaved road) and obstacles (e.g., buildings, trees and bushes, people). Some images present sudden lighting variations and shadows.

The numerical results obtained on the whole subset are reported in Table 3 for both cameras. They have been obtained assuming a typical significance level of 0.1 % ($\alpha = 0.999$) for the cutoff threshold expressed by Eq. (6). It can be seen that both systems achieve good classification performance with accuracy of 86.5 and 88.8 % in the long range and in the short range, respectively.

The capability of the classifier to adapt itself to the changing geometric properties of the ground is shown in Fig. 11. It refers to a short sequence acquired by the long-range stereo module, during negotiation of a mound. Again, points that belong to a cell labeled as ground are denoted by green dots, whereas points falling into cells marked as non-ground are denoted by red dots. Initially, the vehicle travels on relatively horizontal ground, and the mound in the distance is classified as a non-drivable area (Fig. 11a), due to different geometric properties. As soon as enough examples of ground

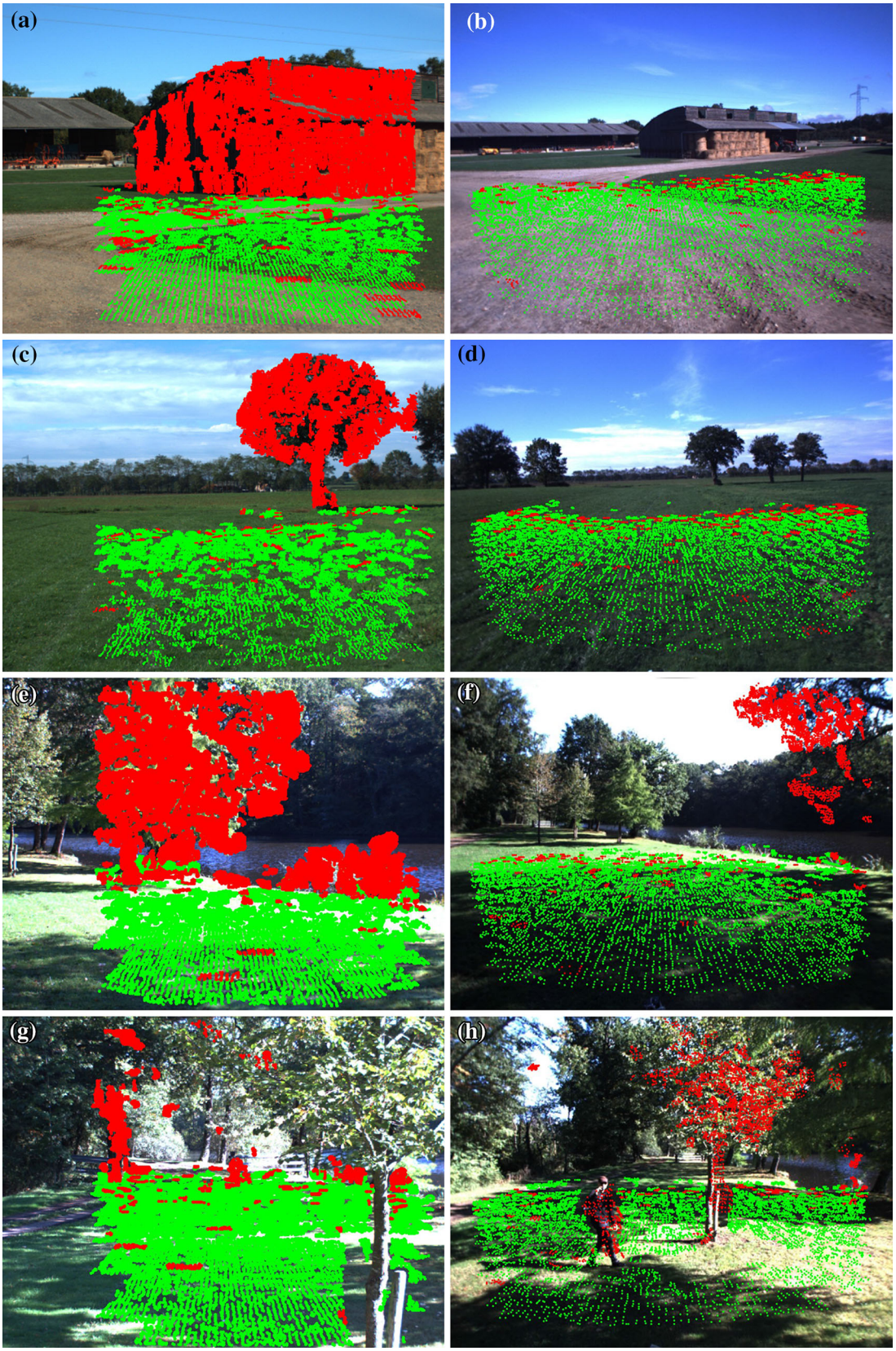

Fig. 10 Results of the geometry-based classifier for some salient images with different terrain types [i.e., unpaved road (**a**, **b**), low grass (**c**–**h**)] and obstacles [i.e., buildings (**a**), trees and bushes (**c**, **e**, **g**, **h**), people (**h**)]. *Left* long-range classification with Flea3 data. *Right* short-range classification with XB3 data. *Green dots* denote ground-labeled points, *red dots* denote points classified as non-ground (color figure online)

Table 3 Performance of the individual long-range (Flea3) and short-range (XB3) classifiers, through comparison with ground-truth data obtained by manual labeling

Camera	Precision (%)	Recall (%)	Specificity (%)	Accuracy (%)	F1-score (%)
Flea3	94.0	84.6	90.1	86.5	89.0
XB3	95.2	90.6	81.1	88.8	92.9

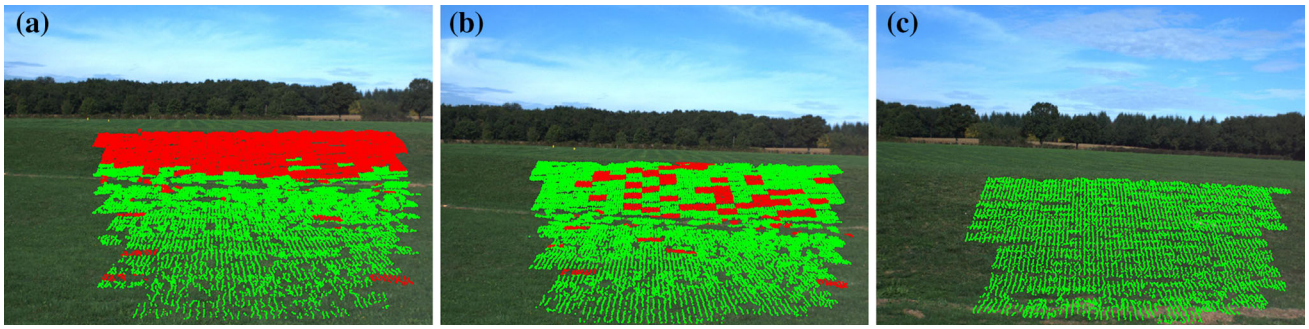

Fig. 11 Short sequence acquired by the Flea3 camera demonstrating adaptation at work. Initially (**a**), the mound in the distance is labeled as non-traversable; as soon as enough examples of ground associated with the mound are incorporated into the training window (**b**), the ground model adapts itself to the new geometric characteristics of the terrain and the mound is marked as traversable (**c**) (color figure online)

associated with the mound are incorporated into the training window (Fig. 11b), the ground model adapts itself to the changing geometry of the terrain and the mound is marked as traversable (Fig. 11c).

6 Conclusion

In this paper, a multi-baseline stereovision system for autonomous navigation in unstructured environments was introduced. The system features a short-baseline and a long-baseline trinocular camera. The former is used to reconstruct nearby points, while the latter is employed to reconstruct distant points. A self-learning framework using geometric features extracted by the 3D stereo data returned by the multi-baseline camera was also presented. It enables the vehicle to detect traversable ground based on a ground model, which is automatically built at the beginning of the robot operation and continuously updated. Experimental results obtained using a test platform in rural scenarios were presented to validate the proposed system. It is shown that using a multi-baseline stereo frame, the predictive capability of the vehicle can be extended to a wide range of visible distances, thus improving the overall flexibility and scalability of the system to different operational conditions.

References

1. Hague T, Marchant JA, Tillett ND (2000) Ground-based sensing systems for autonomous agricultural vehicles. Comput Electron Agric 25(1–2):11–28
2. Rankin A, Huertas A, Matthies L (2005) Evaluation of stereo vision obstacle detecion algorithms for off-road autonomous navigation. In: Proceedings of the 32nd AUVSI symposium on unmanned systems, June 2005
3. Rovira-Más F, Zhang Q, Reid JF (2008) Stereo vision three-dimensional terrain maps for precision agriculture. Comput Electron Agric 60(2):133–143
4. Konolige K, Agrawal M, Bolles RC, Cowan C, Fischler M, Gerkey B (2008) Outdoor mapping and navigation using stereo vision. Exp Robot Springer Tracts Adv Robot 39:179–190
5. Milella A, Reina G, Siegwart R (2006) Computer vision methods for improved mobile robot state estimation in challenging terrains. J Multimed 1(7):49–61
6. Reina G, Milella A (2012) Towards autonomous agriculture: automatic ground detection using trinocular stereovision. Sensors 12(9):12405–12423
7. Okutomi M, Kanade T (1993) A multiple-baseline stereo. IEEE Trans Pattern Anal Mach Intell 15(4):353–363
8. Gallup D, Frahm JM, Mordohai P, Pollefeys M (2008) Variable baseline/resolution stereo. In: IEEE conference on computer vision and pattern recognition, Anchorage, AK, 23–28 June 2008, pp 1–8. doi:10.1109/CVPR.2008.4587671
9. Milella A, Reina G, Foglia M (2013) A multi-baseline stereo system for scene segmentation in natural environments. In: IEEE international conference on technologies for practical robot applications (TePRA), Woburn, MA, 22–23 Apr 2013, pp 1–6. doi:10.1109/TePRA.2013.6556370
10. Broggi A, Cappalunga A, Caraffi C, Cattani S, Ghidoni S, GrisleriP P, Porta P, Posterli M, Zani P (2010) TerraMax vision at the urban challenge 2007. IEEE Trans Intell Transp Syst 11(1):194–205
11. Olson CF, Abi-Rached H (2010) Wide-baseline stereo vision for terrain mapping. Mach Vis Appl 21:713–725
12. http://www.irstea.fr/en/institute. Accessed 26 Feb 2014
13. Broggi A, Caraffi C, Fedriga RI, Grisleri P (2005) Obstacle detection with stereo vision for off-road vehicle navigation. In: IEEE computer society conference on computer vision and pattern recognition, workshop, San Diego, CA, USA, 25 June 2005, p 65. doi:10.1109/CVPR.2005.503

14. Kelly A, Stentz A (1998) Stereo vision enhancements for low-cost outdoor autonomous vehicles. In: International conference on robotics and automation, workshop WS-7, navigation of outdoor autonomous vehicles (ICRA'98), May 1998
15. Manduchi R, Castano A, Talukder A, Matthies L (2003) Obstacle detection and terrain classification for autonomous off-road navigation. Auton Robots 18:81–102
16. Moravec A (1981) Rover visual obstacle avoidance. In: Proceedings of the 7th international joint conference on artificial intelligence, Vancouver, British Columbia, pp 785–790
17. Klarquist W, Bovik A (1997) Adaptive variable baseline stereo for vergence control. In: Proceedings of the 1997 IEEE international conference on robotics and automation, vol. 3, pp 1952–1959
18. Nakabo Y, Mukai T, Hattori Y, Takeuchi Y, Ohnishi N (2005) Variable baseline stereo tracking vision system using high-speed linear slider. In: Proceedings of the 2005 IEEE international conference on robotics and automation, pp 1567–1572
19. Milella A, Reina G, Underwood J, Douillard B (2011) Combining radar and vision for self-supervised ground segmentation in outdoor environments. In: IEEE/RSJ international conference on intelligent robots and systems (IROS), pp 255–260
20. Milella A, Reina G, Underwood J, Douillard B (2014) Visual ground segmentation by radar supervision. Robot Auton Syst doi:10.1016/j.robot.2012.10.001 (in press)
21. Milella A, Reina G, Underwood J (2014) A self-learning framework for statistical ground classification using radar and monocular vision. J Field Robot (in press)
22. Stavens D, Thrun S (2006) A self-supervised terrain roughness estimator for offroad autonomous driving. In: Proceedings of the conference on uncertainty in AI (UAI), pp 13–16
23. Zhou S, Xi J, McDaniel MW, Nishihata T, Salesses P, Iagnemma K (2012) Self-supervised learning to visually detect terrain surfaces for autonomous robots operating in forested terrain. J Field Robot 29(2):277–297
24. Reina G, Milella A, Underwood J (2012) Self-learning classification of radar feautures for scene understanding. Robot Auton Syst 60(11):1377–1388
25. Vernaza P, Taskar B, Lee DD (2008) Online, self-supervised terrain classification via discriminatively trained submodular Markov random fields. In: Proceedings of IEEE international conference on robotics and automation, pp 2750–2757
26. Hadsell R, Sermanet P, Ben J, Erkan A, Scoffier M, Kavukcuoglu K, Muller U, LeCun Y (2009) Learning long-range vision for autonomous off-road driving. J Field Robot 26(2):120–144
27. Konolige K, Agrawal M, Blas MR, Bolles RC, Gerkey BP, Solá J, Sundaresan A (2009) Mapping, navigation, and learning for off-road traversal. J Field Robot 26(1):88–113
28. Bradski G, Kaehler A (2008) Learning OpenCV: computer vision with the OpenCV library. O'Reilly Media, USA
29. Kuthirummal S, Das A, Samarasekera S (2011) A graph traversal based algorithm for obstacle detection using lidar or stereo. In: IEEE/RSJ international conference on intelligent robots and systems (IROS), San Francisco, CA, 25–30 Sept 2011, pp 3874–3880. doi:10.1109/IROS.2011.6094685
30. Duda EO, Hart PE, Stork DG (2001) Pattern classification, 2nd edn. Wiley, New York
31. Mardia K, Kent J, Bibby J (1979) Multivariate analysis. Academic Press, London