Proceedings of the 2023 IEEE International Symposium
on Safety, Security, and Rescue Robotics (SSRR)
Fukushima, Japan, November 13-15 2023

Simultaneous 3D Reconstruction and Vegetation Classification Utilizing a Multispectral Stereo Camera

Johannes Vollet[1], Stefan May[1] and Andreas Nüchter[2]

***Abstract*— Obstacle detection is crucial for ensuring the safety of autonomous robots and their surroundings in unstructured outdoor environments. Objects with minimal lateral dimensions can pose risks to the robot or serve as important elements in the infrastructure it operates in. Detecting these structures becomes particularly challenging when tall vegetation is present. Distinguishing between soft, traversable objects, such as tufts of grass, and potentially lethal solid obstacles is paramount to a robot's ability to operate.**

This paper presents a novel approach that focuses on point cloud generation and vegetation identification to facilitate the safe navigation of autonomous outdoor robots. Our approach uses a single multispectral stereo camera system that employs a novel stereo matching strategy based on binary descriptors for spectrally non-identical image pairs.

I. INTRODUCTION

Autonomous unmanned ground vehicles (UGVs) face significant challenges while navigating through unstructured outdoor scenarios. Assumptions can be made about the surroundings in structured or semi-structured environments, such as industrial facilities or roadways. Obstacles in such environments typically appear as objects protruding from the ground since the terrain is usually planar. Since most of these obstacles are stationary they can be segmented, located accurately, and integrated into the path planning system for avoidance strategies. In field scenarios, the terrain is usually uneven and contains a multitude of obstacles in different shapes and forms. Simple assumptions can be made in scenarios where an appropriate overview of the scene is present. The ground clearance of the UGV at hand is the main determinant. In some cases, the environment can still be approximated as planar. The same conjectures made previously still apply, with respect to the capabilities of the robot. Although such assumptions may apply in certain cases, such as with robotic lawn mowers for gardening, the presence of tall vegetation around the robotic system necessitates a more complex approach.

LiDAR (Light Detection and Ranging) is the most commonly utilized sensor type for obstacle detection on robotic platforms. LiDAR sensors are used to reconstruct the environment in either 2D or 3D as a point cloud. Each voxel of the point cloud represents the distance and bearing from the robot. An autonomous robot operating in tall vegetation may become surrounded by objects. If obstacles are only defined by objects protruding from the ground and exceeding the robot's ground clearance, the system may get stuck with no path to plan. Another issue is the resolution of LiDAR sensors. Detecting small obstacles is challenging since they might not be recognized due to the fixed spacing of the detection (usually greater than 0.2°). Considering their lateral size, the laser scanner may not provide enough data to classify them as obstacles.

Fig. 1: Test system standing in front of a tree trunk.

The research background of this project is an autonomous robot for lawn mowing and inspection on solar farms. Cheaper solar installations are becoming more popular, compared to those commonly found in the past decade. The more affordable solution employs only screw rods to adjust low-lying solar panels' angles. Although this project focuses on this specific area, its application can be extended to other field robotic implementations of UGVs.

Thin obstacles pose no challenge for a teleoperated UGV, as the operator can distinguish between solid objects and vegetation. Unlike a teleoperated UGV, an autonomous robot cannot rely on this form of differentiation and therefore needs a system for the discrimination between possibly solid objects and vegetation.

This paper proposes a novel approach that utilizes only one sensor system for 3D reconstruction of the environment and for discriminating between obstacles and vegetation. The system employs a stereo pair of cameras, each working in a different optical band. The distinction is made via the Normalized Difference Vegetation Index (NDVI). NDVI is commonly utilized in satellite imaging and agricultural applications for chlorophyll content determination. Vegetation rich in chlorophyll reflects infrared light strongly, while absorbing most of the red. This finding led to Equation (1). The formula compares the red light spectrum of a pixel sampled by a sensor with the near-infrared (NIR) spectrum. The equation's

[1] Faculty of Electrical Engineering, Precision Engineering, Information Technology, Nuremberg Institute of Technology, 90489 Nuremberg, Germany

[2] Faculty of Informatics VII Robotics and Telematics, University of Wuerzburg, 97070 Wuerzburg, Germany

output ranges between -1.0 and 1.0, where a value closer to 1.0 indicates an image point rich in chlorophyll.

$$NDVI = \frac{NIR - RED}{NIR + RED} \tag{1}$$

The NDVI has two major benefits for the focus of our research. The equation is simple and fast to calculate, and it is applicable to all vegetation worldwide. This is due to the unique characteristics of chlorophyll, which is the same for all plants. Therefore, no training on large datasets is necessary for classification. However, a disadvantage of this approach is that it requires the initial calculation of image points for this comparison.

To the best of our knowledge, no previous research has attempted to match multispectral images for 3D reconstruction using only two cameras in the visual and NIR domains. This novel approach allows a single stereo camera pair to construct a 3D representation of the scene and simultaneously classify the objects into potentially traversable vegetation and potentially rigid non-vegetation objects.

This paper is structured as follows: Section II provides an overview of related approaches. Subsequently, our approach is outlined in section III. The matching algorithm is currently the main focus of our research. Further analysis and findings will be provided in future publications. Qualitative evaluation of the methodology on informative datasets are found in Section IV. Further discussions on the validity of this approach and its current limitations are found in the same section. The paper's conclusion is presented in Section V.

II. RELATED WORKS

Papadakis conducted a survey on methods for analyzing terrain traversability in 2013 [1]. This survey categorized the methods into proprioceptive, geometric-based, appearance-based, or hybrid approaches. The sensor system in our methodology falls into the latter category. Most approaches mentioned in this category utilize multiple sensors as input, with LiDAR being the most commonly used for ranging.

Manduchi et al. presented a strategy for discriminating between grass and obstacles using a stereo camera system [2]. They derived the range from stereo disparity and trained a Gaussian Mixture Model on shapes of various terrain types to draw conclusions about possible obstacles. In addition, they deployed a ground-looking LiDAR for surface estimation, differentiating between rough and smooth terrain. Bellutta et al. [3] previously employed a comparable method without LiDAR technology. However, they used Expectation Maximization to learn about the obstacles from color data.

As reported in [4] Suger et al. presented a strategy that solely relies on 3D LiDAR. The scans are utilized to construct a 2D grid-based map containing geometric information about the points that belong to each cell. The scans were analyzed for maximum remission, mean remission, standard deviation of intensities, roughness and slope. Using this information, a Random Forest classifier was applied to differentiate between the following classes: street, grass, dirt, or others.

In the publication of Wurm et al. [5] vegetation was classified using a tilted 2D LiDAR and the laser beam remission values. Chlorophyll reflects the infrared light of a laser strongly. Streets, on the other hand, have a much lower reflectivity. The provided data was employed to train a Support Vector Machine (SVM) for the distinction between streets and grassy areas. The SVM was trained in a self-supervised fashion utilizing the vibration data collected by an inertial measurement unit (IMU).

Bradley et al. [6] proposed a comparable method to ours for obstacle detection and classification, which was one of the first to employ NDVI for vegetation classification along with 3D perception. The multispectral camera system, consisting of two cameras and a beamsplitter aligned with the cameras' optical axes was previously reported in [7]. LiDAR provided the 3D data. This approach used Linear Maximum Entropy classifiers trained on geometric features and NDVI values to achieve a more robust classification compared to previous methods.

The most comparable method to the one we present was published by Massidda et al. in [8]. They proposed a stereo system for deployment on a micro aerial vehicle (MAV). This system was designed to find landmarks, such as trees, for aerial navigation. The trees were classified by NDVI and the range was measured with the stereo system. Three calibrated cameras were employed, two of which were filtered for the red band and used for 3D reconstruction. The NDVI was calculated with the help of the third camera in the middle, which was equipped with a NIR filter.

III. PROPOSED APPROACH

In the past years little research focused on the differentiation between vegetation and solid obstacles without machine learning in outdoor robotics, as evident from the related works. LiDAR, which is still an expensive option, is the most common sensor type for 3D reconstruction. However, LiDAR is not capable of robustly detecting laterally small objects, as mentioned in Section I. The publications of Manduchi et al. and Massidda et al. employed stereo systems capable of providing the much denser 3D representation necessary for those objects. In contrast to our approach, [2] only evaluated the 3D representation, while in [8] three cameras were deployed for a common matching algorithm and NDVI calculation respectively. This system is suitable for an airborne vehicle due to the considerable distance from the objects. The baseline between the stereo pair affects the overlap of the images and consequently, the minimum operating distance. For a UGV, a short baseline is preferred. The NDVI classification for vegetation detection was deemed beneficial by Bradley et al. in [6].

Our innovative method utilizes only two cameras for stereo matching and classification via NDVI, which reduces the overall size and cost of the system and enables the deployment on UGVs. A novel matching approach is presented for unstructured outdoor environments for images that are spectrally dissimilar.

A. System Overview

The stereo pair comprises two almost identical cameras. The color image is captured utilizing an IDS UI-3240CP-CHQ Rev.2 camera, while for NIR imaging an IDS UI-3240CP-NIR-GL Rev.2 monochrome camera is used. The RGB-camera employs a MidOpt SP675 shortpass filter with a cutoff around 675 nm for the visual spectrum. The NIR-camera is equipped with a MidOpt BP735 bandpass filter. This filter allows wavelengths from approximately 715 to 780 nanometers to pass through and results in effective separation from the red spectrum of the RGB-camera. Both cameras are fitted with a Tamron M118FM06 lens with a focal length of 6 mm, which relates to 1100 pixels in the calibration data. To ensure the calibration's validity, both cameras are securely fixed to an aluminum plate. The distance between the cameras is 48 mm.

$$z = \frac{f \times B}{D} \tag{2}$$

Equation (2) where f is the focal length in pixels, B is the base distance between the cameras in meters, and D is the disparity, allows to calculate the maximum working distance. A disparity value of 1 is selected for determining the maximum distance. A lower disparity corresponds to points close to infinity. In the system at hand, this value translates to a theoretical maximum operational distance of 52.8 m. It should be noted that the practical working distance is considerably reduced due to the high nonlinearity of the equation. With a distance of 21.4 m between disparity 1 and 2, the corresponding points are deemed highly unreliable.

Fig. 2: Multispectral stereo camera system.

The stereo camera system is displayed in Figure 2. A lens hood is added to prevent lens flare caused by the sun, which would otherwise make certain parts of the image unsuitable for the matching process. A gray card with known color values is attached to the lens hood to allow for a continuous color correction under different lighting conditions, which is needed for the accurate determination of the NDVI and the exposure time of the cameras. For better illumination, a diffusion foil is placed above the curved gray card.

The cameras are connected by a cable for the triggering system for synchronous image acquisition. One camera in free-run mode with fixed update frequency triggers the trigger-input of the other camera via the flash output provided by the digital control signal connector. The host computer is connected to both cameras via USB 3.0. The camera system is calibrated using the Robot Operating System (ROS) camera calibration toolkit, with asymmetric circle calibration targets at two different scales. This approach guarantees good calibration over a large area.

The host system consists of a Lenovo ThinkPad P52 with an 8th-gen, six-core (12-Thread) Intel i7-8850H CPU running at 2.60 GHz. The utilized operating system is Ubuntu 18.04 with ROS Melodic Morenia and OpenCV 3.4.16.

B. Image Acquisition and Processing

Correct exposure of the imaging sensors is a crucial component of the image acquisition. Most cameras are able to calculate their exposure setting on their own, and in many cases this is sufficient for a given application. However, this approach is not suitable for outdoor robotics. In our application, the foreground is the most important component and automatic camera exposure adjustments that consider the entire image are not useful. The sky is usually too bright at daytime, resulting in short exposure times that underexpose the foreground of the image when transitioning from a shaded region to a sunny one. Therefore, the first part of the software involves sky extraction and exposure adjustment, which is exemplified in Algorithm 1 as pseudocode, where I refers to images, V to vectors and A to arrays.

Data: I_{RGB}, I_{NIR}, $V_{EXPOSURE(t-1)}$, A_{CALIB}
Result: $I_{RGB,RECT}$, $I_{NIR,RECT}$, I_{SKY}, $V_{EXPOSURE(t)}$
while $inputImages == TRUE$ **do**
  $I_{RGB,DOWN} \leftarrow$ downscaleImage(I_{RGB});
  $I_{RGB,GAUSS} \leftarrow$ gaussianBlur($I_{RGB,DOWN}$);
  $I_{HSV} \leftarrow$ convertToHSV($I_{RGB,GAUSS}$);
  $I_{SKY} \leftarrow$ extractSky(I_{HSV});
  $I_{RGB,CORR}, I_{NIR,CORR} \leftarrow$ correctImageColors(I_{RGB}, I_{NIR});
  $I_{LUV,RGB}, I_{LUV,NIR} \leftarrow$ convertToLUV($I_{RGB,CORR}, I_{NIR,CORR}$);
  $V_{ERRRORS} \leftarrow$ calculateErrorsToTargetLuminosity ($I_{LUV,RGB}, I_{LUV,NIR}, I_{SKY}$);
  $V_{EXPOSURE(t)} \leftarrow$ calculateNewExposureTimes ($V_{ERRORS}, V_{EXPOSURE(t-1)}$);
  $I_{RGB,RECT}, I_{NIR,RECT} \leftarrow$ rectifyImages($I_{RGB,CORR}, I_{NIR,CORR}, A_{CALIB}$);
end

Algorithm 1: Image acquisition and processing.

The images of both cameras are acquired continuously and simultaneously at a fixed rate.

For sky extraction, the RGB image is downscaled to improve processing time and blurred with a Gaussian kernel to smooth out sensor noise. Next, the image is converted into the HSV color space (Hue, Saturation, Value), where identifying sky colors becomes relatively straightforward. As the sky is primarily bluish-gray, the corresponding regions are extracted from the image and added to a sky mask. However, this color-dependent reasoning also applies to images that are overexposed. Therefore, the lower part of the image is always excluded from the mask. The sky mask will also be utilized in the point cloud generation in the next algorithm. Points located in the sky region are prone to mismatches and should result to a disparity value of 0 and thus to a distance at infinity.

With the approximate area of the sky defined, the exposure time is calculated. First, the current color in the image is determined for both cameras using the gray card. Then, the image colors are adjusted to match the known color of the card. The next step is to convert the images into the LUV color space. To adjust for a predefined target luminosity, the image is divided into a fixed number of regions (in our case 5 x 5) and the mean luminosity of each region is determined, excluding pixels that match the sky mask. To calculate the current luminosity, a weighted sum over the means is performed, where the highest weight is located in the low center of the image. This considers the fact that the area directly in front of the robot is of the utmost importance.

The errors from both images to a predefined target luminosity and the current exposure times are fed into PID controllers. The controllers calculate the new exposure times, which are then transmitted to the cameras.

Another crucial aspect in image acquisition is the process of image rectification. In stereo camera systems, image rectification is essential for the alignment of the epipolar lines of the images and for the matching process along those lines.

C. Stereo Matching

Firstly, it has to be addressed why stereo matching using state-of-the-art methods is not possible with our sensor system. Dense stereo matching algorithms are typically implemented by comparing small patches of the images and search for the best consensus utilizing various refinement techniques. The images are usually captured by identical cameras. Therefore, gradients in contrast, for instance, are used.

The multispectral system produces significantly different images, as shown in Figure 3. These differences are necessary for the NDVI calculation and thus for vegetation classification. For this reason, we require a novel disparity estimation algorithm to extract depth information from those images.

(a) Grayscaled RGB image. (b) NIR image.

Fig. 3: Comparison of the input images.

Our approach is motivated by the presence of natural plants in images that produce many obstructions. This circumstance differs significantly from common datasets, such as the Middlebury dataset [9], which is typically associated with stereo matching publications. The images in Figure 4 exemplify that leafs in the image (Figure 4a) cause a multitude of obstructions, while common datasets have minimal possible occlusion zones (Figure 4b).

(a) Example from our dataset. (b) Teddy (Middlebury [9]).

Fig. 4: Comparison of the datasets.

In the following the matching process is outlined. The process is exemplified as pseudocode in Algorithm 2, where A stands for arrays, PC for point clouds and I refers to images.

Data: $I_{NIR,RECT}$, $I_{RGB,RECT}$, I_{SKY}, A_{CALIB}
Result: $PC_{Vegetation}$, PC_{RGB}
while $inputImages == TRUE$ **do**
  $I_{EN,NIR} \leftarrow$ enhanceImage($I_{NIR,RECT}$);
  $I_{EN,RGB} \leftarrow$ enhanceImage($I_{RGB,RECT}$);
  $I_{LAP,NIR} \leftarrow$ laplaceEdgeDetector($I_{EN,NIR}$);
  $I_{ROB,NIR} \leftarrow$ robinsonEdgeDetector($I_{EN,NIR}$);
  $I_{LAP,RGB} \leftarrow$ laplaceEdgeDetector($I_{EN,RGB}$);
  $I_{ROB,RGB} \leftarrow$ robinsonEdgeDetector($I_{EN,RGB}$);
  $A_{DESC,LOCAL} \leftarrow$ generateLocalDescriptors($I_{LAP,NIR}, I_{LAP,RGB}$);
  $A_{DESC,AREA} \leftarrow$ generateAreaDescriptors($I_{ROB,NIR}, I_{ROB,RGB}$);
  $A_{PREMAT,PIX} \leftarrow$ matchDescriptorsForPixels ($A_{DESC,LOCAL}, A_{DESC,AREA}$);
  $I_{MS,RGB} \leftarrow$ applyMeanShiftSegmentation($I_{EN,RGB}$);
  $A_{STAT,SUPP} \leftarrow$ generateStatisticsForSuperpixels ($I_{MS,RGB}, I_{RGB,RECT}$);
  $A_{MAT,SUPP} \leftarrow$ findMatchesForSuperpixels ($A_{STAT,SUPP}, A_{PREMAT,PIX}$);
  $I_{DISP1} \leftarrow$ firstStageRefinement ($A_{STAT,SUPP}, A_{MAT,SUPP}$);
  $I_{DISP2} \leftarrow$ secondStageRefinement ($A_{STAT,SUPP}, I_{DISP1}$);
  $I_{DISP,SUBPIXEL} \leftarrow$ generateSubpixelDisparityMap ($I_{DISP2}, A_{PREMAT,PIX}$);
  $I_{NDVI} \leftarrow$ generateImageNDVI ($I_{DISP,SUBPIXEL}, I_{NIR,RECT}, I_{RGB,RECT}$);
  $PC_{Vegetation} \leftarrow$ generatePointcloudNDVI ($I_{DISP,SUBPIXEL}, I_{NDVI}, I_{SKY}, A_{CALIB}$);
  $PC_{RGB} \leftarrow$ generatePointcloudRGB ($I_{DISP,SUBPIXEL}, I_{RGB,RECT}, I_{SKY}, A_{CALIB}$);
end

Algorithm 2: Stereo matching.

Since the radiometric differences in both images render direct pixel values unusable, we extract and utilize the edges of the objects in the image. Beforehand the images undergo contrast enhancement and filtering to improve edge detection results. As much of the image information is lost in the conversion into edge representations and small areas may not contain enough information for a robust and dense matching process, our approach differs significantly from contemporary approaches in two key aspects. The first aspect is the utilization of two edge detectors, one for strongly enhancing gradients in the image (Laplace) and the other for

finer granular edge detection (Robinson Compass Masks). The first image is used for a local descriptor with a strong emphasis on the center, while the second is for a descriptor with a wider spread of points for better information on the surrounding area. The two descriptors, which are the second aspect, are binary and have a length of 64 bits. Binary descriptors are calculated quickly as they only require comparisons between individual pixels. The resulting value indicates which pixel has the greater value.

The pre-matcher matches the descriptor pairs for each point between the left and right images. This is efficiently accomplished via the Hamming distance. A weighted sum is formed as the outcome of the pre-matching process based on the result provided by the local and area descriptors. This is done for a maximum disparity value of 100 in the current version, resulting in a minimum operating distance of 0.528 meters.

Because of the loss of information, it is not possible to immediately obtain exact matches for each pixel. Hence, the matching process approximates the disparity region in which each point's disparity lies. For this reason, the enhanced color image is downsampled and processed using mean shift segmentation. Following segmentation, the image is upsampled to restore its original size. This results in labeled superpixels that enable approximating the disparity regions.

Statistics are computed for each superpixel, encompassing mean and variance of the color, count of internal pixels, and adjacent superpixels. This approach is based on the likelihood that superpixels with alike colors are part of the same object and hence possess comparable distances.

During the matching process, histograms are used to determine the minimum matching result for each superpixel. Additional minima located near the minimum matching result are also stored. Areas that won't produce useful matching results are marked for exclusion in the process and are visualized by the black borders around the disparity images in Figure 5. These areas encompass the upper portion of the image where the gray card is present, the outline of the image with half the size of the descriptors, and the portion where a good overlap is not guaranteed. Disparities, count, and mean matching results are calculated and stored for each superpixel. Using this information, a weighted histogram with a small number of bins is created to approximate the most fitting value. The minimum matching difference, which may not be accurate, is determined. All disparities weighted by their count and close to this absolute minimum are included in another histogram. The bin with the highest count is extracted from this histogram, and the mean disparity value of the points inside the bin is calculated. This value represents the initial value for the superpixel being evaluated.

A two-stage refinement process is then performed. Initially, the mean hue and variance calculated in the previous statistics are utilized to identify neighboring superpixels that appear similar to each individual superpixel. A voting scheme is employed to determine the most suitable value for each superpixel such that it is consistent with its neighboring similar superpixels. This method of refinement works for many isolated superpixels. However, if multiple neighboring superpixels are initially matched incorrectly, it may lead to the clustering of erroneous superpixels. As a consequence, a second stage is implemented to detect strong discontinuities in the image. The cluster borders are detected by comparing neighboring pixels and identifying large gradients in the disparity. The detected cluster borders are then filled using a flood fill algorithm. The clusters are eliminated by taking into account similar neighbors. As a result, the disparity map is refined. By using this map of approximate disparity values, it is possible to calculate a subpixel disparity map. The pre-matcher results are used to search for the minimum value of a parabolic approximation around the found disparity for each pixel.

The matching process provides information not only about the disparity, but also about which pixel values of the images need to be compared for the NDVI calculation with Equation (1). Thresholding the NDVI values enables a classification in vegetation and non-vegetation. Finally, color point clouds are created from the disparity image and the NDVI results, and the RGB image respectively, using the Point Cloud Library (PCL) and the calibration matrix of the system.

IV. Experimental Results And Discussion

The following section presents qualitative results from an informative set of experiments. Most of the experiments were conducted during the spring and fall seasons. Figure 5 illustrates the scene in the first row, the corresponding disparity image in the second row, and the classification results as an overlaid image and point cloud in the third and fourth row, respectively. For a better visualization vegetation is marked in blue on the overlaid image, while plants and thus traversable terrain is shown in green in the point cloud. Non-vegetation is marked in red in both cases. Although it was mentioned in Section III-A that distant points are not reliable, they were not removed to represent the entire point cloud, except for parts that match the sky mask, determined in Algorithm 1. Note that the stereo pair was mounted on a tripod, rather than on the test system, in columns three and four. Although the heights were nearly identical, it is possible for the system to be vertically tilted.

The first column corresponds to the scene shown in Figure 1. The tree trunk is distinctly visible in both the point cloud and the disparity image. Furthermore, the patches devoid of grass coverage have been accurately classified. On the trunk some patches are classified as vegetation. Those relate to moss growing on the bark. Our current approach reconstructed the tree that is further away relatively well. The limitations of our methodology are also clearly evident. Incorrectly matched superpixels are present on the left side of the tree in the foreground, and the building on the right side is also not correctly matched. It should be noted that artificial objects with few features, such as buildings, are problematic for all passive stereo systems, leading to incorrect ranging, especially in the far field.

The second column displays a scene featuring tall vegetation. Matching such an image is particularly challenging for

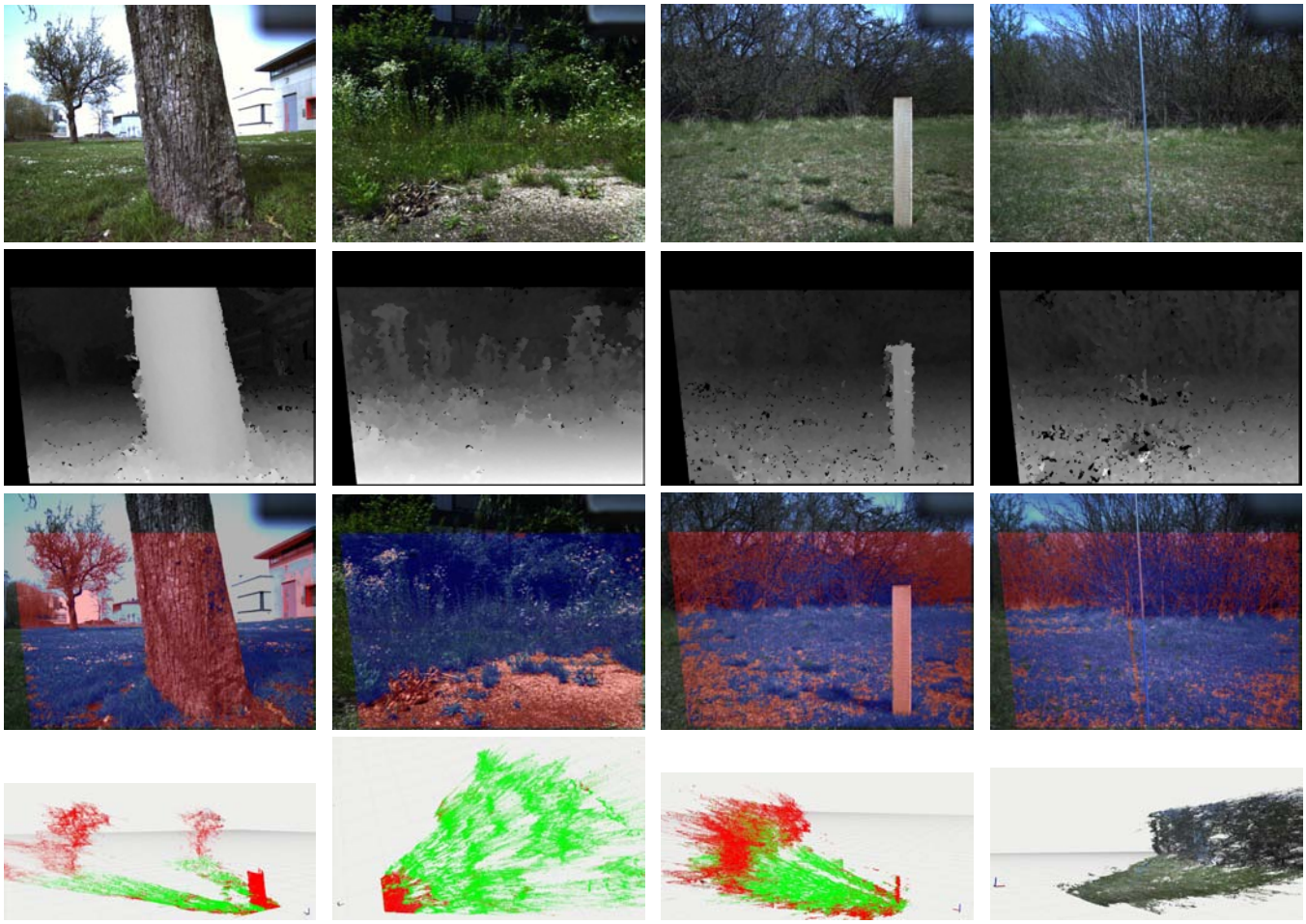
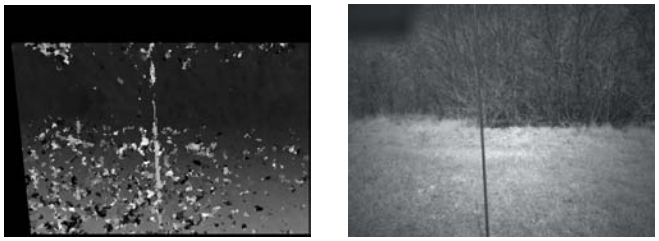


Fig. 5: Qualitative comparison of informative datasets. First row image of the scene, second row disparity image and third row overlaid NDVI result on the RGB image (blue for vegetation and red for non-vegetation). The point clouds in the forth row are either with the classification in vegetation (green) and non-vegetation (red) (columns 1-3) or in RGB (column 4).

many algorithms, because the individual plants are subject to many occlusions. Our approach provides an accurate representation of the scene in the disparity image. Several observations have to be made regarding the classification. Although the bushels of grass in the foreground are clearly distinguishable and surrounded by gravel classified as non-vegetation, non-vegetation classified voxels are present in the upper middle region of the point cloud, above the correctly classified plant. These points correspond to the blossoms, which exhibit no chlorophyll activity, as evidenced by the overlaid image. In a real-world pathfinding situation, this may create issues and must be addressed after the initial classification step by analyzing the geometric context of those voxels.

The third column of images shows a small post placed in a field in front of a hedge with little plant activity in mid-March. While it displays a fairly good matching result, it has the same problem as the tree example (first column) on the left edge of the post. In addition, it shows small errors in the center right area of the post due to overexposure of some parts. Overexposure is a common issue in outdoor scenarios with natural light. The large difference in reflected light from the ground and the post is not beneficial to accurate exposure calculation. It should be noted that these erroneous voxels are relatively few compared to the number of correct voxels of the post, and can be filtered out from the entire point cloud if desired. In the background, the hedge is well represented. However, matching highly unstructured natural objects is challenging due to occlusions and similarities between the small branches.

The final column showcases current limitations of the approach. An 8 mm screw rod is positioned in front of the camera, with the hedge in the background. The screw rod is incorrectly represented, which results in the wrong matching of the surrounding areas, as evidenced by the disparity image. The hedge in the background and superpixels not in the vicinity of the screw rod are correctly matched. As demonstrated in the third row, the mismatches also lead to misclassification, where parts of the ground are classified as non-vegetation, while the lower part of the screw rod is classified as vegetation. For this reason, the point cloud is depicted in RGB in this example.



(a) First stage disparity image. (b) NIR-image.

Fig. 6: Screw rod experiment.

This example illustrates that small objects are prone to being easily discarded, particularly during the second phase of the refinement process. Figure 6a depicts the output of the first stage of the refinement. The screw rod is clearly visible, as well as several incorrectly matched clusters of superpixels. This suggests that image smoothness, whilst important in the current algorithm for eliminating mismatched areas, is too restrictive in this case. The missing part of the screw rod above the middle in Figure 6a is explained using the NIR image depicted in Figure 6b. Due to a significant difference in the background between the RGB and NIR image, the weight of the area descriptor is too high in this particular case. No special treatment is applied to the object borders in the current process. As a result, the disparity image is oversmoothed, leading to the discarding of laterally small objects which is also visible in the tree example, where single leafs of grass in front of the trunk are not correctly classified. The oversmoothing also leads to mismatches in the occlusion zones of larger objects. This issue is present in the example of the tree and the post.

The experiments demonstrate that the approach is valid, but the current state of the algorithm is not yet satisfactory. In many instances, the superpixels are matched accurately, resulting in a sound subpixel accuracy despite the highly unstructured environment. This is particularly validated by the experiment that involved tall vegetation. The current refinement process presents the major issue.

The number of superpixels and the average processing time of the experiments are set forth in the Table I. The use of five CPU threads is due to the fact that many parts of the algorithm have to be performed on four images, with one waiting thread. The average times are taken from 100 consecutive runs. Over all the experiments, the mean processing time is 3.87 seconds utilizing one thread and 2.62 s with multiple threads.

Dataset	Superpixels	Single Thread	Five Threads
Tree	49901	3.64 s	2.44 s
Tall vegetation	59300	3.60 s	2.35 s
Post	63031	4.07 s	2.91 s
Screw rod	63643	4.15 s	2.79 s

TABLE I: Average processing times.

The matching process may seem complex, but it is efficiently computed on a, at the time of writing, five year old mobile CPU. The computation time was provided using a single thread, to enable better comparison with other approaches. The performance improvement from multithreading is lower than expected. The reason for this is that the individual process steps are currently fast to compute, but memory accesses take relatively long in comparison.

V. CONCLUSION

Although there are limitations with the current matching algorithm, particularly in the refinement stage, our approach demonstrates that multispectral stereo camera pairs are able to effectively create a dense 3D reconstruction of highly unstructured outdoor scenes and classify vegetation and non-vegetation objects at the same time. To our knowledge, this methodology is the first to use a single stereo pair of cameras with different light spectra for this purpose on an UGV.

Future research will be focused on border-aware matching, with variable weight of the local and area scores, refinement of the disparity images, as well as the optimization of the data structure for performance improvement. Quantitative evaluation of the disparity and NDVI representations will also be provided with an optimized matching algorithm.

ACKNOWLEDGMENT

The research was funded by the German Central Innovation Program for small and medium-sized enterprises, known as “Zentrales Innovationsprogramm Mittelstand”, with the funding number 16KN083034.

REFERENCES

[1] Panagiotis Papadakis, ”Terrain traversability analysis methods for unmanned ground vehicles: A survey,” Engineering Applications of Artificial Intelligence, 2013, 26 (4), pp.1373-1385, doi: 10.1016/j.engappai.2013.01.006.

[2] Manduchi, Roberto, Andres Castano, Ashit Talukder and Larry H. Matthies, “Obstacle Detection and Terrain Classification for Autonomous Off-Road Navigation,” Autonomous Robots 18 (2005): pp. 81-102, doi: 10.1023/B:AURO.0000047286.62481.1d.

[3] P. Bellutta, R. Manduchi, L. Matthies, K. Owens and A. Rankin, ”Terrain perception for DEMO III,” Proceedings of the IEEE Intelligent Vehicles Symposium 2000 (Cat. No.00TH8511), Dearborn, MI, USA, 2000, pp. 326-331, doi: 10.1109/IVS.2000.898363.

[4] B. Suger, B. Steder and W. Burgard, ”Terrain-adaptive obstacle detection,” 2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Daejeon, Korea (South), 2016, pp. 3608-3613, doi: 10.1109/IROS.2016.7759531.

[5] K. M. Wurm, R. Kümmerle, C. Stachniss and W. Burgard, ”Improving robot navigation in structured outdoor environments by identifying vegetation from laser data,” 2009 IEEE/RSJ International Conference on Intelligent Robots and Systems, St. Louis, MO, USA, 2009, pp. 1217-1222, doi: 10.1109/IROS.2009.5354530.

[6] D. M. Bradley, R. Unnikrishnan and J. Bagnell, ”Vegetation Detection for Driving in Complex Environments,” Proceedings 2007 IEEE International Conference on Robotics and Automation, Rome, Italy, 2007, pp. 503-508, doi: 10.1109/ROBOT.2007.363836.

[7] D. M. Bradley, S. Thayer, A. T. Stentz, and P. Rander, “Vegetation detection for mobile robot navigation,” Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, Tech. Rep. CMU-RI-TR-04-12, February 2004.

[8] C. Massidda, H. H. Bülthoff and P. Stegagno, ”Autonomous vegetation identification for outdoor aerial navigation,” 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Hamburg, Germany, 2015, pp. 3105-3110, doi: 10.1109/IROS.2015.7353806.

[9] D. Scharstein and R. Szeliski, ”High-accuracy stereo depth maps using structured light,” 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings., Madison, WI, USA, 2003, pp. I-I, doi: 10.1109/CVPR.2003.1211354.