Depth Estimation and Navigation Route Planning for Mobile Robots Based on Stereo Camera

Ajay Kumar Kushwaha[1(✉)], Supriya M. Khatavkar[1], Dhanashri Milind Biradar[2], and Prashant A. Chougule[1]

[1] Bharati Vidyapeeth (Deemed to be University) College of Engineering, Pune, India
akkushwaha@bvucoep.edu.in

[2] Sharad Institute of Technology College of Engineering, Yadrav, Ichalkaranji, India

Abstract. The extraction of three dimensional (3D) information from digital pictures, such as those acquired by a CCD camera, is known as computer stereo vision. Examining the relative locations of things in the two panels allows 3D information to be derived by comparing information about a scene from two viewpoints. We employed two cameras in the suggested system to identify obstructions in front of the autonomous vehicle or robot using the disparity idea. The absolute difference between the two pictures is calculated and utilized to regulate the motion of the vehicle/robot. Edge pixels retrieved from two pictures are matched, and a dense disparity map is generated by filling in the gaps between two consecutive edge pixels. A system has been proposed that permits the identification of an ideal path in real time between a starting point and a desired position in a congested environment using stereo vision, followed by a path planning algorithm and navigation enabling easy vehicle/robot traversal.

Keywords: Depth estimation · Disparity map · Stereo vision · Autonomous vehicles · Robot · Block matching · Sum of absolute difference (SAD)

1 Introduction

Recently, technology is steadily moving toward automation with little human intervention and reaching ease of operation in a wide range of disciplines. Since autonomous robots can automate a variety of labor-intensive occupations in the industrial environment and increase productivity, the number of robots used in manufacturing has rapidly increased, and this trend is predicted to continue in the future. Autonomous vehicles and robots may also do activities without the need for human intervention by combining hardware and the right algorithm to achieve the needed functionality in a range of contexts, which has led to their widespread use in several industries.

For navigation, a variety of sensors are used, such as ultrasonic sensors, laser scanners, Lidars, and others. This study looked at how to use stereo vision as the main sensor to create a consistent global map for an unstructured interior environment in real time,

N. Gupta et al. (Eds.): IC4S 2022, LNICST 472, pp. 180–191, 2023.
https://doi.org/10.1007/978-3-031-28975-0_15

which will aid in a robot's autonomous movement. Due to its numerous significances in several applications, such as Industrial Automation, ADAS, and robots, visual perception of an unknown environment has become essential [1]. Localization of the robot or vehicle, as well as sufficient map and path planning, are essential for autonomous traversal, with path planning being critical in implementing an ideal accident-free path forward. Precedent to these activities, we performed object recognition and range estimation for every item via the vision configuration.

Any sort of uncertainty perceived in a world model should be dealt with by the path planning method, which should also be capable of reducing the impact of objects on the robot when it is traversing a small space. These frequently use different map representations as their foundation, including metric maps, topological maps, hybrid maps, and others [2]. For these applications, several object recognition and classification algorithms built on stereo matching approaches [3], classic image processing techniques as well as deep learning-based methods [4] have been developed. We have used computer vision for this.

Stereoscopy is a technology for capturing and displaying 3D pictures by merging two or more photographs taken from slightly different perspectives, it may create the illusion of depth. There are two ways to capture stereoscopic images: with specialist two-lens stereo cameras or with systems that combine two single-lens cameras. The distance between the camera and the object of interest in the picture can be calculated using stereoscopic images. One of the most important qualities of any autonomous ground vehicle is its ability to discriminate between obstacles and how reliable and complete it is the perception of the environment.

After identifying an item and measuring distance from it, a system of stereo vision provides a pair of stereo images that may be used to measure distance. Avoidance performed by any of the controlling devices upon receiving the detection decision from the stereo system [5–8]. To grasp the different techniques published in the literature in recent years, a taxonomy of stereo matching methods is devised.

2 Methodology

Figure 1 shows the block diagram of the proposed system for depth estimation and navigation path planning using stereo vision, which is briefly described in this section.

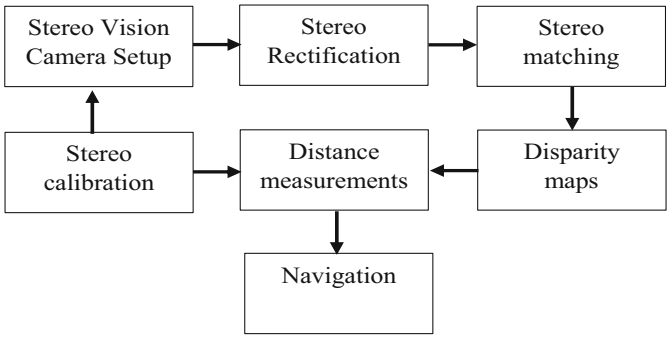

Fig. 1. Proposed diagram for depth estimates and navigation path planning

2.1 Stereo Calibration

When working with stereoscopy, it is critical to calibrate the cameras and get the necessary intrinsic and extrinsic characteristics. This is an example of a pinhole camera. A scene view is created by employing a perspective transformation to project 3D points into the picture plane.

$$s\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} r_{11} & r_{12} & r_{13} & t_1 \\ r_{21} & r_{22} & r_{23} & t_2 \\ r_{31} & r_{32} & r_{33} & t_3 \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix} \tag{1}$$

When we calibrate the camera, we get certain precise numbers that we may use to estimate distances in length units rather than pixels. The purpose of calibration is to determine the intrinsic and extrinsic characteristics of the camera. A chessboard pattern is a common method for calibrating a camera. The chessboard patterns as shown in Fig. 2. The chessboard design is easy to calibrate since it is flat, so there are no depth problems, and it is easier to extract corner points because they are well defined. The

Fig. 2. Chess board patterns

corners are all on the same line. To get a better calibration, a variety of positions are employed [9–11].

Stereo calibration is identical to single camera calibration; however, it includes more stages and provides all intrinsic and extrinsic characteristics. We must define two image vectors and locate the chessboard in each photo.

2.2 Stereo Rectification

For rectification of frames received from stereo cameras, intrinsic and extrinsic properties discovered during calibration are employed. Due to the imaging principle of the camera and the construction of the device, the left and right pictures acquired by the binocular vision system normally contain some image distortion. As a result, the identical pixel in the left and right photos may not be on the same pole line. This will make subsequent stereo matching more difficult, resulting in increased time consumption and mismatch. To increase the accuracy of stereo matching, the picture must be corrected to obtain precise co-plane and line alignment of the left and right images in theory. Calibration and rectification of stereo images as shown in Fig. 3.

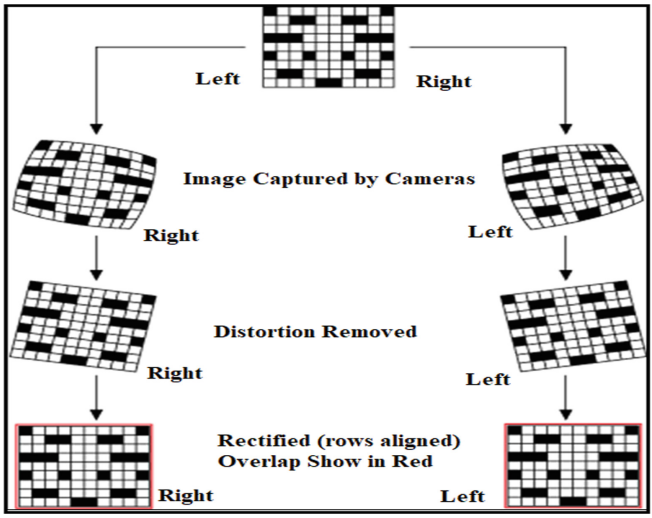

Fig. 3. Calibration and rectification of stereo images

2.3 Stereo Synchronization

Many stereo synchronization algorithms have been presented during the last two decades. All approaches are classified as sparse stereo or dense stereo matching. All approaches are classified as direct matching, custom-built filters, or network learning models. Until date, the majority prominent categorization approach has been global and local. One of the most significant responsibilities of a machine vision system is calculating the

distance between distinct points or primitives in a scenario and the cameras location. The most common method for obtaining depth information from intensity images is to use two synchronized camera signals, from which the disparity maps, also known as depth images, are created by matching the two stereo images point by point [3–9, 11, 12].

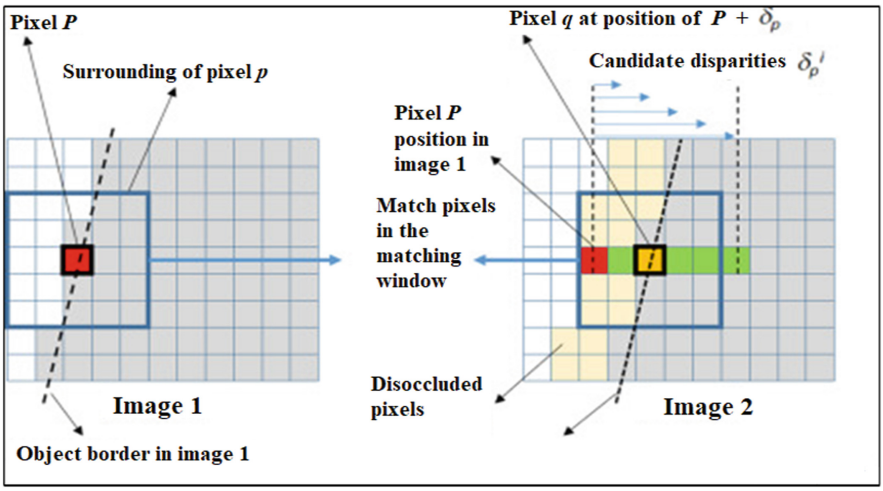

Fig. 4. Matching windows in stereo matching

To calculate the pixel p disparity, local stereo matching compares a pixel's surroundings in the left picture to significantly translated regions in the right image. The processing of each individual pixel, which ignores the context of the complete image, results in noisy disparity images. Areas of non-textured images that are impacted by input noise of any kind are especially susceptible to this (e.g., light gleaming, slightly varying colours over adjoining camera views, etc.). In Fig. 4, the surrounding areas of pixels p and q in the left and right images are compared using local stereo matching algorithms, with q translated across a candidate disparity p in comparison to p. N = 256 and 65,536 potential disparities are assessed for each pixel p in 8-bit and 16-bit depth maps, respectively, and the candidate disparity with the lowest matching cost is given to pixel p. [8]. On the other hand, accurate, local stereo matching algorithms pay close attention to the matching window form, lining up the edges with object borders.

2.4 Disparity

Disparity is the measured parallel shift in pixel-coordinates between the positions of a particular object in a pair of stereo pictures. By identifying the object in both the left and right photos, disparity is calculated. $d = xl - xr$, where xl and xr stand for the item's parallel locations in the left and right images, respectively, and d represents the disparity. Items close to the camera will be in a different position than those farther away. The relationship between disparity and depth allows us to calculate the actual

separation between two objects. Figure 5 shows two images taken using a stereo camera setup along with the distance between the objects. Each image has a different object in two dimensions. The distance between the cameras and the object can be measured by calculating the disparity. Distance and disparity are inversely connected.

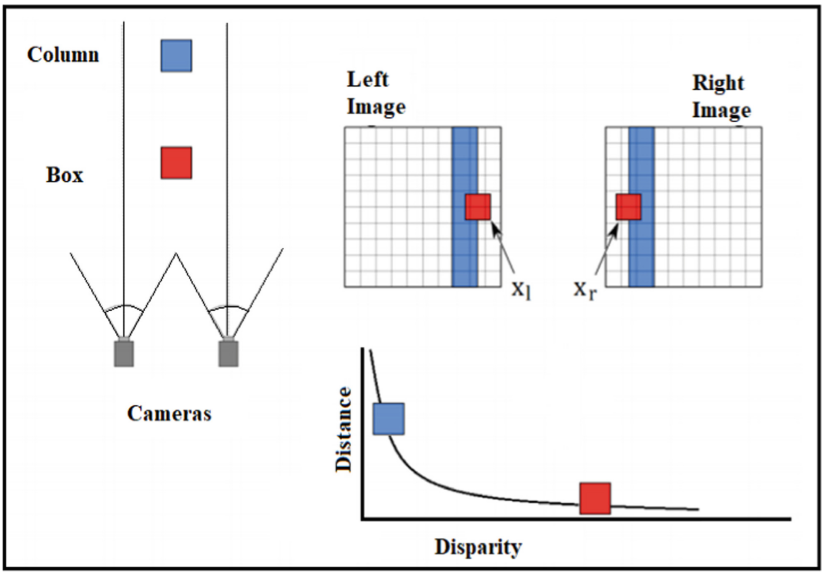

Fig. 5. The pair of photographs taken and the disparity between the objects

2.5 Disparity Map

A disparity map is a 2D matrix in which each quantity represents the pixel disparity value by a single value. By visualising each number as a coloured pixel, the disparity map can be displayed as a grayscale image. High disparity levels produce brighter pixels, whilst low disparity values produce darker pixels, as shown in the image. The disparity map is shown as a grayscale image in Fig. 6. Dark features are farther away from the cameras than bright features. Features that are too far away from the camera to be associated with are represented by the black patches.

Fig. 6. Disparity map for stereo pairs

2.6 Stereo Correspondence

The correspondence between two corrected stereo pictures is referred to as stereo correspondence. A disparity map is produced by computing the difference between the features of the left and right images. Stereo Block Matching to compute stereo correspondence, Python and OpenCV create a quick and efficient stereo block-matching technique. The semi-global block matching-method is the function that was implemented. The photos are first gone through excluding process to improve quality and make it easier to locate characteristics. In order to identify characteristics that complement those of the left and right corrected images, the approach iterates between the images using a SAD window (sum of absolute differences). The algorithm looks through the equivalent rows in the right image to find a match for each trait found in the left picture. The algorithms use the SAD window to match blocks of pixels rather than individual pixels. By doing this, each batch of photos will process more quickly, which is beneficial for real-time applications. Post filtering is used to prevent bad similarity matches from occurring. The class Stereo SGBM in OpenCV offers this capability. The disparity function in Python uses the same method. Consequently, a disparity map is produced.

2.7 Total Absolute Differences

In digital image processing, the sum of absolute differences (SAD) is a metric for picture block similarity. It is calculated as the sum of the differences between each pixel in the native block and its corresponding pixel in the contrast block. The block matching motion estimation subsystem uses the SAD method continually. The macro block uses the SAD technique to calculate the definite differences between the picture's (Template image) and its matching pixels' (Search image) pixels, and these differences are then averaged to produce the similarity block. The only arithmetic operations used in the procedure are addition and shifting. The SAD approach is particularly the quickest and is extensively used in block motion evaluation and object discovery due to its simplicity. It performs independent computation checks on each pixel in the block, making the implementation

process simpler and more parallel. This study presents the 4 × 4, 8 × 8 SAD technique for video compression motion estimation. The architecture can conduct full locomotion searches on essential manifold of 4 × 4 and 8 × 8 block dimension. Figures 7 and 8 depict the diagrams of blocks of 8 × 8 SAD and Ladder of SAD respectively.

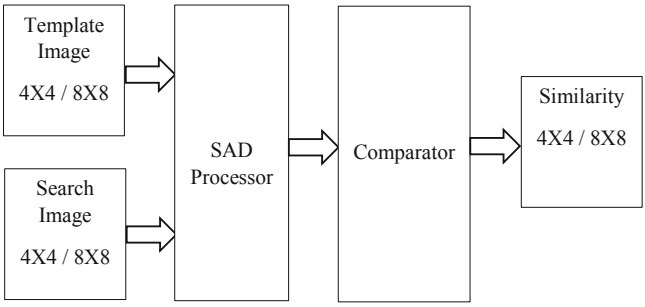

Fig. 7. Diagram of 8 × 8 SAD block

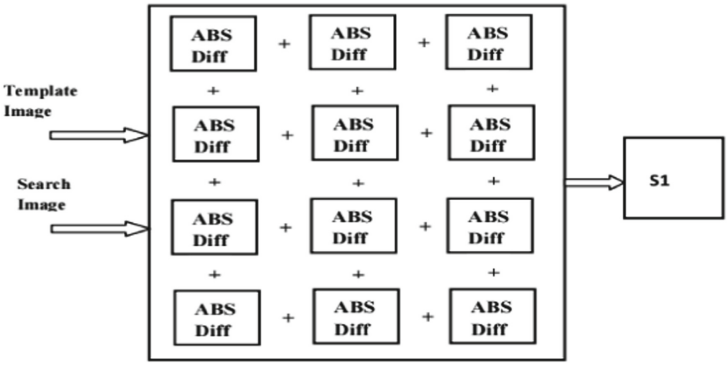

Fig. 8. Ladder of SAD block

Representative stages elaborated in completely parallel SAD architecture are:

- Carry out complete difference of all the pixels (of a block of video).
- Perform addition of all the complete dissimilarity.
- Pick block with smallest contrast value.

The sum of absolute differences (SAD) approach is an easy way to determine how similar template image T and sub-images in source picture S are to one another. The absolute difference between each pixel in T and its corresponding pixel in the sub-images being compared in S is determined. These distinctions are added together to form a basic similarity measure. Assume a 2-D m * n template, $T(x, y)$ is to be matched inside a $S(x, y)$ source image of size $p \times q$, where $(p > m \; and \; q > n)(x, y)$" [7]. The

SAD distance is determined for each pixel position (x, y) in the picture as follows:

$$SAD(x, y) = \sum_{k=0}^{(m-1)} \sum_{l=0}^{(n-1)} |S(x + k, y + l) - T(k, l)| \tag{2}$$

2.8 Distance Evolution

Range evaluation via stereovision is commonly utilized since, for a given environment, two separate perspectives are produced, which aids in determining depth values. Due to the disparity values obtained being inversely proportional to the depth of a certain scene, disparity maps are utilized to estimate the depth of a given scene. The bounding boxes produced by object detection were employed to estimate range for the detected items, with the centroid of the bounding box designating the location of the object in the environment at that precise moment [4]. The following are the distance estimate equations that were utilized based on this information:

$$\mathrm{D} = \frac{\text{baseline} \times \text{focal length}}{\text{disparity value}} \tag{3}$$

$$D = 562.44 \times d^3 + 1426.83 \times d^2 + 1300.22 \times d - 494.35 \tag{4}$$

Here, d constitute of disparity values acquired from SAD method. This equation was created using ground truth data of noticed object detection interval and disparity values, where 'd' represents observed disparity values. These two equations were obtained to calculate the range of objects that were identified.

2.9 Physical Mapping and Navigation Route Planning

Building a map and improving it at orderly time spans that are synced with the robot's movement is a critical task for navigation. Mapping an unfamiliar area aid in determining the position of the robot in relation to its surroundings. This approach aids in the estimation of landmarks, which aids in the route planning process of the robot. The suggested technology performs mapping using vision, with items classified as barriers serving as markers for the robot. "To map the environment in our scenario, we use occupancy grid maps, which are made up of discretized cells that each reflect the occupancy of a certain obstacle. These assist in locating nearby open spaces that can be used for safe movement or to accomplish a particular objective. Each grid cell is given an integer value that describes its state, much like an occupancy grid map (empty or occupied). Obstacle-occupied cells are given a high integer value, i.e., 1, whereas empty cells are given a value of 0" [4]. A typical illustration of an occupancy grid map is shown in Table 1.

Following the completion of the techniques outlined in the preceding segments, produced data is received and provide to route planning methods, which generate control points that the robot must act in accordance with in order to get to the target location. These waypoints were then utilized to provide actuation orders to the robot, allowing it to move securely.

Table 1. Occupancy grid map

1	1	1	0	0	0	0	0
1	1	1	0	0	0	0	0
1	1	1	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
1	0	0	1	1	1	1	1
1	0	0	0	1	1	1	1
1	0	0	0	1	1	1	1
1	0	0	0	1	1	1	1

3 Result and Discussion

It is depended on the methods mentioned in preceding segments, a network map was created to indicate the habitation of every network by the entity and was supplied to the planner to generate the course that the robot travelled. The range estimates for each identified barrier were calculated using Eqs. (2) and (3), after which some observations were taken, revealing that the suggested interval-disparity mapping equation, Eq. 3, proved to be more precise than Eq. 2.

Figure 9 depicts the stereovision system's output GUI. A initial camera image is shown in GUI Image 1. An image from the second camera is displayed in Image 2. The

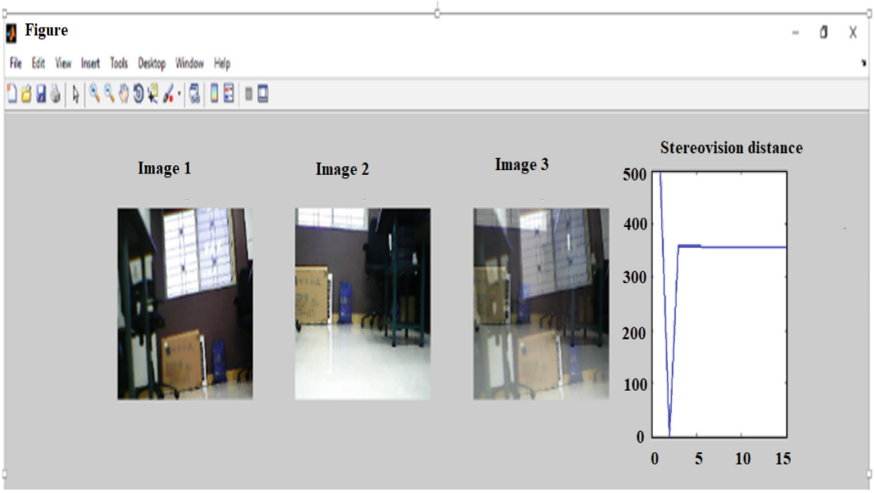

Fig. 9. Output GUI of stereovision system

composite image of images 1 and 2 is seen in image 3. The distance travelled by the object is displayed on a stereovision distance graph.

Table 2. Comparison for range estimation with equations

Ground truth (cm)	Equation 2 (cm)	Equation 3 (cm)
31	41	31
61	56	64
91	61	93
121	64	124
151	79	148
181	89	180
211	121	208
241	127	241
271	146	274

Table 2 mention comparison for range estimation with equations. The suggested method, in conjunction with the distance-disparity mapping equation, proved to be efficient in presented scenario.

4 Conclusion

The procedures mentioned here for autonomous route planning of a robot/vehicle in an interior territory are conducted in real-time using the provided object perception structure and the suggested distance estimate approach. The performance of the suggested method as well-organized approach for route planning in limited contexts was determined through analysis of path planning techniques. Due to restricted processing recourses, a high-fidelity pair of cameras was used for depth measurement, helping to improve maps while robot motion was being performed. When superior quality analytical resources are available, estimation can be performed by utilizing deep learning structures with maps improved using probabilistic route-maps, and route planning procedures can be evolved using more systematic planning methods for vigorous presentation of our robot in settings with severe constraints.

References

1. Vanne, J., Aho, E., Hamalainen, T.D., Kuusilinna, K.: A high-performance sum of absolute difference implementation for motion estimation. IEEE Trans. Circuits Syst. Video Technol. **16**, 876–883 (2006)
2. Salzman, O., Halperin, D.: Asymptotically near-optimal RRT for fast, high quality motion planning. IEEE Trans. Rob. **32**, 473–483 (2016)

3. Bansal, V., Balasubramanian, K., Natarajan, P.: Obstacle avoidance using stereo vision and depth maps for visual aid devices. SN Appl. Sci. **2**(6), 1–17 (2020). https://doi.org/10.1007/s42452-020-2815-z
4. Phan, R., Androutsos, D.: Robust semi-automatic depth map generation in unconstrained images and video sequences for 2D to stereoscopic 3D conversion. IEEE Trans. Multimed. **16**, 122–136 (2014)
5. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Trans. Pattern Anal. Mach. Intell. **39**, 1137–1149 (2017)
6. Chen, L., Fan, L., Xie, G., Huang, K., Nüchter, A.: Moving-object detection from consecutive stereo pairs using slanted plane smoothing. IEEE Trans. Intell. Transp. Syst. **18**, 3093–3102 (2017)
7. Alsaade, A.: Fast and accurate template matching algorithm based on image pyramid and sum of absolute difference similarity measure. Res. J. Inf. Technol. **4**, 204–211 (2012)
8. Stankiewicz, O., Lafruit, G., Domański, M.: Multiview video: acquisition, processing, compression, and virtual view rendering. In: Academic Press Library in Signal Processing, vol. 6, pp. 3–74 (2018)
9. Murray, D., Little, J.J.: Using real-time stereo vision for mobile robot navigation. Auton. Robot. **8**, 161–171 (2000). https://doi.org/10.1023/A:1008987612352
10. Yurtsever, E., Lambert, J., Carballo, A., Takeda, K.: A survey of autonomous driving: common practices and emerging technologies. IEEE Access **8**, 58443–58469 (2020). https://doi.org/10.1109/ACCESS.2020.2983149
11. Bányai, T., Cservenák, Á.: Logistics and mechatronics related research in mobile robot-based material handling. In: Jármai, K., Cservenák, Á. (eds.) VAE 2022, pp. 428–443. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-15211-5_36
12. Kushwaha, A.K., Kumar, A.: Sinusoidal oscillator realization using band-pass filter. J. Inst. Eng. (India): Series B **100**, 499–508 (2019)