Accurate Movement Detection of Artificially Intelligent Security Objects

Ahad Hamed Najim Al Shibli, Riham Ahmed Amur Al-Harthi, and Ramesh Palanisamy

Abstract — A number of intelligent applications demand multiple item detection and tracking as vital elements. Although object search links the identified object through a series of frames, object detection pinpoints the thing's location in a scene. Over the last few decades, a wide range of approaches has been created, which may be divided into 2D and stereo-based 3D techniques. When used in limited circumstances, most of these strategies yield trustworthy findings. These limiting presumptions are used to determine the number of complex elements that object detection and tracking naturally entail. Environmental factors, object appearance, flow density, backdrop colour intensity information, the amount of time an object is present in the scene, object occlusion, a scene's maximum number of things, etc., are among the most often held presumptions. In real-time applications, these approaches' dependability is not assured. A modern surveillance system needs reliable object identification and tracking in an open area.

Keywords — AI, Detection, Tracking, Security Objects, Location.

I. INTRODUCTION

The input video sequence is pre-processed regarding resizing, denoising, camera calibration, etc. Identifying the moving object is one of the fundamental and critical steps in object detection, usually carried out by segmenting motion in the scene.

In general, there are three fundamental methods for segmenting motion: background subtraction [3]-[6], temporal differencing [7]-[9], and strategies based on the optical flow [10], [11]. The most popular method for modelling the backdrop in relation to foreground picture attributes is background subtraction. Mean or median-based models [11]-[14], Gaussian distribution models [15]-[17], Mixture of Gaussian (MOG) models [18]-[20], Stereo vision-based disparity models [21], [22], and others are often used, backdrop models. One of the practical techniques for backdrop reduction uses stereo vision to combine 3D information with 2D disparity points.

It first creates a disparity map by matching two video sequences of the same scene from various angles. Foreground object identification uses the disparity map that was obtained together with 3D world coordinates [21], [22]. Similar to background subtraction, the temporal differencing approach sets the background model following the preceding frames. The 2D projection of the 3D velocity map on the camera plane is determined using optical flow-based motion segmentation [10]. As a picture is changed from one image to the next in a series due to this process, a velocity field is created in the image [23]. Compared to background subtraction and temporal differencing, this method offers significant benefits. For instance, discontinuity in the optical flow aids in separating foreground and background areas as well as numerous object regions. Despite the fact that this strategy is more effective than the earlier ones, it is computationally costly and noisy sensitive. It is noted that the original hypotheses and particular application of the algorithm have the greatest influence on the dependability of the aforementioned strategies.

The object detection procedure will be started following the separation of moving zones from the backdrop. Foreground blobs [24], [25], template matching [26], [27], explicit 3D form model [28]-[30], statistical shape model [27], [31], low-level features [32]-[34], multi-cue-based approach [9], [35], [36], and other methods are often used for object recognition. Techniques based on foreground blobs investigate a group of blobs acquired by background removal using analytical methods. It operates on the presumption that items are divided up into separate segments in the foreground and that object occlusion won't ever happen in the picture. Real-time applications might not support this supposition. Template-based approaches employ one or more 2D object models or templates for object detection. The central concept

Fig. 1. A generalized block diagram of a video surveillance system.
is to look for the produced template in an image's surrounding areas. This method's main issue is making the template flexible enough to accommodate varying scales and orientations. Explicit 3D shape-based procedures, in contrast to earlier techniques, identify the pose and direction using a 3D object model. Creating the model is more straightforward than earlier techniques. Moreover, it uses a hypothesis-and-verify methodology to improve the item detection process' precision. Using a single 2D statistical model, statistical shape model-based algorithms identify changes in object appearance.

After training on initial samples, the resulting model is tested on an inclusive feature space for identifying similar forms. Low-level feature-based approaches detect object and non-object areas using pattern classification algorithms and several shifting windows of varying widths across the picture at different resolutions. The main issues with this approach are the recognition of objects with variable scale, posture, occlusion, and orientation. In order to increase the detection accuracy while adding additional computational burden, multi-cue-based techniques incorporate several characteristics such as skin colour, face, and numerous body part detectors.

According to observations, the variety in objects' local and overall appearances presents the object detection module with its biggest obstacle [37]. As a result, not all types of things may be detected using an object detection approach that uses a particular characteristic or object appearance. Another challenge is object posture since different poses cause the item's overall shape to change in a number of ways [38], [39]. Moreover, an item may be seen from various angles and distances from the camera [38]. Because of this, closer objects will seem quite different from farther away [40], leading to higher occlusion and decreased accuracy. On the other hand, shadows or nearby objects may also have an impact on object templates.

At the object classification step, recognized items are categorized into pre-established classes, such as people, animals, automobiles, etc. The methods that have been developed for classifying objects may be generally divided into two categories: silhouette-based methods and appearance-based methods. Ways based on silhouettes are more effective than techniques based on appearances, and it has been noted that none of these methods exhibits robust behaviour in an unrestricted context [41]. Object and non-object verifications are completed in the warranty and refinement module once the objects have been categorized into pre-defined classes, and then thing tracking is performed.

The definition of object tracking is the construction of the material connection between identified items from frame to frame. This procedure may temporally identify the segmented sections, which also yields related data like the trajectory, speed, and direction of moving objects in the observed area. Tracking processes are typically employed to improve the effectiveness of motion segmentation, object recognition, and object classification. The tracking algorithm uses feedback control to update the object location continuously. Modern tracking techniques may be divided into two categories: continuous and single detect-and-track.

II. OBJECTIVES OF THE RESEARCH
The following vital objectives influenced the creation of a powerful stereo vision-based object recognition and tracking system.

i. Developing a stereo matching method that is illumination invariant for disparity estimation.
ii. Improving the efficiency of the object-detecting method in an obstruction
iii. Illumination-invariant stereo-matching for multiple object recognition and tracking.

III. CONTRIBUTIONS
This paper made three main scientific contributions: disparity estimation, multiple object detection, and multiple object tracking. However, there are several sub-contributions in each module, as shown below.

The gap that exists estimation module: The following disparity estimation algorithms have been created to address the main issues, such as depth discontinuities and radiometric differences in the stereo picture. A form adaptable local support region, Hybrid Correlation based stereo data cost for light invariant disparity estimation, and Correlation Fusion based stereo data cost for robust disparity estimation.

A. Multiple Object Detection Module
The following sub-modules create an effective multiple object detection method. Effective 3D Region of Interest (ROI) creation and refining methods employing RANSAC and Mode filtering for ground plane estimation and extraction. Plan-view Significance map creation for ground plane occluded object detection. A technique for multiple-item detection that uses depth layering and linked component-based labelling.

B. Multiple Objects Tracking Module
The following sub-modules are designed to create a robust multiple objects tracking approach. Creating a Kalman filter for multiple object tracking; using an object detection rollback loop and a Kalman filter for multiple object tracking.

IV. LITERATURE REVIEW
In an unrestricted setting, monitoring many objects presents a number of real-time difficulties. Two system types, monocular and stereo/multi-camera-based approaches, were primarily used to address these difficulties. Monocular camera systems have solved most challenges, but certain crucial ones, such as multiple object occlusion, object posture, changing lighting conditions, etc., are dealt with by binaural or non-linear and non-structures.

The sparse disparity information is derived from subsequent images in a stereo or multi-camera system that records the scene from many angles. To extract depth information, the resulting disparity is employed [46], [47]. This depth information may then be used for further processing, such as object identification and tracking. Comprehensive baseline systems and short baseline systems are the two main categories for multi-camera systems depending on the spacing between the cameras.

DOI: http://dx.doi.org/10.24018/ejece.2023.7.3.530
When not previously calibrated, comprehensive baseline systems offer a more flexible viewing angle than short baseline systems. Thus, most cutting-edge stereo vision-based tracking algorithms choose it without hesitation. The resultant disparity map has less noise since it estimates disparity using a sparse collection of feature points. Using 16 cameras to determine object location in 3D space, the region-based stereo method and M2 tracker [47] are created using this broad foundation architecture.

In order to track several objects simultaneously on the ground plane, Huang and Essa [20] devised a combination technique that combines two plan-view statistics, such as the Height map and Occupancy map. On the raw Occupancy and Height map, a refining approach is provided that can help eliminate some of the less essential items. To track the many objects, these enhanced maps are integrated. Recently, Zhang et al. [16] suggested a fusion-based method that fuses ground plane item appearance with plan-view statistics. The newly developed methodology is more effective than all the previous ones. However, it was unable to perform precise tracking since it lacked appearance features during the close engagement.

V. OBJECT OCCLUSION

Identifying objects under close interaction or occlusion is one of the critical problems that must be resolved for real-time applications. It has been shown by more recent techniques that it can be determined by using depth information from a stereo camera [9]. Objects are separated by the 3D data in a depth image during close contact or occlusion, and it enhances the object model by fusing scale, shape, and appearance information. The first effort was made this way by Elafi et al. [51], who used depth pictures to detect and track things directly. This technique was improved once further by the use of ground plane estimation and 3D projection [52]. The planar region is initially found using 3D points and the calibration parameters, and the direction normal to the area is determined.

This technique employs plan-view statistics known as an occupancy or density map for detection and tracking, and it combines numerous stereo views. The main issue with this method is depth noise since it estimates disparity using a traditional stereo matching algorithm included within the stereo camera. Another plan-view statistic known as the Height map was added by Huang and Essa [20] to further this effort. Compared to an occupancy map, it keeps the object shape information since it contains the highest point of each vertical bin of a planar histogram. The primary flaw in the height map is its inability to accurately detect items whose height is below the threshold while simultaneously considering the noise in the input image.

VI. RADIOMETRIC VARIATION ON DISPARITY ESTIMATION

Most algorithms developed during the past few decades operate under the presumption that adjacent pixels would have comparable colour values. So, for corresponding point estimates, these algorithms use a data cost that incorporates the intensity or colour difference of the raw picture. For the stereo picture, which lacks a matching equivalent colour value, this strategy might not work. Few attempts have been made to solve this issue in this direction. One of the crucial elements is the variation in illumination, more precisely, the radiometric difference between stereo images. Changes in camera settings, exposure variations, picture noise, non-Lambertian surfaces, etc., may bring this variation. Most stereo methods currently in use under this variation are weak at identifying the appropriate point, which results in inaccurate disparity estimates.

Most modern algorithms operate on the premise that adjacent pixels would have comparable colour values. To account for the intensity or colour difference of the original picture, these algorithms apply a data cost for the related point estimate. This method could not work well because the stereo picture doesn't have a matching suitable colour value. Not many attempts have been made to solve this issue in this manner. One of the critical elements is the variation in illumination, particularly the radiometric difference between stereo images. Non-Lambertian surfaces, exposure changes, picture noise, camera settings, etc., might cause this variation.
been found that the accuracy of the detection and tracking approaches depends on the quality of the disparity map. It has also been noted that very few attempts have been undertaken to create disparity maps of more outstanding quality. It instead employs a traditional stereo-matching algorithm followed by pricey refining methods. Estimates disparity using the traditional stereo-matching algorithm included from the Digiclops stereo vision camera [19], [10]. Using plan-view statistics, obtained discrepancies are exploited to detect occluded objects [22], [40]. Inhomogeneous and deep discontinuity zones, it is well-known that traditional stereo-matching methods do not yield correct discrepancies. Also, the disparity map will contain unknown areas because of its sensitivity to slight variations in radiometric conditions between stereo images. Modern occluded object recognition techniques are also vulnerable to depth noise since they are reliant on plan-view statistics. Thus, an effective stereo-matching technique is needed to manage radiometric changes and depth discontinuities during disparity estimation. This disparity estimate method avoids costly refining techniques, and resilience in managing object occlusion in the plan-view is automatically increased.

The findings mentioned above have inspired the development of a stereo vision-based object recognition and tracking method that can address several real-time challenges inside a single framework. Fig. 1 provides a block schematic of the developed approach. The primary contribution of this thesis is the development of effective disparity estimate methods. Before developing two robust correlation measures for correcting local and global radiometric fluctuations during disparity estimation, it first created a shape adaptive local support area-based disparity estimation approach to handle depth discontinuities. The thesis' second contribution created a method for obscured item recognition utilizing a plan-view Significance map and a 3D ROI. Thirdly, using a continuous detect-and-track architecture, a multiple-object tracking approach was constructed by fusing a Kalman filter with a detection rollback loop.

REFERENCES


