Top-down Spatial Attention for Visual Search: Novelty Detection-Tracking Using Spatial Memory with a Mobile Robot

Nevrez Imamoglu[1], Enrique Dorronzoro[2,3], Masashi Sekine[1], Kahori Kita[3], Wenwei Yu[3]

[1]Graduate School of Engineering, Chiba University, Japan, [2]Department of Electronic Technology, University of Seville, Spain, [3]Center for Frontier Medical Engineering, Chiba University, Japan;

nevrez.imamoglu@chiba-u.jp, enriquedz@dte.us.es, sekine@office.chiba-u.jp, kkita@chiba-u.jp, yuwill@faculty.chiba-u.jp

Abstract

Assistive robotics technologies have been growing impact on at-home monitoring services to support daily life. One of the main research fields is to develop an autonomous mobile robot with the tasks detection, tracking, observation and analysis of the subject of interest in the indoor environment. The main challenges in such daily monitoring application, thus in visual search, are that the robot should track the subject successfully in several severe varying conditions. Recent color and depth image based visual search methods can help to handle part of the problems, such as changing illumination, occlusion, and etc. but these methods generally use large amount of training data by checking the whole scene with high redundancy to find the region of interest. Therefore, inspired by the idea that spatial memory can reveal novelty regions for finding the attention points as in Human Visual System (HVS), we proposed a simple and novel algorithm that integrates Kinect and Lidar(Light Detection And Ranging) sensor data to detect and track novelties using the environment map of the robot as a top-down approach without the necessity of large amount of training data. Then, novelty detection and tracking is achieved based on space based saliency map representing the novelty on the scene. Experimental results demonstrated that the proposed visual attention based scene analysis can handle various conditions stated and achieve high accuracy of novelty detection and tracking.

Keywords: Novelty detection and tracking, Visual attention computational models, Space based spatial saliency, Robotics, At-home monitoring.

1. INTRODUCTION

The significance and necessity of assistive technologies in medical and service industry have been growing as a research field for daily life [1-8]. Researchers should introduce projects to improve the people’s or society’s quality of life; especially, for the people who are in demand such as elderly or people with disabilities. Recently, with the increase in population, it is obvious that the work load of service and medical industry such as the clinics and hospitals increasing rapidly that can decrease the efficiency and quality of the service to the people in demand [1-3]. As an emerging topic in the assistive intelligent systems field, in our previous work [3, 9, 10], we have been working on developing at-home bio-monitoring mobile robot system for motor function impaired persons (MIPs) with the tasks tracking, behaviour detection and analysis, and etc. (Figure 1), which have been becoming an important concept in medical field for those who have difficulties leaving their house such as elderly people or MIPs [3, 9, 10]. These systems can benefit by providing more convenient and comfortable way of health care or daily support for the patients and reducing the workload of therapists that can lead to time and economy efficiency concerning the quality of life [3, 9, 10]. In addition, they can provide seamless safety conditions for the people at-home by enabling full observation of their behaviors in any condition, in which smart camera systems or wearable sensor may have problems such as blind spots or lack of data for analysis [3, 9, 10].

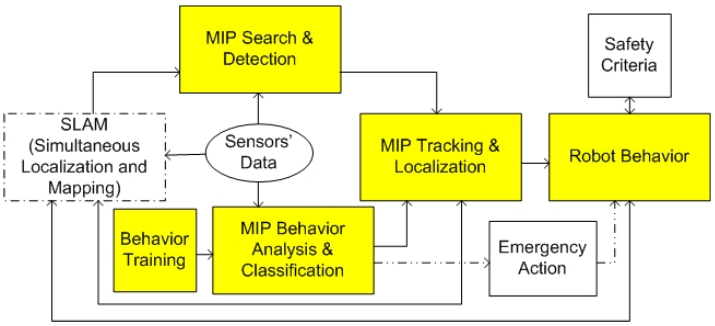
MIP Search & Detection
Safety Criteria
SLAM (Simultaneous Localization and Mapping)
Sensors' Data
MIP Tracking & Localization
Robot Behavior
Behavior Training
MIP Behavior Analysis & Classification
Emergency Action

Figure 1: Necessary Modules for At-Home Monitoring Autonomous Mobile Robot System for Mips Observation (Highlighted Modules are the Topics of Interest in our Project)

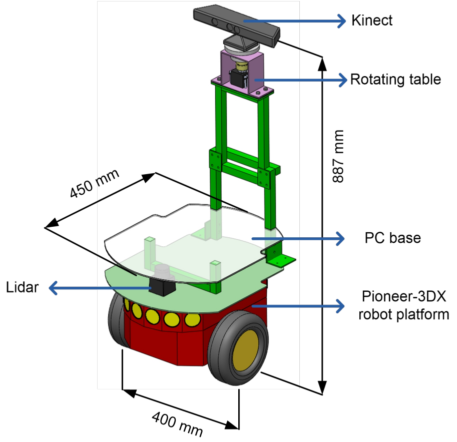

Figure 2: Bio-monitoring Mobile Robot Platform with Sensors to Provide the Tasks Defined in Figure 1

The required modules for an autonomous mobile robot system with monitoring skills are given in Figure 1 where the highlighted modules are the current research interests such as detection, tracking, subject’s activity recognition, and analysis of walking pattern using Kinect RGB-D camera sensor. For such an at-home monitoring mobile robot (Figure 2), if the visual tracking module of the system is not robust enough, the whole system will be ineffective regardless of how good or accurate activity recognition or human model the system has. It is obvious that visual search for detection and tracking of subject of interest is the main core of at-home monitoring mobile robot system: it starts or supports all the at-home monitoring tasks, including subject localization, measurement, activity recognition, and robot control.

In this work, to achieve robust detection and tracking of the novelties on the scene with low complexity algorithms and high novelty detection accuracy based on range sensors, we proposed a fast and novel algorithm by integrating 2D (laser range finder: Lidar) and 3D (Kinect) sensor information. The proposed algorithm detects and tracks the points with the highest novelty on the scene by using the robot environment memory as a top-down approach inspired by the concept of spatial memory and visual attention mechanism in human visual system during visual search. So, we achieved to detect and track subject on the scene as a novelty of the environment extracted by using the proposed space-based saliency algorithm.

The paper is organized as follows. Section 2 demonstrates the related work as background information along with the idea and contribution of this work. Section 3 describes the proposed space-based saliency model based on the spatial memory, experimental results and discussions were given in Section 4, and finally, concluding remarks were stated in Section 5.

2. BACKGROUND AND CONTRIBUTION OF THE STUDY

In order to achieve robust observation of human subjects at home, it is very critical to achieve effective visual search for the robot with subject detection and tracking tasks. The main challenges in such daily monitoring application, thus in visual search, are that the robot should detect and track the subject successfully in several severe conditions, such as: (i) low quality data with distortion or noise, (ii) partial occlusion, (iii) highly changing illumination conditions, and etc.

2.1 RELATED WORKS IN LITERATURE

2.1.1 Background on Color Image based Detection and Tracking

Regarding the detection and tracking, many algorithms have been developed using color image sensors to handle these problems [11-13]. There are several ways to utilize detection algorithms; using interest point detectors, segmentation models, background and/or foreground modeling, and using supervised classifiers with extracted features from the images [13]. For example, Gupta *et al.* [14] took advantage of SURF [15] descriptors, inspired by SIFT [16] method, for matching features from a reference subject for detection and tracking of person by a mobile robot. However, SIFT, SURF, or similar interest point based algorithms on color camera images depend on texture details, enough image resolution, and scene quality [17]. Therefore, these models are sensitive to the high illumination changes, low texture details, and low resolution images [17]. Nguyen *et al*. [18] proposed a method to use interest points to find the search points for the objects (e.g. human), then using their proposed texture descriptor non-redundant local binary patterns for matching by a Bayesian approach to find the likelihood [18]. But, it still has the disadvantages of interest point detectors, and it utilizes window scanning on the image with several scales and positions, which can be time consuming for real-time applications.

Color histogram information can also be utilized to detect and track the region of interests on images. Ning *et al*. [19] proposed the Scale and Orientation Adaptive Mean Shift Tracking (SOAMST) algorithm, which also offers invariance to rotations and scale, and it also updates search region for each frame. Moreover, it can accomplish tracking even under low resolution and low quality image conditions so it can be a good candidate for indoor real-time robot based subject tracking applications [19]. On the other hand, it also has problems with high illumination changes, and the tracking may fail if the background region has similar representation to reference frame.

In sum, there are many powerful tracking algorithms using color images as expressed work or more in literature [14-19]. However, in general, for at-home daily observation, they can be easily affected by the image data or environmental conditions such as low resolution image, low texture quality of the image, and high illumination changes, subject pose or gait differences, and etc. Therefore, there are also applications to incorporate color image data with other sensory information for detection and tracking applications.

2.1.2 Background on Multi-Sensor Approaches for Detection and Tracking

With the advancement of sensor devices in imaging technologies and new algorithms in computer vision applications, different sensory type representations of the scene can be provided such as depth or disparity images, 3D point clouds, thermal images, and etc [12, 20-23]. For instance, Talha and Stolkin [20] proposed an adaptive system that utilize particle filter based subject tracking by fusing thermal and visible spectra camera. Using thermal cameras along with color images, it is possible detect and track human subjects accurately; however, thermal camera systems generally has low resolution field of vision, and high-quality and high-resolution thermal imaging solutions are too expensive to be considered in at-home bio-monitoring robot projects.

On the other hand, range sensor applications for range imaging and 3D point cloud representations, are more appropriate choice considering the cost and performance. Therefore, applications have been developed for detection and tracking by using color and depth data integration [21-23]. For example, Garcia *et al*. [21] proposed particle filter model by representing the particle states with 3D world coordinates and particle features with depth gradients and polar representation of color averages from HSV color space Cartesian definition [21]. It is also possible to detect and track specific objects by using prior knowledge from disparity and depth images of the region of interest. In the study of [22], detection and tracking algorithm is proposed using stereo vision mounted on a mobile robot by using the head and shoulder shape information such height and width ratio from disparity [22]. In addition, Liu et al. [23] make use of 3D point cloud data obtained from Kinect to find the possible candidate points by selecting points with local height maxima compared to their surroundings [23]. Then, for learning the human appearance model with SVM classifier, color-height joint histogram features are extracted from various height and human head appearance (front, back, and etc.) [23]. Then, appearance model is used to track the human detected on the scene [23].

In sum, depth or 3D data fusion with color based features can greatly improve the detection and tracking performance of the system considering the illumination changes and object representation. However, the models that combines color and depth data can fail under high illumination conditions due to the change in color appearance model. And, most of the models requires large amount of training data depending on the detection task with supervised approaches. In addition, due to searching and checking all the interest points on the image, these models generally have redundancy. Especially for detection cases, all the images goes under high-level feature extraction, feature matching, or feature classification procedure, which is redundant and time consuming task on the visual search. Even though there are models to eliminate the redundant information as in [23], they lack of using the knowledge on the environment, which can greatly deal with the redundancy especially for the indoor applications regarding the at-home monitoring for daily support.

2.1.3 Background on Visual Attention (Saliency Map) based Detection and Tracking

Visual Attention (VA) mechanism is an important part of Human Visual System (HVS) by avoiding redundant data on the scene and popping out the significant information on the scene for the benefit of selective attention process [24, 25]. Visual attention mechanism works from two perspectives such as spatial attention (space based attention) and object base attention, where spatial attention is spatial features such as contrast and orientation but object based attention relies on the object structure such as shape [26-28]. By using these relations, VA mechanism can benefit scene analysis and visual search by using low-level features for task independent situations or prior information with decision making process if there is task or knowledge dependency during the visual search [24, 25]. Many computational models have been developed to mimic VA for the benefit of computer vision applications; however, most of the saliency models are based on the 2D color images [24, 25]. And, there are few applications to create saliency maps for specific detection and tracking approaches [29-31]. Zhang *et al*. [30] proposed a context aware saliency model for the key object discovery and tracking application by introducing spectral affinity to saliency computation and saliency based particle filter tracking [30]. Yang et al. [31] proposed a more task specific approach to create saliency maps for specific objects such as bicycle, car and person.

On the other hand, several approaches combined the 3D data integration to color based models for saliency computation [32-37]. As stated by Wang et al. [32], most of the 3D based models utilizes depth image from three perspectives [32]: i) depth weighting [33], ii) depth saliency [34], and iii) stereo vision models that compute and use disparity [35]. All the studies stated until now, they were generally space based approaches by comparing one position to another based on 2D color images or depth images for 3D models. There are few studies to utilize space and object based attention models or 3D real world space by using all three *XYZ* dimensions to find attentional objects [36, 37]. For example, Garcia and Frintrop [37] proposed a model for attentional 3D Object Detection by using Kinect RGB-D scene, where they do combine clustered 3D data and color image saliency to create a 3D object saliency map [37]. However, these approaches do not tell the exact novelty of the scene since they do not use spatial memory of the scene. So, attention regions or attentional objects found from these algorithms can both includes novelty or previously existing irrelevant part of the scene. And, the result still redundant to give valuable information such as what is new in the scene or what is most likely to be to object or subject of interest for a monitoring system ,which is crucial for surveillance of people living alone and in demand. Therefore, there is a need for a better 3D approach for novelty detection to observe subject robustly in indoor environment.

2.2 IDEA AND CONTRIBUTION OF THE WORK

For the detection and tracking tasks, low level and high level features on 2D/3D space or objects, such as color contrast, orientation, shape, object appearance model, and etc., are certainly necessary for an attention mechanism to detect novelties on the scene either for bottom-up or top-down mechanisms. However, current models neglect knowledge or effect of the space based or spatial attention [26-28], regardless of object features during visual search, which can decrease redundancy on attention significantly. As stated and referenced by Chun and Wolfe [38], experience and memory also affect attention while the attention has its contribution to experience and memory [38, 39]. In addition, Oh and Kim [40] also demonstrated the influence of spatial memory on visual search. Most importantly, Chun and Wolfe [38] also states that, people tend to pay attention to the regions, items or objects with novelty on the scene [38, 41]. These physiological studies [26-28, 38-41] prove that novelty based on spatial information, which can decrease redundancy and guide human attention during visual search, can also be an important factor for attention process independent of object features. Therefore, inspired by the idea of these spatial memory and space-based attention on HVS, we proposed a computation model to employ spatial memory as the knowledge of occupied region in the environment. We tried to detect novelties, specifically person of interest for the at-home monitoring mobile robot, which does not belong to the part of occupied regions of environment. We implemented a detection and tracking algorithm based on finding the novelty on the scene with the idea of space-based attention.

On the other hand, similar to the saliency approaches that generally combine spatial information in 3D and 2D color information, the most similar studies other than saliency map are related to the semantic mapping or semantic representation [8, 42-44] of the scene which may also associate visual search and visual memory; however, most of the applications are based on the analysis of data for segmentation [42] or creating 3D semantic maps (semantic SLAM) [8, 43, 44] rather than novelty detection. For example, in the study [44], the system integrates different sensory information such that robot uses 2D global mapping of the environment for localization, and with Kinect, 3D local information is transformed into global frame to find out objects and their changes in the position if any during long-term observation in indoor [44]. So, these models do not detect novel regions temporarily exist in the scene such as persons.

In addition, regarding the novelty detection and tracking for monitoring, semantic scene analysis approaches do not consider the attention or saliency concept to define visual priority of the regions, or they do not search for novelty on the scenes, with high likelihood to be the subject of interest in a known environment, to decrease redundancy in visual search. Therefore, by integrating different sensory information as in [44], we proposed a novel space-based saliency approach to detect and track novelties for at-home monitoring task to support people in demand such as elderly or MIPs who may be living alone. To the best of our knowledge, there have not been any models developed for novelty detection based on spatial memory of the real world scene, since most of the attention based approaches are either 2D model or depth data processing of the visual field for depth saliency computation. In addition, the algorithm can be integrated into many existing color or depth based state-of-the-art models for detection and tracking applications to decrease the search points on the scene. Because the algorithm can decrease the search points or regions by providing the novel regions as task-dependent high priority attention regions on the scene with the aid of spatial memory based attention model. In this work, we have tested the model with different walking patterns, illumination and environmental conditions on various recorded data. Experimental results demonstrated that proposed visual attention model can handle various conditions stated by having high accuracy of novelty detection and tracking.

3. PROPOSED NOVELTY DETECTION AND TRACKING

The idea is to generate a space based spatial saliency map, in which the detected novel point is tracked along the continuous frames. The flowchart of the proposed model is given in Figure 3. In the proposed model, space-based saliency map is computed by comparing the readings from Kinect sensor (Figure 2) mounted on the rotating Table with the prior information of occupied region on the environment memory from Lidar sensor (Figure 2) along with the depth and size heuristics of the candidate regions.

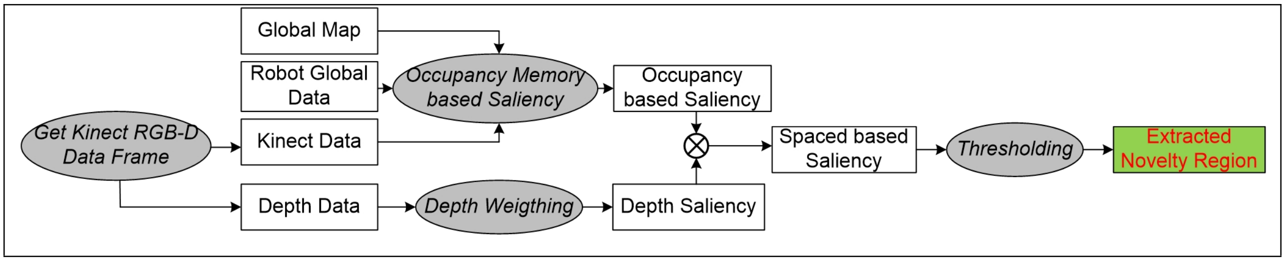

Figure 3: Flowchart of the proposed novelty detection and tracking algorithm

To be able to transform the local Kinect data to global environment map, first, the robot initially should build the map of room or house as a spatial memory of the environment. For robot handling and mapping tasks, we take advantage of Robot Operating System (ROS) [45], in which ROS incorporates with the ROSARIA package [46] to read the robot odometry data. And, prior to the main task of novelty detection and tracking, robot generates a 2D map of the environment with Gmapping package in ROS as described in [47, 48] by combining the odometry and Lidar (Figure 2) sensory information with 5cm grid resolution. Then, built map can be used for localization and robot path planning ,as demonstrated in Figure 4(a), with ROS navigation stack that includes localization, costmap, global and local path planer. Robot pose is estimated by using the Adaptive Monte Carlo localization algorithm proposed in [49] from the global map generated with Gmapping [47, 48]. And, 2D occupancy grid map as the cost map [50] is employed to define the cost of grids on the map, which consists of obstacle information, static map, and dynamic sensory information for update process.

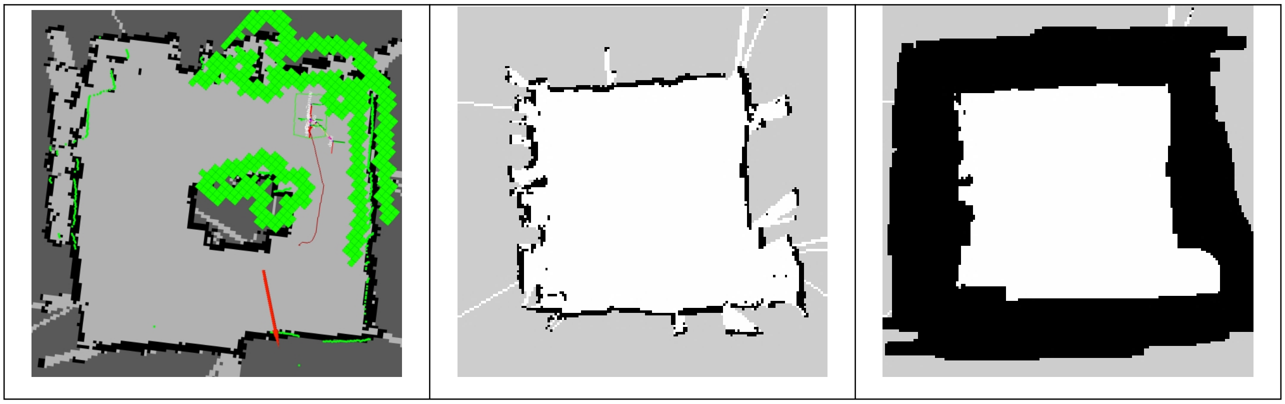

Figure 4: (a) ROS localization and path planning on the global map built by robot exploration, (b) global 2D Occupancy map (c) updated occupied memory map for the regions where novelties can not exist (black pixels are occupied)

After having the global 2D map as the environment memory for the robot, we can use the global occupancy map and robot pose to find out the space based saliency. Initially, irrelevant places should be removed or added as occupied region to memory such as table plane or regions that subject can not appear. Therefore, the locations such as tables or irrelevant regions with their surfaces on the global map (Figure 4(b)) were added to occupied regions of the memory (Figure 4(c)) where the subjects can not be observed. Then, next step is to process Kinect 3D data for occupancy memory based saliency detection (Figure 3). First, we obtain the Kinect 3D data, and remove the floor and ceiling from the observation to reduce the irrelevant points for the saliency calculation by using height threshold values on 3D points (Figure 5). Also, since the Kinect data is 3D and the global map is 2D without any height values, Kinect data is reduced to 2D dimension by avoiding the height information. And, the occupancy memory based saliency (see Figure 3) of the points on filtered Kinect 3D data can be defined as:

$$\mathbf{S}_o\left(\mathrm{X}_k^l \middle| \mathbf{O}_m^g, p_r^g, \theta_r^g, \alpha_k^g, \mathrm{X}_k^l\right) \tag{1}$$

where $\mathbf{S}_o$ is the occupancy based saliency values of the observed points on Kinect local data given the local and global information such as $\mathbf{O}_m^g, p_r^g, \theta_r^g, \alpha_k^g$, and X_k^l . In Equation (1), X_k^l is the local kinect data with horizontal and depth distances as the Kinect center is the origin, and depth (*x*) values of the locally detected occupied regions on the scene, $\mathbf{O}_m^g$ is the global occupancy map, p_r^g is the robot global position on 2D axis, θ_r^g is the global pose of the robot on $\mathbf{O}_m^g$, α_k^g is the Kinect pose differs from robot pose due to the rotating table as shown in Figure 2.

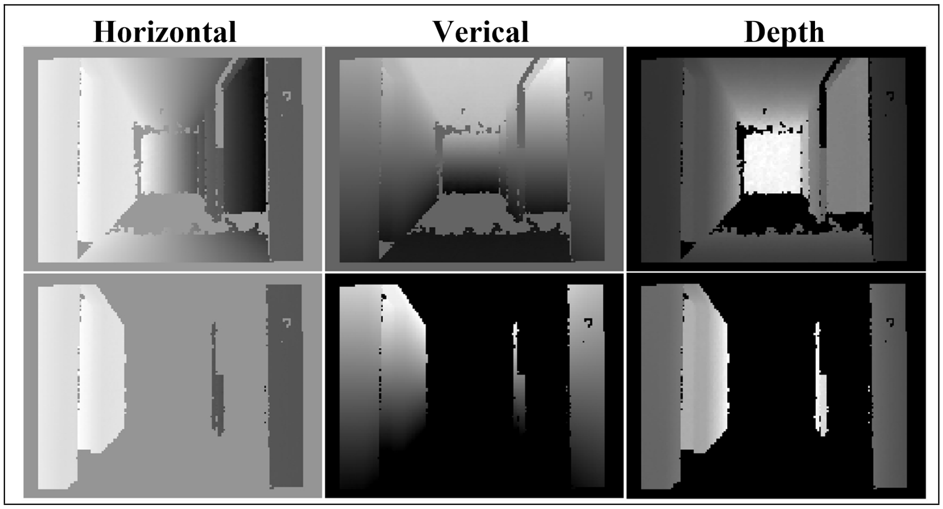

Figure 5: (top row) Kinect 3D data (bottom row) Kinect 3D observation after removal of floor and ceiling

To be able to calculate the occupancy based saliency in Equation (1), Kinect local data should be transformed into global values (Figure 6) regarding the global 2D map to be able to find out novelty points as in Equation (2) [51]:

$$\begin{bmatrix} K_{Gx} \\ K_{Gy} \end{bmatrix} = \mathbf{A} \times \begin{bmatrix} K_{Lx} \\ K_{Ly} \end{bmatrix} + \begin{bmatrix} T_{Gx} \\ T_{Gy} \end{bmatrix} \tag{2}$$

$$\mathbf{A} = \begin{bmatrix} s_x \cos(\alpha) & -s_y \sin(\alpha) \\ s_x \sin(\alpha) & s_y \cos(\alpha) \end{bmatrix} \tag{3}$$

$$\begin{aligned} T_{Gx} &= R_{Gx} - d\cos(\theta) \\ T_{Gy} &= R_{Gy} - d\sin(\theta) \end{aligned} \tag{4}$$

where K_{Gx} and K_{Gy} are the transformed Kinect global *x* and *y* positions on global 2D map, K_{Lx} (depth data in Fig.6) and K_{Ly} (horizontal data in Figure 5) are the local Kinect data on the relevant *x* and *y* axis. K_{Lx} and K_{Ly} is resolution is adjusted to 5cm resolution to match the 2D map resolution since each pixel representation 2D map image corresponds to 5cm grid regions. α is the Kinect pose on global map calculated by the robot pose and rotating table angle. **A** is the transformation matrix given in Equation (3) [51], in which s_x and s_y are the scaling coefficients on x and y axis. T_{Gx} and T_{Gy} are the translation values due to the difference between robot center and Kinect position on the robot (Figure 2). Translation values are defined as in Equation (4) where R_{Gx} and R_{Gy} are the robot position in global 2D map, and *d* is the distance of Kinect position to the robot center, and θ is the robot pose on global map.

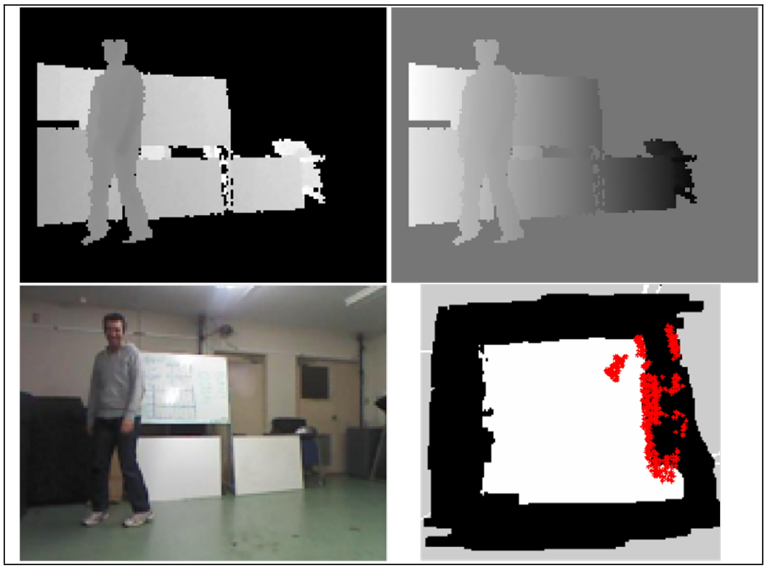

Figure 6: (top row) Kinect 3D local data depth and horizontal images respectively (bottom row) Color image of Kinect 3D observation, and $K_G(x,y)$ transformed points are shown on 2D map with red marks

For each transformed Kinect global points, the occupancy based saliency can be defined by find the distances of the global Kinect points to closest occupied regions on the global map as in Equation (5) with the given global transformed Kinect points ($K_G(x,y)$) and the environment memory ($\mathbf{O}_m^g$).

$$\mathbf{S}_o\left(K_{Gx}, K_{Gy}\right) = \arg\min\left(K_G\left(x,y\right), \mathbf{O}_m^g\right) \tag{5}$$

$$\begin{aligned} & \textit{foreach } K_G\left(x,y\right) \\ & \qquad \mathbf{S}_o\left(K_{Gx}, K_{Gy}\right) = \min\left(\operatorname{norm}_{l^2}\left(\mathbf{O}_m^g, K_G\left(x,y\right)\right)\right) \\ & \textit{end} \end{aligned} \tag{6}$$

where $S_o(K_{Gx},K_{Gy})$ is the occupancy based saliency of the given $K_G(\underline{x},y)$ point with height h based on the environment knowledge $\mathbf{O}_m^g$ as in Equation (5). The definition of *argmin* function to calculate the saliency can be expressed in Equation (6), where the comparison of the Kinect points is done by l^2-*norm* of the vectors defined with the each Kinect point to the each map point; in other way, Euclidean distance of the Kinect points to the each occupied region in spatial memory.

Then, next step is to define the depth saliency to give more priority to the closer novelty regions by increasing the occupancy based saliency value of each $K_G(\underline{x},y)$. This operation is done by defining depth weights as in Equation (7).

$$\mathbf{S}_D = \frac{1}{\mathrm{N}\left(\mathbf{X}_k^{depth}\right)+1} \tag{7}$$

$$S_S\left(K_{Gx},K_{Gy}\right) = S_D\left(K_{Gx},K_{Gy}\right) \times S_o\left(K_{Gx},K_{Gy}\right) \tag{8}$$

$$S_S\left(K_{Gx},K_{Gy}\right) = \begin{cases} S_S\left(K_{Gx},K_{Gy}\right) & if\ S_S\left(K_{Gx},K_{Gy}\right) > T \\ 0 & otherwise \end{cases} \tag{9}$$

where **S**$_D$ is the depth saliency map, which is used for weighting the occupancy based saliency map S_o, and N(.) is the normalization function for local depth data $\mathbf{X}_k^{depth}$ of Kinect sensor. For the calculated **S**$_D$ and S_o, space-based saliency map, **S**$_S$, can be calculated by Equation (8), where the saliency values less than a given threshold as in Equation (9) to remove irrelevant regions on extracted salient regions.

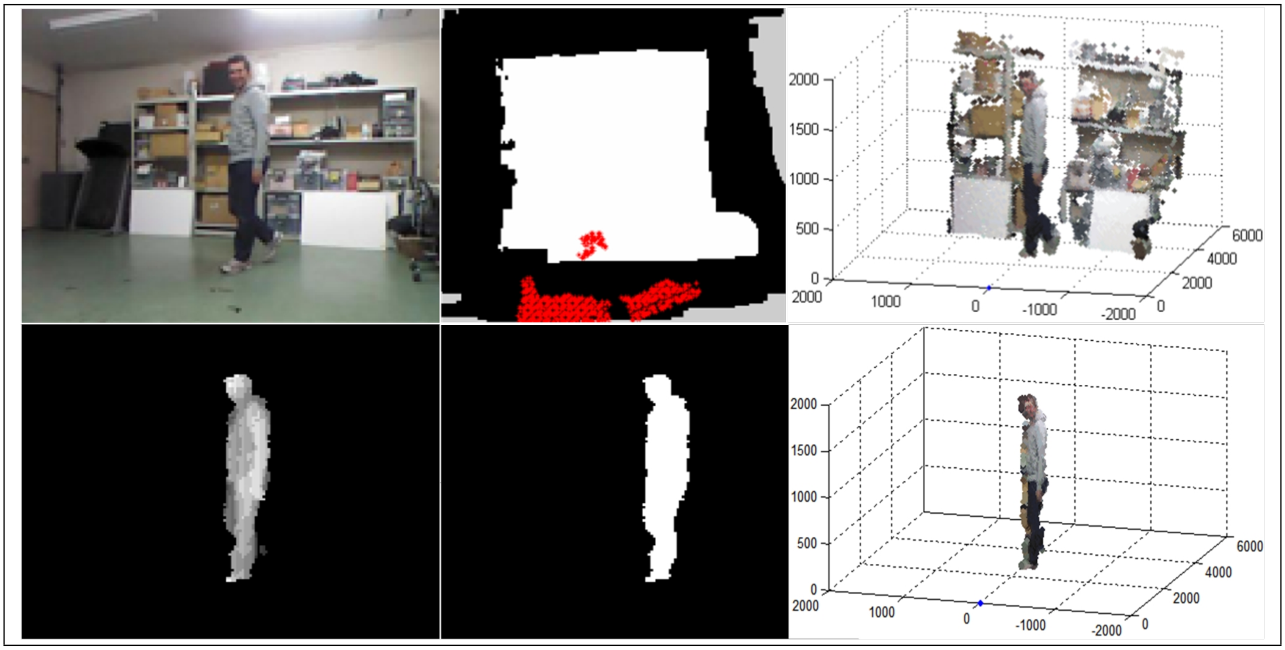

Figure 7: (top row) (a) Color image of Kinect 3D observation, (b) KG(x,y) transformed points are shown on 2D map with red marks, (c) representation of the Kinect data in 3D (bottom row) (a) space based saliency maps with height information of the bed, (b) extracted novelty region from space based saliency map, (c) space based saliency region in 3D

Since our study is to develop algorithms for at-home bio-monitoring, our priority is based on the large subjects (human), where small salient regions on the saliency map can be removed with a defined threshold based on 3D real data clustered regions or 2D image regions. Finally, novelty region extraction is simply done by segmenting all extracted regions, and finding their size on 2D image representation, then the region with the largest size is selected as the novelty region to be tracked as in Figure 7.

In sum, using the prior information, the redundancy during visual search was decreased to find out the novelty of the scene, and then, some simple heuristics such as the closer depths and size of the candidate regions were used for the primary attention region decision as the tracking region on the scene, which are also consistent with the behaviour of HVS.

4. EXPERIMENTAL RESULTS AND DISCUSSION

Before going through an extensive analysis of the algorithm, we started our test with a small number of test data from discrete data of some specific selected cases, in which robot were placed on the map randomly as Subject 1 (S1) was moving on the scene or a scene without any subject. The spatial memory for this test is the global map obtained in a hall of the building (Figure 8).

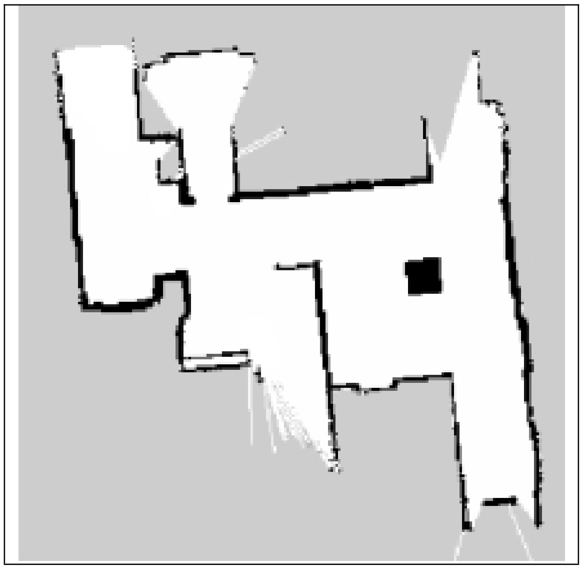

Figure 8: Occupancy map layout for the initial test with Subject 1

The dataset includes 55 frames of data including RGB color image, 3D Kinect data, robot pose, and Kinect pose. Among the 55 images, 22 data is without the subject as expressing the case that the scene does not have any novelty to be detected. And, 23 data includes the subject with different conditions of the subject and environment such as: (i) subject is fully visible, (ii) subject is partially visible (partial occlusion), (iii) subject is in various distance to the robot, and (iv) scenes are under various illumination conditions. Then, using the proposed algorithm in previous section, we analyzed each frame to detect novelties on Dataset-1. In Figure 9, some sample color images of Kinect data are given for the stated conditions with their space based saliency and extracted novelty region outputs.

It can be seen that when there is no novelty (subject) in the scene as in Figure 9(a), space based saliency and extracted novelty regions yields nothing as output. On the other hand, subject can be extracted as novelty from the space based saliency maps (Figure 9(f-h)) as in Figure 9(j-l) by using the spatial memory (global map in Figure 8) robot global data, and Kinect data. With the initial analysis, we tested detection performance in a simple environment with different situations like partial occlusion or partially visible cases (Figure 9(c-d)).

In Table 1, detection results are given with defined data (Dataset-1). For the 23 data with novelty (subject visible), number of true positives (TP) are 23 without any detection error, and the algorithm yielded no false positives (FP), which means that it did not extracted any non-novelty region as novel area instead of the subject. Also, for the 22 data without any novelty, number of true negative (Figure 9(a, e, i) cases) is 22 without any error, and the algorithm did not stated subject as as a part of spatial memory or non-novelty region.

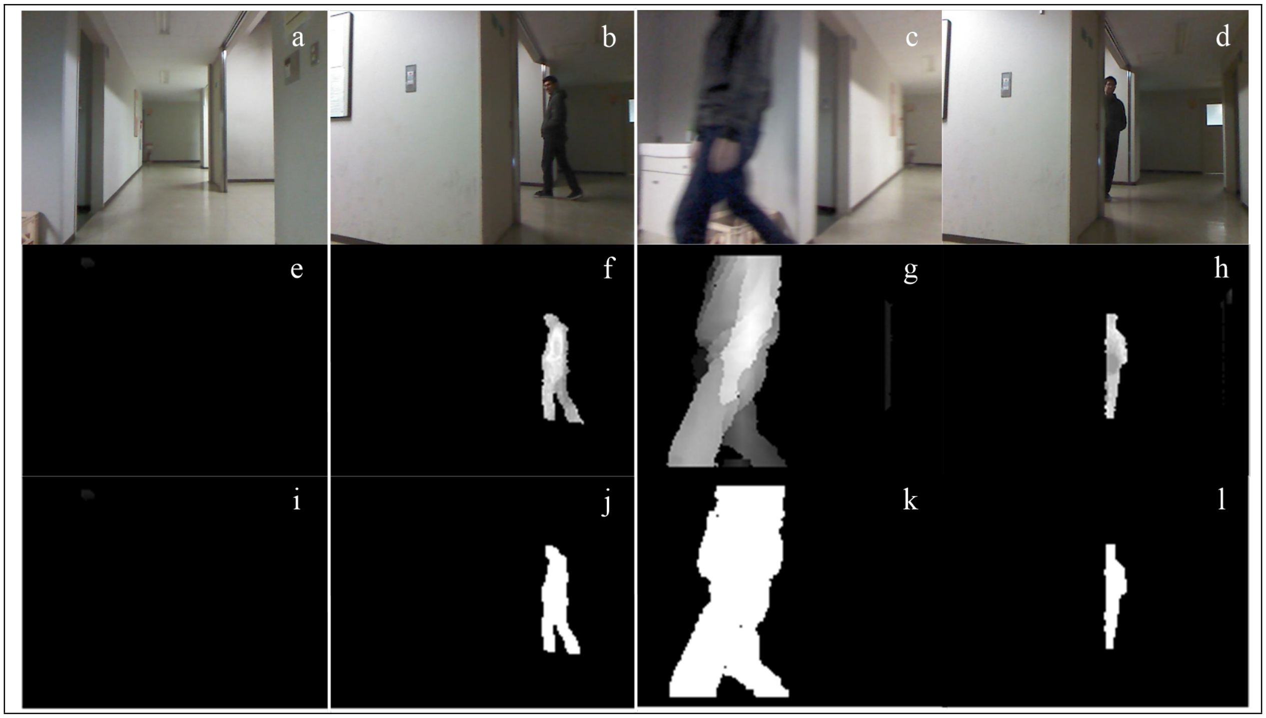

Figure 9: (a) sample image without subject or novelty, (b-d) sample images with subject or novelty representing different conditions such as subject visibility or occlusion and illumination changes, (e-h) space-based saliency maps of the scenes in (a) to (d) respectively, (i-l) extracted novelty regions from saliency maps in (e) to (h)

Table 1: Detection results of the algorithm for Dataset-1

	Conditions	# frames	Accuracy
Novelty Exist (S1) & Without Novelty (55 frames)	TP	23	100%
	FP	0	
	TN	22	100%
	FN	0	
	Overall correct detections	55	100%

In addition, we also analyzed how the proposed algorithm can decrease redundant data for future processing based on novelty region. Our data consists of 120x160 pixel resolution, so it means that the number of points in point cloud is 19200. However, with the initial floor and ceiling removal some of these points neglected. With this removal on Dataset-1, average number of points to be processed decrease to 10883 points, where around 43.32% of the points were removed as redundant data during the initial process. Then, we calculated the average number of points remained on Dataset-1 after proposed novelty detection algorithm applied. On Dataset-1 with novelty existence cases (23 frames as given in Table 1), the average number of interest points decrease to around 2283 points with the proposed novelty extraction, which means that 11.89% of the data requires attention, and 88.11% of the points are redundant in average regarding the Dataset-1.

Experiment on randomly selected data with specific cases gives promising results to go for the next step with a more realistic environment and continuous data, where the subject moves in the room randomly to test detection and tracking performance of the algorithm. In the following section, detection and tracking performance of the systems is tested with more complex conditions.

4.1 EXPERIMENTS FOR COMPARISON

The proposed algorithm is tested with several conditions and compared with selected tracking algorithms by using a different dataset (Dataset-2) including more subjects and scenarios. Since the main purpose of the project is to observe the elderly or people with impairment, detection and tracking model should be able to track persons with different walking patterns in various indoor environmental conditions. For the purpose of testing various simulated conditions, first, we generated a test dataset that consists of 6 recordings from 2 subjects with various illumination conditions, color variations and different walking patterns such as normal walking, simulated impaired walking, and simulated elderly walking.

Table 2: Properties of Dataset-2

Data	#image frames	Subject color	Subject exist in all frames	Light Changes	No-light condition
S2IW	2220	Black	Yes	No	No
S2EW	1245	Green	Yes	No	No
S2NW	1180	Black	Yes	No	No
S3IW	2120	Red	Yes	No	No
S3EW	2000	Red	Yes	No	No
S3NW	2390	Gray	No (172 frames)	Yes (decreased)	Yes (dark room)
Total:	**11155**				

The conditions for each recording data are given in Table-2. First of all, it should be noted that each data includes walking patterns inside the room with changing viewing point or angle. In Table-2, for the datasets that includes 11155 frames in total, recording name abbreviations can be described as; (i) S2IW: Subject 2 and impaired walking, (ii) S2EW: Subject 3 and elderly walking, (iii) S2NW: Subject 3 and normal walking, (iv) S3IW: Subject 3 and impaired walking, (v) S3EW: Subject 3 and elderly walking, (vi) S3NW: Subject 3 and normal walking (Figure 10).

During recordings of the datasets, subjects were wearing different color clothes for each recording set such as red, green, black and gray colors are the subjects color appearances for color tracking models (Figure 10). Red and green were chosen to be the easy to track color while black and gray colors are selected to test the robustness of the color tracker with similar background or highly changing illumination during subject motion. Moreover, in five dataset (S2IW, S2EW, S2NW, S3IW, S3EW), the conditions are set static such that (i) subject exists in all frames, and (ii) same light conditions are used during recording. However, based on the position of the subject, illumination value around the subject will be different due to the distance to the light sources. Moreover, the distance between the subject and robot will affect the texture details too. Hence, there are still dynamical conditions that make tracking difficult for color image data. On the other hand, one of the dataset (S3NW) is prepared with changing environmental and subject appearance conditions. For example, for the S3NW dataset, three different illumination conditions are used; (i) all the lights are on, (ii) illumination is decreased by closing some lights, and (iii) dark room environment as the all lights are off. In addition, because the proposed algorithm is based on novelty detection, existence and non-existence of the novelty region (subject in our case) should be tested too. Hence, in S3NW, 172 frames of 2390 frames do not represent any novelty as the subject is not on the visual area of the sensor.

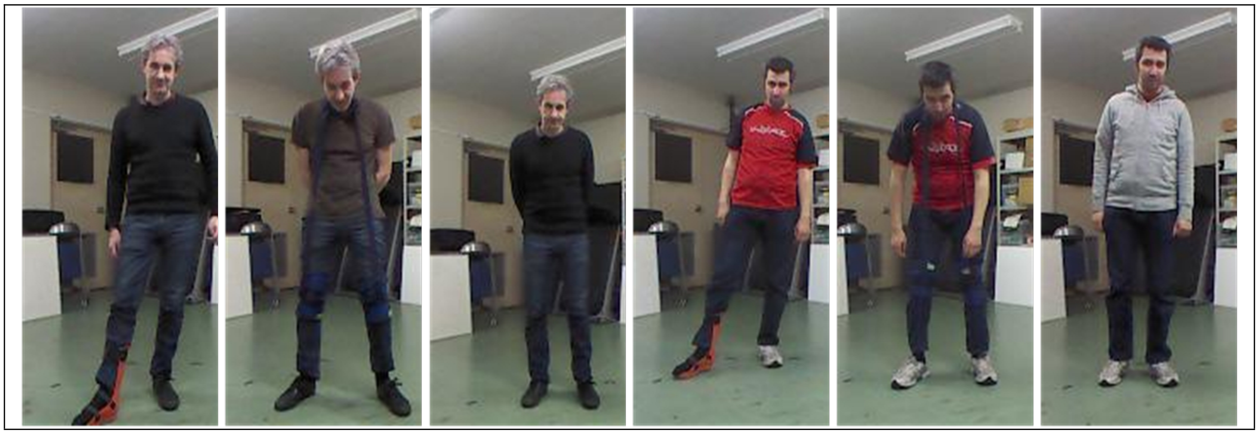

Figure 10: Sample color images for each dataset (a) S2IW: Subject 2 and impaired walking, (b) S2EW: Subject 2 and elderly walking, (c) S2NW: Subject 2 and normal walking, (d) S3IW: Subject 3 and impaired walking, (e) S3EW: Subject 3 and elderly walking, (f) S3NW: Subject 3 and normal walking

4.1.1 Dataset-2 Tests and Comparison Results

With the given dataset, color tracking, depth improved tracking and proposed novelty detection and tracking algorithm are tested and compared. For the color tracker, we have chosen one of recent algorithms in the field, which is Scale and Orientation Invariant Adaptive Mean-Shift Tracker (SOAMST) [19] that can handle low resolution and texture images (e.g. 120x160 pixel images as in our case). The algorithm is easy to integrate it into real-time applications which require low computational cost. In addition, SOAMST [19] is proven to be quite robust to scale, orientation and illumination changes [19] as long as the region of interest (RoI) exists and is trackable on each frame. However, if the illumination value changes a lot that can alter the color information on RoI, SOAMST may not achieve tracking robustly as also stated prior in the related literature review section.

Due to the fact that, Kinect sensor provides depth information locally given the robot position, to compare with the proposed model, we also integrated depth likelihood map to improve the accuracy of tracking similar to the study in [10]. Although depth likelihood map integration, depth improved SOAMST (DI-SOAMST), can improve the result of color tracking [19], it is still not robust enough with the motion of subject and the robot. Moreover, when tracking fails ones in SOAMST and DI-SOAMST, it is very difficult to recover since both algorithms use previous frame's tacking points as a reference to continue next tracking process. If the previous frame's reference tracking point is false or not the subject of interest, then it is highly probable that next process output for the tracking may fail too.

Table 3: Comparison of the Tracking Algorithms with the Dataset-2

Dataset-2	SOAMST [19]			DI-SOAMST [19, 10]			Proposed Model		
	Detection & Tracking			Detection & Tracking			Detection & Tracking		
	Success	Track Fail	Detect Fail	Success	Track Fail	Detect Fail	Success	Track Fail	Detect Fail
S2IW	1668	552	NA	2220	0	NA	2220	0	0
S2EW	968	307	NA	1245	0	NA	1235	0	5
S2NW	1042	138	NA	1180	0	NA	1180	0	0
S3IW	2120	0	NA	2120	0	NA	2120	0	0
S3EW	2000	0	NA	2000	0	NA	2000	0	0
S3NW	1633	585	NA (172)	1484	734	NA (172)	2390	0	0
Total	***9431***	***1582***	***172***	***10249***	***734***	***172***	***11150***	***0***	***5***
Accuracy	***85.60% (Tracking only)***			***93.32% (Tracking Only)***			***99.96% (Detect & Track)***		

In Table 3, the results of tracking tests are given for SOAMST, DI-SOAMST, and proposed novelty detection algorithm. It should be noted that SOAMST and DI-SOMAST tracking only models, where they can not detect or recognize whether the novelty exist or not (subject in the scene or not). Therefore, their detection and tracking results were not count if the subject does not exist in the scene. And, their detection performance (Detect Fail in Table 3) are assigned as not available (NA). On the other hand, the proposed model is detection and tracking algorithm so tracking and detection failures are obtained from the tests.

First of all, from results of S2IW and S2NW, it can be seen from Table 3 that walking pattern may also affect tracking performance of the color based approaches. In these two cases, Subject 2 has the same color appearance model and the layout is the same, the only difference between S2IW and S2NW is the walking pattern (Table 2), where S2IW is simulated impaired walking (Figure 10(a)) and S3NW is the normal walking (Figure 10(c)). And, while S2NW has 83.70% accuracy by having more stable walking gait, S2IW has lower performance as 75.14% tracking accuracy due to the high motion and gait change of the subject that can affect appearance model. On the other hand, color based tracking by using SOAMST can yield very high accuracy independent of walking pattern if the subject appearance model is very distinctive from the environment as in S3IW and S3NW (Table 2) with 100% tracking accuracy (Table 3). And, changing illumination conditions can decrease the tracking performance down to the 73.63% as in S3NW case. In sum overall tracking only performance of the SOAMST is 85.60% from all cases. On the other hand, DI-SOAMST showed promising results by handling SOAMST failures in five cases (S2IW, S2EW, S2NW, S3IW, S3EW) by giving 100.00% tracking performance (Table 3) without any failure when the light condition of the room does not change as in S2IW, S2EW, S2NW, S3IW, S3EW data (Table 2). For the S3NW data, DI-SOAMST is also affected from the changing light conditions of the room (Table 2). DI-SOAMST also failed to track subject in many frames of S3NW data due to loss of appearance model of the subject on the scene with highly changing illumination conditions or dark room. In general, it improved the tracking performance of SOAMST from 85.60to 93.32% as given in Table 3.

By having problem with detection after subject disappearance and reappearance, or with changing illumination conditions, as a color based approaches, both SOAMST and DI-SOAMST is not good enough to handle all day (day and night) monitoring task considering subject detection and tracking. However, Table 3 demonstrates that proposed space based novelty detection and tracking algorithm is a good candidate and a reliable model for monitoring by having 99.96% detection and tracking accuracy among 11150 frames of all cases. Also, detection and tracking output of a frame is not affected by the detection or tracking errors in previous frames, which is not the case for SOAMST, DI-SOAMST or many color based state of the art tracking approaches.

In sum, proposed space based saliency model for novelty detection and tracking task yielded promising results for at-home monitoring mobile robot project since all the datasets in indoor environment and their results have very high accuracy. Moreover, this algorithm can be extended or integrated to other algorithms easily if more complex approaches are necessary for indoor surveillance tasks such as multiple person tracking and identification, activity recognition, change detection, and etc.

5. CONCLUSION

In this study, we demonstrated that the idea of spatial working memory during visual search as in Human Visual Systems (HVS) can be used to calculate saliency map from real world data. Therefore, we proposed an algorithm to create space-based saliency map for novelty detection and tracking by fusing two different sensors, in which Lidar is used for 2D global mapping and localization, and Kinect is used for perceive the local 3D data of the scene. Then, using prior environment knowledge, we showed that it is possible to detect and track subjects as the novelty of the scene by paying attention to the region that is different from the spatial memory, in other words, global map. With this approach, an efficient and fast model is obtained for indoor mobile robot based tracking by having more than 99.00% detection and tracking accuracy in tested Datasets.

As a future work, this model can be improved by using 3D global mapping and localization or height information integration for more complex environments instead of 2D approach even though current approach is good enough to detect people for monitoring walking activities. Also, it can be integrated with semantic map algorithms to work on a more semantic level. Moreover, multiple person detection and identification algorithms in more complex situation can be handled faster by decreasing the redundancy of RGB-D based algorithms with the proposed novelty detection.

References

[1]. Rahidi, P. and A. Mihailidis, *A survey on ambient-assisted living tools for older adults*. IEEE Journal of Biomedical and Health Informatics, 2013. **17**(3): p. 579-590.

[2]. Gross, H.-M., et al., *Progress in developing a socially assistive mobile home robot companion for the elderly with mild cogntive impairment*. IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS), 2011. p. 2430-2437.

[3]. Nergui, M., et al., *Human gait behaviour classification on lower body triangular joint features*. IASTED Int. Conf. on Signal and Image Processing (SIP), 2012. p. 212-219.

[4]. Cesta, A., et al., *The ROBOCARE assistive home robot: Environment, features, and evaluation*, The ROBOCARE Technical Reports, 2004. RC-TR-0906-6.

[5]. Jayawardena, C., et al., *Design, implementation and field tests of a socially assistive robot for the elderly*. The Fourth IEEE RAS/EMBS Int. Conf. On Biomedical Robotics and Biomechatronics, 2012. p. 1837-1842.

[6]. Li, R. and M. A. Oskoei, H. Hu, *Towards ROS based multi-robot architecture for ambient assisted living*, IEEE Int. Conf. on Systems, Man, and Cybernetics, 2013. p. 3458-3463.

[7]. Simpson, R. C., et al., *NaVChair: An assistive wheelchair navigation system with automatic adaptation*, Assistive Technology and AI, Lecture Notes in Artificial Intelligence (LNAI), 1998. **1458**: p. 235-255.

[8]. Wei, Z., W. Chen, and J. Wang, *Semantic mapping for smart wheelchairs using RGB-D camera, Journal of Medical Imaging and Health Informatics*, 2013. **3**(1): p.94-100.

[9]. Myagmarbayar, N., et al., *Human Activity Recognition Using Body Contour Parameters Extracted from Depth Images*. Journal of Medical Imaging and Health Informatics, 2013. **3**(3): p. 455–461.

[10]. Imamoglu, N., et al., *An Improved Saliency for RGB-D Visual Tracking and Control Strategies for a Bio-monitoring Mobile Robot*. Communications in Computer and Information Science, 2013. **386**: p. 1-12.

[11]. Liu, H., S. Chen, and N. Kubota, *Intelligent video systems and analytics: A survey*. IEEE Trans. on Industrial Informatics, 2012. **8**(1): p. 49-60.

[12]. Luo, R. Cc. and C.-C. Chang, *Multisensor fusion and integration: A review on approaches and its applications in mechatronics*. IEEE Trans. on Industrial Informatics, 2013. **9**(3): p. 1222-1233.

[13]. Yilmaz, A., O. Javed, and M. Shah, *Object tracking: A survey*. ACM Computing Surveys, 2006. **38**(4) - Article 13: p. 1-45.

[14]. Gupta, A.M., et al., *An on-line visual human trcking algorithm using SURF-based dynamic object model*. IEEE Int. Conf. on Image Processing (ICIP), 2010. p. 3875-3879.

[15]. Bay, H., et al., *Speeded-up robust features (SURF)*. Computer Vision and Image Understanding, 2008. **110**: p. 346-359.

[16]. Lowe, D.G., *Distinctive image features from scale-invariant keypoints*. International Journal of Computer Vision (IJCV), 2004. **60**(2): p. 91-110.

[17]. Borji, A., et al., Adaptive object tracking by learning background context. IEEE Int. Conf. on Computer Vision and Pattern Recognition (CVPR), 2012. p. 23-30.

[18]. Nguyen, D.T., et al., *Object detection using non-redundant local binary patterns*. IEEE Int. Conf. on Image Processing (ICIP), 2010. p. 4609-4612.

[19]. Ning, J., et al., *Scale and orientation adaptive mean shift tracking*. IET Computer Vision, 2010. **6**(1): p. 52-61.

[20]. Talha, M. and R. Stolkin, *Particle filter tracking of camouflaged targets by adaptive fusion thermal and visible spectra camera data*. IEEE Sensors Journal, 2014. **14**(1): p. 159-166.

[21]. Garcia, G.M., et al., *Adaptive multi-cue 3D tracking of arbitrary objects*. Pattern Recognition, Lecture Notes in Computer Science (LNCS), 2012. **7476**: p. 357-366.

[22]. Jia, S., et al., *Robust human detecting and tracking using varying scale template matching*. IEEE Int. Conf. on Information and Automation, 2012. p. 25-30.

[23]. Liu, J., et al., *Real-time human detection and tracking in complex environmentss using single RGBD camera*. IEEE Int. Conf. on Image Processing (ICIP), 2013. p. 3088-3092.

[24]. Fang, Y., et al., *Saliency detection in the compressed domain for adaptive image retargeting*. IEEE Trans. on Image Processing (IEEE TIP), 2012. **21**(9): p. 3888-3901.

[25]. Imamoglu, N., W. Lin, and Y. Fang, *A saliency detection model using low-level features based on wavelet transform*. IEEE Trans. on MultiMedia, 2013. **15**(1): p. 96-105.

[26]. Fink, G.R., et al., *Space-based and object-based visual attention: shared and specific neural domains*. Brain, 1198. **120**: p. 2013-2028.

[27]. Logan, G.D., *The CODE theory of Visual Attention: An integration of space-based and object-based attention*. Psychological Review, 1996. **103**(4), p. 603-649.

[28]. Mozer, M.C. and S.P. Vecera, *Space-and object-based attention*. Neurobiology of Attention, Academic Press, 2005. p. 130-134.

[29]. Frintrop, S., et al., *A component based approach to visual person tracking from a mobile platform*. International Journal of Social Robotics, 2010. **2**(1), p. 53-62

[30]. Zhang, G., Z. Yuan, and N. Zheng, *Key object discovery and tracking based on context aware saliency*. International Journal of Advanced Robotic Systems, 2013. **10**(15): p. 1-12.

[31]. Yang, J. and M.-H. Yang, *Top-down visual saliency via joint CRF and dictionary learning*. IEEE Int. Conf. on Computer Vision and Pattern recognition (CVPR), 2012. p. 2296-2303.

[32]. Wang, J., et al., *A computational model of stereoscopic 3D visual saliency*. IEEE Trans. on Image Processing (IEEE TIP), 2013. **22**(6): p. 2151-2165.

[33]. Chamaret, C., et al., *Adaptive 3D rendering based on region-of-interest*. Proc. of SPIE, 2010. **7524**: p. 75240V.

[34]. Ouerhani, N. and H. Hugli, *Computing visual attention from scene depth*. IEEE 15th International Conf. on Pattern Recognition, 2000, **1**: p. 375 -378.

[35]. Fang, Y., et al., *Saliency detection for stereoscopic images*. IEEE Trans. on Image Processing (IEEE TIP), 2014. **23**(6): p. 2625-2635.

[36]. Begum, M., et al., *Object- and space-based visual attention: An integrated framework for autonomous robots*. IEEE/RSJ Inter. Conf. on Intelligent Robots and Systems (IROS), 2008. p. 301-306.

[37]. Garcia, G.M. and S. Frintrop, *A computational framework for attentional 3D object detection*. Proc. of the Annual Meeting of the Cognitive Science Society, 2013.

[38]. Chun, M.M. and J.M. Wolfe, *Visual Attention*. Handbook of Sensation and Perception (Chapter 9), Edited by E. B. Goldstein, Blackwell Publishing, 2005. p. 273-310.

[39]. Chun, M.M. and K. Nakayama, *On the functional role of implicit visual memory for the adaptive deployment of attention across scenes*. Visual Cognition, 2000. **7**: p. 65-81.

[40]. Oh, S.-H. and M.-S. Kim, *The role of spatial working memory in visual search efficiency*. Psychonomic Bulletin & Review, 2004. **11**(2): p. 275-281.

[41]. Johnston, W.A., et al., *Attention capture by novel stimuli*. Journal of Experimental Psychology, 1991. **119**(4): p. 397-411.

[42]. Zhou, Y., et al., *Region based high-level semantics extraction with CEDD*. 2nd IEEE Int. Conf. on Network Infrastructure and Digital Content, 2010. p. 404-408.

[43]. Nuchter, A. and J. Hertzberg, *Towards semantic maps for mobile robots*. Robotics and Autonomous Systems, 2008. **56**: p. 915-926.

[44]. Mason, J. and B. Marthi, *An object-based semantic world model for long-term change detection and semantic querying*, IEEE/RSJ Inter. Conf. on Intelligent Robots and Systems (IROS), 2012. p. 3851-3858.

[45]. *Robot Operating System (ROS)*, http://wiki.ros.org/

[46]. *ROSARIA*, http://wiki.ros.org/ROSARIA

[47]. Grisetti, G., C. Stachniss, and W. Burgard, *Improved techniques for grid mapping with rao-blackwellized particle filters*. IEEE Transactions on Robotics, 2007. **23**: p. 34-46.

[48]. *ROS-Gmapping*, http://wiki.ros.org/gmapping

[49]. Thrun, S., W. Burgard, and D. Fox, *Probabilistic robotics*. The MIT Press Cambridge, 2005.

[50]. *Costmap*, http://wiki.ros.org/costmap_2d

[51]. *Transformation Matrix*, http://mathforum.org/mathimages/index.php/Transformation_Matrix