Towards Visual Mapping in Industrial Environments - A Heterogeneous Task-specific and Saliency Driven Approach

J. Rafid Siddiqui[1], Henrik Andreasson[2], Dimiter Driankov[3] and Achim J. Lilienthal[4]

Abstract— **The highly percipient nature of human mind in avoiding sensory overload is a crucial factor which gives human vision an advantage over machine vision, the latter has otherwise powerful computational resources at its disposal given today's technology. This stresses the need to focus on methods which extract a concise representation of the environment inorder to approach a complex problem such as visual mapping. This article is an attempt of creating a mapping system, which proposes an architecture that combines task-specific and saliency driven approaches. The proposed method is implemented on a warehouse robot. The proposed solution provide a priority framework which enables an industrial robot to build a concise visual representation of the environment. The method is evaluated on data collected by a RGBD sensor mounted on a fork-lift robot and shows promise for addressing visual mapping problems in industrial environments.**

I. INTRODUCTION

Although there has been significant advancements in visual mapping, some basic questions still remain in inquisitive minds. How can the tremendous amount of visual data received through sensors be filtered into a concise representation? How should an autonomous system, for example, a mobile robot, make use of such representation so that it could diligently perform the very specialized tasks for which it has been designed and yet be able to generalize its understanding in order to handle the dynamic nature of the environment? The advancement in Convolutional Neural Network (CNN) exposes some of the limits of human perception. The success of CNN on huge datasets for image classification tasks challenges the human vision by exposing its limits on one hand while leaving many questions in the mind on the other hand [1], [2]. It indicates that the capacity to recognize large classes of objects is maybe not the main strength of human vision. The reason why humans are better in tasks such as object perception is perhaps due to intelligent use of resources by building a concise representation of the environment. It is perhaps the selection criterion used for quantifying the content of visual input which differentiates human vision from machine vision.

The process of visual memory in humans is mostly coupled with the quality of the stimuli. As the brain is modulated by a chemical based reward system, which can be associated with any stimuli, quantification of such quality becomes subjective. However, there are some consistent patterns of stimuli preference due to evolution. Such preferences can be broadly classified into two major categories: a) task-specific b) stimuli driven. It is this prioritized stimuli which passes the strong barrier of human visual filtration framework and get register as long term memory. The reason for such a stringent filtration mechanism can be explained from an evolutionary stand point. While being under constant threat from predators and requiring robust detection of food sources for survival, it was essential to be able to limit the sensory information to the most crucial parts. Task specific object perception happens when humans change the visual attention towards certain pre-learned object, which is the subject of most object detection methods. Stimuli specific object perception occurs when object draws the visual attention towards itself and is the subject of most saliency based methods. These strategies (i.e. task-specific and stimuli driven) combined together can give a satisfactory answer to the aforementioned questions.

The work reported in this paper is an effort to combine the aforementioned two domains of perception and generate a coherent model for visual representation of the environment, which could be useful for certain practical applications (e.g. a fork-lift robot operating in a warehouse environment). While there exists a reoccurring pattern among different warehouses due to a common infrastructure (e.g. pallets, shelves, etc.), there is also a large difference among them due to variety of goods handled in each warehouse. One way of tackling this duality is to develop a hybrid perception framework. The major contribution of this paper is twofold: a) a novel visual mapping architecture is proposed which combines the task-specific perception with salience perception that builds a 3D occupancy map in real time b) a novel method for salient region detection which makes use of global as well as local saliency cues while incorporating some of the evolutionary indicators of saliency detection. Modeling of task specific perception is achieved by training classifiers for the most important objects in the warehouse environment. The stimuli driven perception is achieved by computing a saliency measure for the various segments of the scene (Section II). The visual map builds and updates the occupancy model of the environment as well as its visual appearance by combining task and saliency information (Section III). The results of the evaluations are presented and discussed in Section V and related work can be found in Section IV followed by conclusions in Section VI.

[1] Rafid Siddiqui is post-doc at Örebro University, Sweden. rafid.siddiqui@oru.se

[2] Henrik Andreasson is associate professor at Örebro University, Sweden. henrik.andreasson@oru.se

[3] Dimiter Driankov is professor at Örebro University, Sweden. dimiter.driankov@oru.se

[4] Achim J. Lilienthal is professor at Örebro University, Sweden. achim.lilienthal@oru.se

II. Perceptual Filtration

Perceptual filtration is the process of systematic reduction of sensory input into meaningful abstractions which could enable the building of a concise visual representation of the scene. This section describes the process used in order to extract task specific targets as well as salient objects.

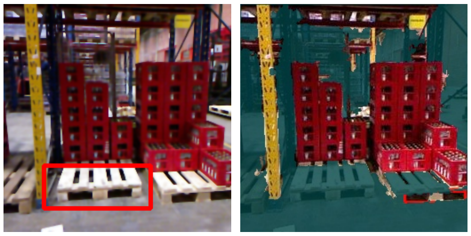

Fig. 1: Perceptual Filtration: Task-specific and Salient Regions.

A. *Task Specific Perception*

The task specific object perception is a learned behavior therefore, it can be modeled as supervised learning problem. In the recent past there has been substantial progress in the performance of supervised object recognition classifiers. First prominent break-through of the last decade happened when Viola and Jones [5] proposed their method based on boosted classifiers. A more recent development happened when CNN showed some remarkable results on the image-net benchmark challenge, which contains millions of real world images of objects categorized in hundreds of classes [3], [6]. Although CNN is still state-of-the-art and perhaps the best choice for huge datasets with hundreds of classes to learn, it is prone to over-fitting when the training set is small, which is mostly the case in real-world scenarios. There has been models proposed that tries to circumvent this problem by reducing the size of network [4], however, the required size for the training images remains large.

Along side the progress in CNN based methods, there has also been substantial improvements in Support Vector Machine (SVM) based classifiers. Notably, object detection based on learned deformable object parts has provided a more generalized interpretation of object representation [7]. The method works by finding the object parts represented by Histogram of Oriented Gradients (HOG) at multiple scales and building a constraint graph of the parts. This strategy of describing objects by its parts comes close to human object perception and allows rich visual information which can be exploited in later stages of a visual system [7]. However, processing time is the major drawback of this technique.

Most object detectors work by extracting a subset of overlapping candidate windows which are grouped into two sets; one positive and other negative. These sets are used to train a binary classifier that predicts to which class a particular window belongs. The false positives are controlled by an iterative refinement method where the wrong classifications are added to the negative set and the model is updated. While performing the training, only a subset of windows are taken for computational efficiency, which badly affects the optimization solution. This problem is mitigated by optimizing the classifier on all the candidate windows as in the Maximum Margin Object Detection (MMOD) method [8], which provides a faster optimization solution. This substantially reduces the false positive rates and therefore makes the object detector reliable for real-time applications such as robotics. Since the application of this work is targeted for a mobile robot platform operating in a warehouse environment which need to identify a set of critical objects for performing its operations, MMOD is used for task specific object detection. The simplicity of MMOD for training an unknown object, requiring only few examples, makes it a good choice for highly demanding applications such as industrial robots. More specifically, a concise set of object classes are identified which is needed by a fork-lift robot to perform its operation. For every given training image of an object class, HOG features are extracted and a structured SVM classifier is trained on the features. Given the learned model, object detection is performed on new images and the output is forwarded to the mapping system, which integrates the information into a visual map. An example of task-specific and saliency perception is shown in Figure 1.

B. *Stimuli Driven Perception*

Saliency contains an element of subjectivity due to its dependence on the color cue. This is in contrast to popular belief that saliency is only an innate property of the object. Not all salient objects are salient to everybody. However, there are certain patterns which are common across humans and which can be modeled. Despite a drastic difference in the number and distribution of cones cells; Long (L), Medium(M) and Short (S) across people, a general color perception exist that remains almost constant (e.g. pure red is considered as a sign of danger almost over the globe) [12]. Shiny objects as well as highly distinctive colors or orientations from the surroundings appear to pop-out. This observation has been the basis of almost all of the saliency based object perception methods. The underlying principle of center-surround maps used in most of the state-of-the-art methods [9], [10] is grounded in biological evidence of cells in Lateral Geniculate Nucleus (LGN) which has similar structure. However, as pointed in [15] most of these methods suffer from a bias which tends to assume fixation as a function of an object's property. While on the contrary the saccadic eye movement based imaging is on one hand very different than camera imaging which captures the whole FOV at once rather than in saccades, and on the other hand there are usually more than one salient objects in a natural scene. This is in contrast to single centered object images for which most these methods perform better. Additionally, the concept of pixel-level salient object segmentation is different compared to how humans performs it [15]. If a human fixate on a salient object, it is not as if the whole FOV gets dark and it is no longer possible to fixate on other salient objects at that moment. The saliency methods, especially the ones which tends to find salient object based on color contrast tend to ignore the fact that some colors bind more solidly with long term memory than others [13].

Building upon these observations, a new color cue based saliency method is proposed which computes saliency by taking into account local saliency, global saliency and color precedence. The saliency method is incorporated into a mapping system using a random fixation strategy which not only allows the system to become real-time but also helps to keep the overall map size small.

1) Color Preprocessing: Color representation is a challenging task, hence the reason of the large amount of color spaces available. Biologically inspired opponent color space is grounded in the opponent theory of colors which compliments the tri-stimulus theory of colors and therefore it is well suited for saliency related tasks. Although there are different formulations for converting RGB space to opponent space the basic framework remains the same. RGB color is first transformed to an intermediate format (i.e. CIE XYZ), which then is converted to a biologically inspired color space denoted LMS space. LMS space is further converted to Red-Cyan, Blue-Yellow, and White-Black contrasting channels. A RGB color is transformed to XYZ space and LMS space as follows:

$$\begin{pmatrix} L \\ M \\ S \end{pmatrix} = \begin{pmatrix} 0.7328 & 0.4296 & -0.162 \\ -0.703 & 1.6975 & 0.0061 \\ 0.0031 & 0.0136 & 0.9834 \end{pmatrix} \begin{pmatrix} X \\ Y \\ Z \end{pmatrix} \tag{1}$$

The LMS space is further converted to opponent channels as follows:

$$\begin{pmatrix} O1 \\ O2 \\ O3 \\ O4 \end{pmatrix} = \begin{pmatrix} 1 & -0.5 & -0.5 & 0 \\ -0.5 & 1 & -0.5 & 0 \\ -0.5 & -0.5 & 1 & 0 \\ 0.33 & 0.33 & 0.33 & 0 \end{pmatrix} \begin{pmatrix} L \\ M \\ S \\ 1 \end{pmatrix} \tag{2}$$

The image encoded in opponent color space is used in global and local saliency computations. An example of opponent color space is depicted in Figure 2.

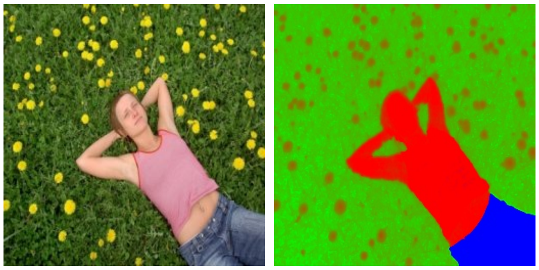

Fig. 2: Visualization of Opponent Space in RGB colors.

2) Global Saliency: Salient objects have least in common, whereas the background usually have homogeneous texture spread across the image. The global saliency computation is based on this assumption. Salient regions are often small-sized and have high variation while background on the contrary tends to repeat over large regions. This pattern can be easily filtered by a low-pass filter. In this work, a Log-Gabor filter is used. Log-Gabor filters have a significance as they model the processing done by cells in LGN. A log Gabor filter can be defined as follows:

$$\mathcal{G}(\boldsymbol{x}) = \exp\left(\frac{-1}{2\gamma^2}\left(\log\frac{\|\boldsymbol{x}\|}{f_0}\right)^2\right) \tag{3}$$

where $\boldsymbol{x} = (x, y)$, f_0 and γ are spatial coordinates of filter in frequency domain, initial frequency and bandwidth of the filter respectively. The global saliency map is hence given as:

$$\mathcal{S}_g(O_i) = O_i * \mathcal{G}(\boldsymbol{x}) \tag{4}$$

where O_i represent the image channel in opponent color space and $*$ is the convolutional operator. Examples of the global saliency is visualized in Figure 5.

3) Local Saliency: Local saliency is often computed for individual pixels based on their difference to the colors of surrounding pixels. The computation of pixel-wise saliency is not only computationally expensive but it is also counter intuitive because regions with similar colors tend to have same saliency. Therefore, pixels are grouped into homogeneous regions based on their color using a mean-shift segmentation [11]. The obtained scene segments form the input to the local saliency computation. The central idea is that salient objects have contrasting color with respect to their surrounding. This phenomenon can be modeled by computing the color contrast between regions. More specifically, for every scene region i and neighbor region j the distance in opponent color space is used to compute saliency as follows:

$$s_i = \frac{1}{n}\sum_{j=1}^{n}\|\boldsymbol{c}_i - \boldsymbol{c}_j\|^2 \exp\left(-\sum_{j=1}^{n}\frac{\|\boldsymbol{u}_i - \boldsymbol{u}_j\|^2}{2\sigma^2}\right) \tag{5}$$

where $\boldsymbol{c_i}$, $\boldsymbol{c_j}$, $\boldsymbol{u_i}$, $\boldsymbol{u_j}$ are; color in opponent space of region i and j, spatial position of the region i and neighbor j respectively.

As discussed earlier, certain colors contribute more to saliency, for example, warm colored objects are perceived as salient more often than objects with cool colors. Therefore a saliency computation method should also take into account color precedence. A biological explanation of such behavior lies in the way cone cells behave under different illumination conditions. Under lit conditions, a Photopic vision is practiced by the cells. Which means that under Photopic vision, L cells gets more preference than S cells and therefore the red becomes the most dominant color and blue the least dominant color. This is the reason for the sky being the least salient region of the scene in day light. Under low light conditions, a Scotopic vision is practiced by the cone cells which enhances the output of S cells thus resulting in a reversed preference. Under medium conditions, colors compete for resources, which is called Mesotopic vision. Such a preference can be modeled by measuring color temperature. Color temperature is the temperature of a black-body radiator at the same hue as the light source. A measure for color temperature is Color Correlation Temperature (CCT). In order to embed the color preference in the system, local saliency map is computed as follows:

$$\mathcal{S}_l = \sum_{i=1}^{n} e^{-\frac{\mathcal{T}(c_i)^2}{2\sigma^2}} s_i \mathcal{M} \tag{6}$$

where $\mathcal{M}_{uv} = \begin{cases} 1 & u, v \in \mathcal{R} \\ 0 & otherwise \end{cases}$ represents the mask region which has image coordinates u and v. $\mathcal{T}(c_i)$ is the CCT function which computes color temperature. Note that color temperature is reversed to represent the saliency, because humans' perception of warm colors is inversely related to physical temperature of the color which is highest at blue.

4) Saliency Map: Given the local and global saliency maps, a joint saliency map is computed by integrating both maps. A combined saliency map is thus given as:

$$\mathcal{S} = \alpha \sum_i \mathcal{S}_g^2(O_i) + (1-\alpha)\mathcal{S}_l^2 \tag{7}$$

where α is the weight factor between local and global saliency.

Salient region extractions along with local and global saliencies is shown in Figure 5.

5) Fixation Strategy: Human saliency is not only dependent on certain properties of the object but is also affected by the region which has already been observed. This is accomplished by saccadic eye movements. There is lack of consensus in whether this eye movement is solely dependent on low level features, which tends to generate saliencies, or if it is an intentional act. Therefore, in this work the fixation is modeled as a hybrid strategy which combines the strength of randomness as well as the precision of intentional act. A number of candidate scene sections are generated where the image is divided into a set of large overlapping candidate sub-regions by a sliding window approach. A sample from the candidate sections is drawn randomly which represents the focus region. The coherent regions in the focus region are evaluated for their respective saliency. Regions that have saliency higher than a threshold are selected as fixated salient regions and are then used in the next stage where they are integrated into the visual map. A sample of focus region and corresponding saliency region extraction steps are shown in Figure 3.

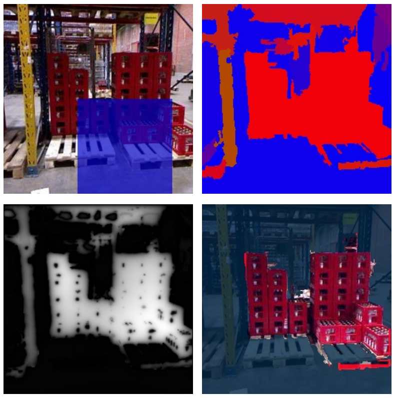

Fig. 3: Salient Fixation: Focus region, color segmentation, saliency map and detected salient region respectively.

III. Visual Map

Visual mapping is often tightly coupled with the occupancy of the scene. The representation is typically done either by accumulating raw sensor data or by maintaining a grid map estimated from multiple sensor readings, where each grid cell represents the occupancy status of a region in the real world. These sensor readings or occupancy cells are then labeled by finding a pattern which is present in the neighborhood. Such techniques can suffer from scaling problem due to redundancy which exist in the real-world data. The reason for this is because occupancy is usually unnecessarily tied to visual interpretation of the scene. While it is important to know the occupancy in the scene for a robot to plan the pathway to its goal, it is not necessary for a visual interpretation of the environment (i.e. visual memorization in terms of key objects/regions in the scene and associations that exist among them). Furthermore, many mapping solutions are proposed by keeping a human observer in mind and not the machine for which it is in-fact proposed. Perhaps humans do not save the information about peculiar obstructions in the pathway to long term memory. One counter argument would be that the memorization of specific spatial details of the environment perhaps simplify the planning process or enable the generation of plans offline. However, the pre-build highly constrained plans may not work due to the dynamic nature of the environment. Humans make short term node-node loosely coupled dynamic plans which are not tied to the structural peculiarities. This work presents a simple strategy which addresses the problem of visual mapping in a warehouse environment. The proposed visual map consists of two major components: a) Scene Occupancy b) Perceptual Regions. A complete architecture of the proposed method is given in Figure 4. Scene occupancy consists of information about obstructions in the scene and can be modeled using occupancy mapping. Due to its simplicity and efficiency an octree [14] representation has been used in this work. Initially, an empty octree is created. Each RGB and depth pair from a Kinect Sensor is processed to create a labeled point-cloud containing task-specific and salient regions. Using the labeled point-cloud and the robot pose, the occupancy and labels in the visual map are updated. A set of new occupied cells are added to the tree for new unseen regions and already occupied but unassigned nodes are labeled with the detected task-specific object.

Perceptual regions consist of task-based and stimuli driven perception along with spatial information. Every perceptual region is a self-sufficient coherent representation which consists of information about the appearance and position of

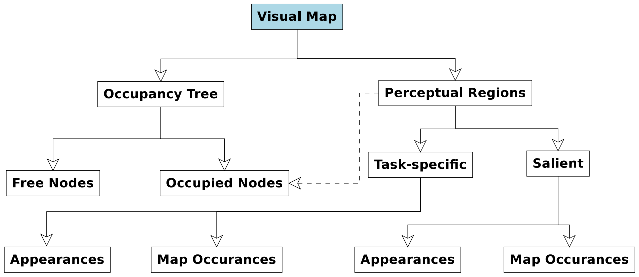

Fig. 4: The architecture of the proposed Visual Map

a particular stimuli (i.e. an object in case of task-based perception and a salient region in case of stimuli-driven perception). An overview of such an organization is given in Figure 4. The reason for keeping scene occupancy alongside the visual interpretation is only for visualization purposes since the occpuancy map is not required for computing perceptual regions. A salient region occupies an unknown label as the robot is unaware of the corresponding name of the object (or region), however, any salient region could be transformed into a known one by a semi-supervised step. This framework also enables independent development of two subsystems. The complete system has been made highly modular therefore, each component (i,e. task-based perception, saliency perception and occupancy) can be improved independently in order to generate better mapping system.

IV. Related Work

The proposed technique, which attempts to bring multiple domains of visual interpretation (i.e. 3D metric mapping, saliency, object detection and abstract scene representation) closer makes it difficult to point out similar work with sufficient overlap. Nevertheless, the closest domains of visual mapping solutions in terms of the underlying motivation and concept are: scene summarization and semantic mapping based on object/place associations. Scene summarization approaches tend to seek an abstract summary of the visited places by finding spatial properties of objects in the scene. In [16] an image based semantic summarization approach is proposed which tries to capture the visual experience of a robot by clustering images into ordinary and novel categories. Similarly in [17] an algorithm for selecting the most important parts of a scene, which are sufficient to build a representation of the scene is proposed. In [18] a method for finding novel objects in the scene is proposed which utilizes the given information about the known objects including and the map of the environment. In [19] and [20] topological place associations are taken as basis for developing an abstract representation of the environment.

V. EXPERIMENTAL RESULTS

The goal of this work has been to build a representation, a visual map, to assist a warehouse robot which perform loading and unloading operations. A fork-lift robot operating in a warehouse of a superstore has been used to collect the data. The robot has odometry sensors along with various of vision sensors where data from a Kinect sensor has been used in this work. The depth images are pre-processed in order to reduce the effect of noise at object edges as well as at the reflected parts of the scene. The points which have high depth variance in the neighborhood (in this work $>$ 0.5m for the 8 closest readings) are removed. A bilateral filter is applied on the depth images in order to smooth out depth values while keeping the object boundaries intact. For each RGB image, task specific perception and stimuli driven perception is performed and corresponding perceptual regions are updated. The result of perceptual regions along with depth is transformed into a point-cloud by projective geometry of the camera which is used to update the occupancy map. The performance of the task specific perception is evaluated on a set of 122 labeled images selected randomly from the recorded sequences among which two third are used for training. A set of 100 images with manually labeled salient regions has been used as ground truth for the evaluation of stimuli performance. The percentage of correctly identified salient pixels for each marked region is used as measure for evaluation. The performance of the task-specific and stimuli driven detection systems can be seen in the Figure 7. As mentioned earlier, one of the reason for choosing MMOD is that it generates lower false alarms. While evaluating the detection system, it became clear that MMOD indeed had tendency of generating lower false alarms (i.e. in comparison to Viola and Jones), as it generated only a couple of false positives on the 200 images of warehouse environment which did not contain any of the target objects. After observing this behavior, we decided to stress test the algorithm by increasing the number of negative images. Therefore, we collected a set of 5000 indoor images from the SUN dataset [21]. The algorithm has been run multiple times and Receiver Operating Curves (ROC) has been obtained by varying the margin of the support vector and training the model. A comparison in terms of ROC curves with Viola-Jones is depicted in Figure 7. The results verify the claim that MMOD indeed has fewer false detections than its counterpart. The reason that ’pillar’ class have slightly more steep curve than other objects, could be explained by the fact that negative images used in the experiments (i.e. from the SUN dataset) were indoor images of various real-world places which indeed contained pillar or pillar like structures. The output of the task-specific as well as stimuli driven systems on the warehouse dataset can be visualized in Figure 6. It can be observed that the system tends to give preference to highly salient as well as distinctive objects in the scene which are indeed the kind of objects which ought to be put in long-term memory (i.e. the visual map). The system is tested on recorded sequences and the final map snapshot is obtained which is visualized in Figure 8. The figure also presents a 3D model based representation of the perceptual regions. The runtime analysis of individual components of the system as well as the total processing time is given in Figure 9. The saliency and map update components have

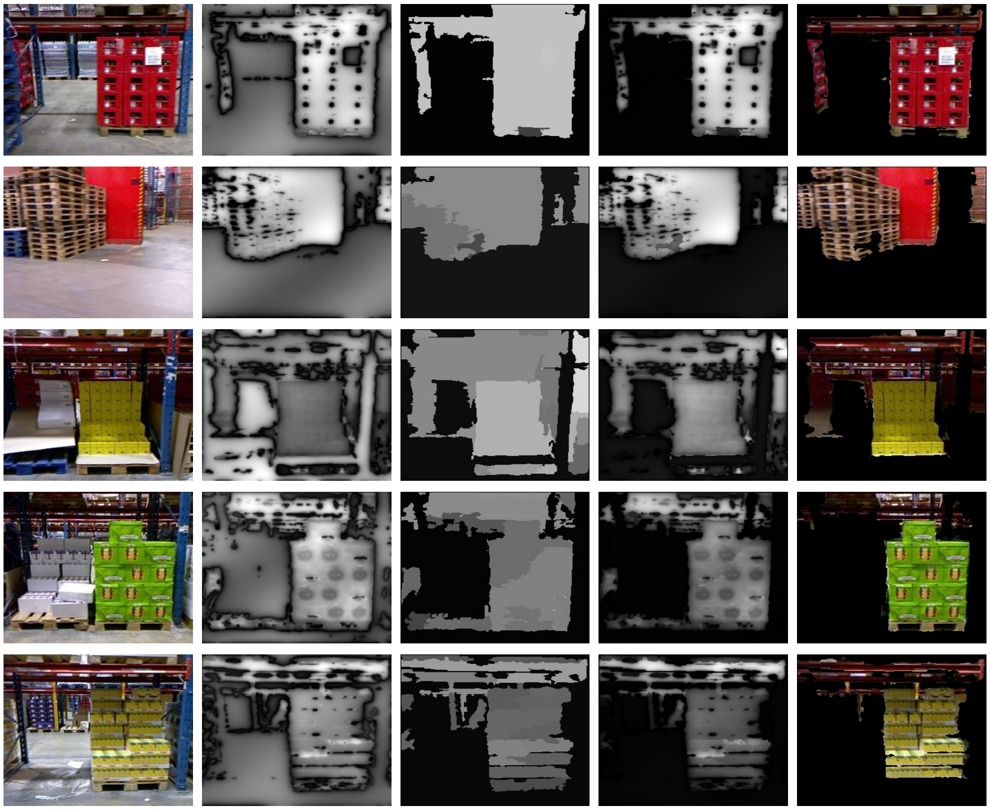

Fig. 5: Visualization of Salient Region Extractions. Column 1: original image, Column 2: global saliency, Column 3: local saliency, Column 4: joint saliency map, Column 5: salient regions

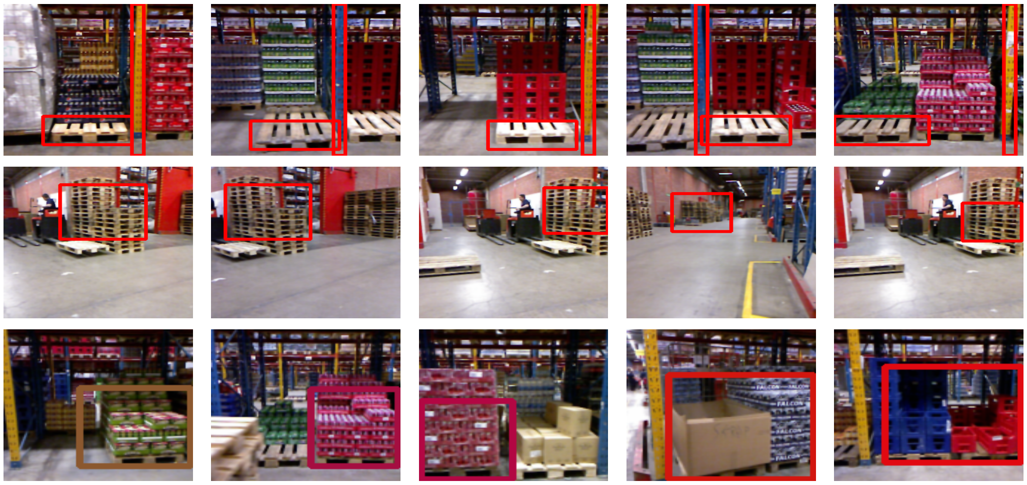

Fig. 6: Visualization of Perceptual Region Detection for Mapping. Row 1: pallets/pillars, Row 2: depositories and Row 3: salient regions.

large variance, this is due to variations in the segmentation step which generates arbitrary number of regions which directly affects the runtime of local saliency and perceptual region computations. There are two major portions of map process; perceptual region update and occupancy update. If occupancy map is turned off the computational resources for that component could be taken by feature representation technique which encapsulates the appearances of the perceptual regions. Currently, mean color of the region and average HOG are used to memorize the appearance of a perceptual region. The system is implemented in Robot Operating System (ROS).

VI. CONCLUSIONS

The real world is often dynamic and contains an abundance of information, most of which is unnecessary to

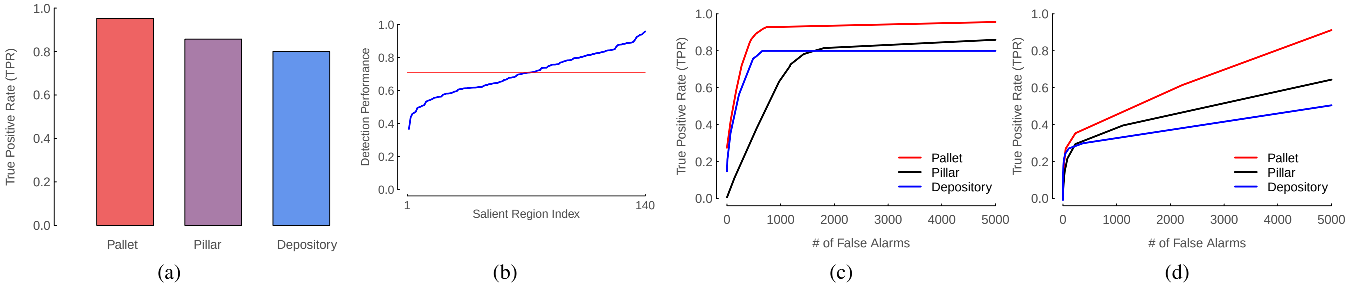

Fig. 7: Performance Evaluation: (a) Performance of Task-specific Region Extraction. (b) Performance of Salient Region Extraction - Blue: Percent overlap of the detected region with ground truth, Red: mean line. (c) ROC curves of MMOD detector. (d) ROC curves of Viola-Jones detector.

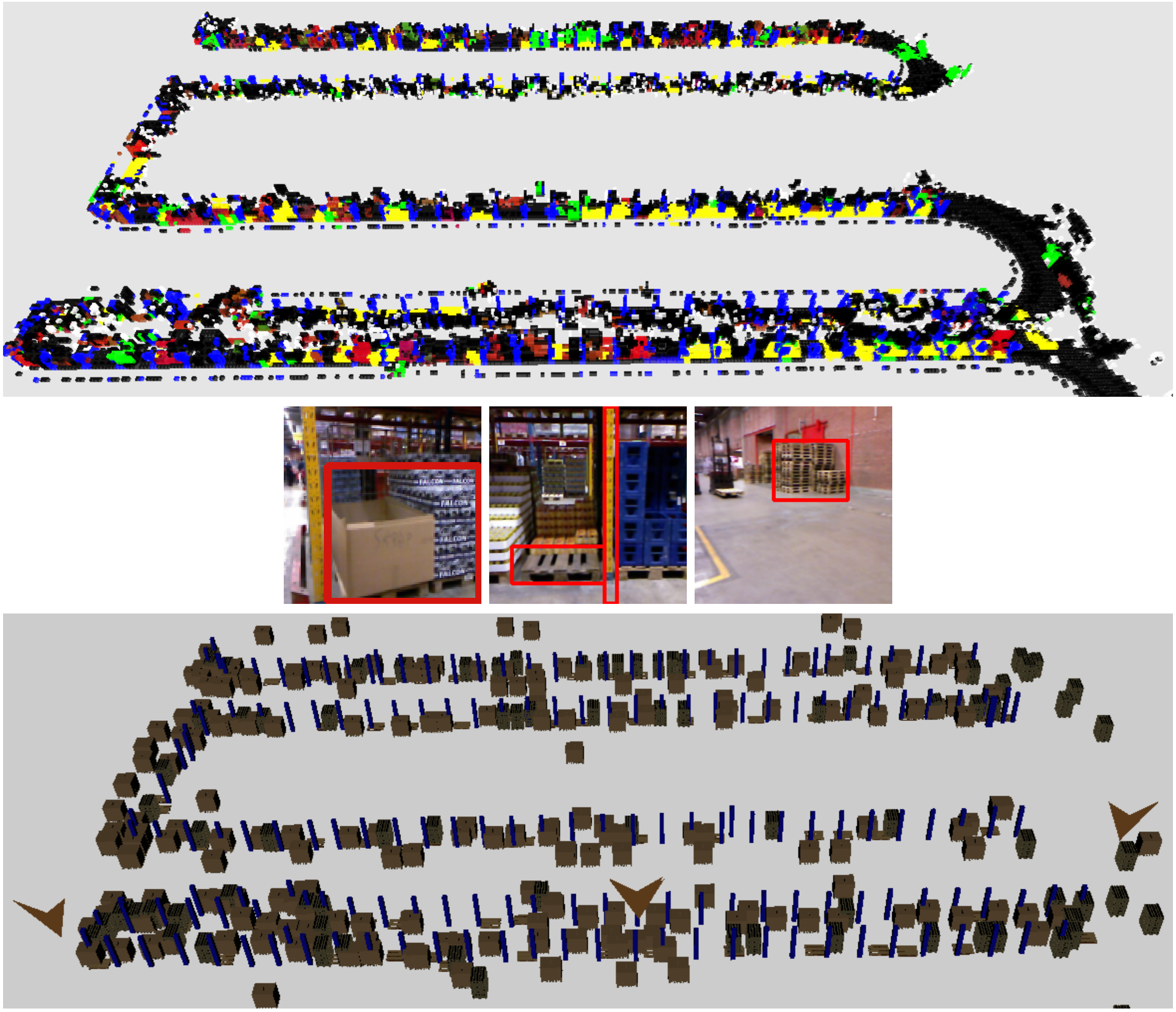

Fig. 8: Visual Map, Top: Occupancy representation (black='occupied space', yellow='pallets', green='pallet depositories', blue='pillars' and all other colors depict salient regions). Middle: detection of salient regions; pallet, pillar and depository respectively. Bottom: Perceptual regions represented with 3D models, the arrows indicates the position towards the objects shown in the middle row.

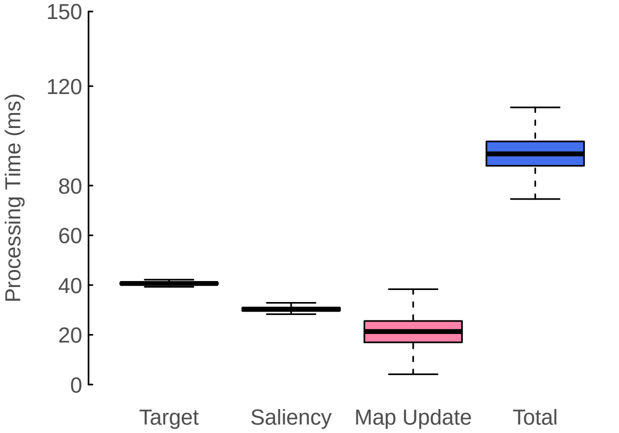

Fig. 9: Runtime of task-specific, saliency computation and map update step respectively.

perform a specific type of task. On one hand, the robot is expected to perform highly specialized repetitive tasks such as loading/unloading specific objects and on the other hand, it is also expected that robots seamlessly work in a dynamic environment and become aware of its surrounding yet not getting overwhelmed by the amount of sensory data to be processed every second. Therefore, a robot must have a general framework for deciding what should be kept in long-term memory. An attempt towards providing an answer to this basic yet highly challenging question is performed in this work. A visual map system is proposed which takes decision based on task specific and salient nature of a stimuli and builds a representation of the environment which can be used not only to simplify the map but also to increase the effectiveness and generality of the developed robots. A video presenting the results is available [22].

This work is part of an ongoing research project ”Semantic Robots” funded by *KK-Foundation*.

REFERENCES

[1] He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. in Computer Vision and Pattern Recognition (CVPR), 2015.

[2] Ciresan, D., U. Meier, and J. Schmidhuber. Multi-Column Deep Neural Networks for Image Classification. in IEEE Conference on Computer Vision and Pattern Recognition (CVPR),Rhode Island, 2012, pp.3642-3649.

[3] Russakovsky, Olga, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, et al. ImageNet Large Scale Visual Recognition Challenge., International Journal of Computer Vision, 2015, pp.1-42.

[4] Lin, Min, Qiang Chen, and Shuicheng Yan. Network In Network. Neural and Evolutionary Computing, 2014, pp.1-10.

[5] Viola, Paul, and Michael Jones. Robust Real-Time Object Detection., International Journal of Computer Vision, 2001.

[6] Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. ImageNet Classification with Deep Convolutional Neural Networks. in Advances in Neural Information Processing Systems 25, 2012.

[7] Felzenszwalb, P.F., R.B. Girshick, D. McAllester, and D. Ramanan. Object Detection with Discriminatively Trained Part-Based Models. in IEEE Transactions on Pattern Analysis and Machine Intelligence 32, no. 9, 2010, pp.1627-1645.

[8] King, Davis E. Max-Margin Object Detection. in Computer Vision and Pattern Recognition, 2015.

[9] Achanta, R., S. Hemami, F. Estrada, and S. Susstrunk. Frequency-Tuned Salient Region Detection. in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2009, pp.1597-1604.

[10] Cheng, Ming-Ming, Guo-Xin Zhang, N.J. Mitra, Xiaolei Huang, and Shi-Min Hu. Global Contrast Based Salient Region Detection. in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2011, pp.409-416.

[11] Comaniciu, Dorin, and Peter Meer. Mean Shift: A Robust Approach Toward Feature Space Analysis. in IEEE Transactions on Pattern Analysis and Machine Intelligence, 24, no. 5, May 2002, pp.603-619.

[12] Hofer, Heidi, Joseph Carroll, Jay Neitz, Maureen Neitz, and David R. Williams. Organization of the Human Trichromatic Cone Mosaic., The Journal of Neuroscience: The Official Journal of the Society for Neuroscience 25, no. 42, 2005, pp.9669-9679.

[13] Kuhbandner, Christof, Bernhard Spitzer, Stephanie Lichtenfeld, and Reinhard Pekrun. Differential Binding of Colors to Objects in Memory: Red and Yellow Stick Better than Blue and Green., Frontiers in Psychology 6, 2015.

[14] Hornung, Armin, Kai M. Wurm, Maren Bennewitz, Cyrill Stachniss, and Wolfram Burgard. OctoMap: An Efficient Probabilistic 3D Mapping Framework Based on Octrees. Autonomous Robots 34, no. 3, 2013, pp.189-206.

[15] Li, Yin, Xiaodi Hou, C. Koch, J.M. Rehg, and A.L. Yuille. The Secrets of Salient Object Segmentation. in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp.280-287.

[16] Paul, R., D. Rus, and P. Newman, How Was Your Day? Online Visual Workspace Summaries Using Incremental Clustering in Topic Space. in IEEE International Conference on Robotics and Automation (ICRA), 2012, pp.4058-4065.

[17] Paul, R., D. Feldman, D. Rus, and P. Newman, Visual Precis Generation Using Coresets., in IEEE International Conference on Robotics and Automation (ICRA), 2014, pp.1304-1311.

[18] Kollar, T., and N. Roy. Utilizing Object-Object and Object-Scene Context When Planning to Find Things., in IEEE International Conference on Robotics and Automation, 2009, pp.2168-2173.

[19] Murphy, L., and G. Sibley. Incremental Unsupervised Topological Place Discovery., in IEEE International Conference on Robotics and Automation (ICRA), 2014, pp.1312-1318.

[20] Pronobis, A., and P. Jensfelt. Large-Scale Semantic Mapping and Reasoning with Heterogeneous Modalities., in IEEE International Conference on Robotics and Automation (ICRA), 2012, pp.3515-3522.

[21] J. Xiao, J. Hays, K. Ehinger, A. Oliva, and A. Torralba., ”SUN Database: Large-scale Scene Recognition from Abbey to Zoo”, IEEE Conference on Computer Vision and Pattern Recognition, 2010.

[22] J. R. Siddiqui, H. Andreasson, D. Driankov, and Achim. J. Lilienthal.: ”Video Result”, 2016, [Online] Available: https://drive.google.com/file/d/0B-4NEXz3rBTxVVdGS0JpMHZyeE0/view?usp=sharing