Received: 4 July 2021 | Revised: 3 December 2021 | Accepted: 6 December 2021

IET Image Processing

DOI: 10.1049/ipr2.12399

The Institution of Engineering and Technology WILEY

ORIGINAL RESEARCH PAPER

Structure-aware multiple salient region detection and localization for autonomous robotic manipulation

Sudipta Bhuyan[1] | **Debashis Sen**[2,4] | **Sankha Deb**[3,4]

[1] Advanced Technology Development Centre, Indian Institute of Technology Kharagpur, West Bengal, India

[2] Dept. of Electronics and Electrical Communication Engineering, Indian Institute of Technology Kharagpur, West Bengal, India

[3] Dept. of Mechanical Engineering, Indian Institute of Technology Kharagpur, West Bengal, India

[4] Centre of Excellence in Advanced Manufacturing Technology, Indian Institute of Technology, Kharagpur, India

Correspondence
Debashis Sen, Department of Electronics and Electrical Communication Engineering, Indian Institute of Technology Kharagpur, West Bengal 721301, India.
Email: dsen@ece.iitkgp.ac.in

Funding information
Tata Steel

Abstract

This paper proposes a multiple salient region detection and localization approach for unstructured industrial robot work environments with arbitrarily located and orientated objects. Different from the existing, the authors' novel technique to detect multiple salient regions performs locally adaptive center-surround operations on proto-object partitions obtained through color consistency and spatial proximity analysis. The multi-scale center-surround operations are done by masks that are local structure-aware yielding regions with precise and accurate boundaries as required for robotic manipulation. First, experiments to evaluate the multiple salient region detection performance are carried out using four standard databases having images with multiple salient objects. Quantitative result analysis using F-measure, shuffled F-measure, shuffled AUC and MAE, and subjective result inspection suggests that the proposed approach is in general better at collectively detecting multiple salient regions than the state-of-the-art, including those based on deep learning. Then, real-life experiments involving robotic manipulation are carried out to demonstrate the utility of the multiple salient region detection method. For robotic manipulation, object localization is improved after salient region detection by employing a fast shadow detection algorithm proposed based on hue analysis, and recognition through existing matching techniques is applied only at the localized salient regions. The benefit of the novel multiple salient region detection approach in the robotic manipulation system is shown using localization and pose estimation accuracy, rates of detection and recognition, positional and angular errors, and processing speed.

1 | INTRODUCTION

Owing to the revolution in robotics, manufacturing industries are increasingly using robots for a number of complex operations such as assembly, bin picking, material sorting etc. Repeatability, as well as high precision, have been achieved using off-line/on-line programmed robots to perform tasks in structured environments of mass production systems [1]. Autonomous machines/robots performing active processes like detection and recognition have also been employed to execute complex tasks in unstructured environments [2]. However, to perform the tasks quickly and reliably, they must have human-like sensory abilities, especially in vision. It is well accepted that the ability of humans in executing complex tasks significantly surpasses that of robots, although efficiency decreases while performing the tasks repeatedly for a long time.

Hence, providing an autonomous industrial robot human-like capabilities is of utmost importance as manufacturing industries require customizable mass production systems, which are quick, reliable, and agile enough to handle material handling diversity and variation in product design [1]. In an unstructured industrial environment, the primary necessity is to detect and localize each object quickly and accurately in the scene viewed by the robot. In vision, speed and accuracy are contradictory requirements that humans achieve by attending to the environment in a selective manner, which is called "selective visual attention" [3] for object detection and localization [4]. Salient object detection algorithms have been used to automatically select the informative parts in an image [5–7]. Human visual attention is simulated by saliency detection, which can be used to significantly reduce the time for searching and recognizing an object in an image. A brief discussion on

existing multiple salient region detection methods is given in Section 2.1.

In recent times, deep learning-based end-to-end object recognition has been attempted for industrial applications a few times [8–10] with reasonable success. However, their performance, in general, relies heavily on accurate and vast labeled data [11], which are hard to generate for an industrial environment and vary from one environment to another. Therefore, many of the latest methods for object recognition in robotic manipulation are still based on local feature point matching [12–14]. All the feature points in the entire image are exhaustively searched to match them with those of the template, and the performance of such an approach deteriorates due to the presence of background clutter in an unreliable industrial environment and similar industrial objects [2]. Another popular approach of object recognition in robotic manipulation for industrial applications is based on shape matching. Several shape analysis based methods have been proposed for object recognition [15–21], where the entire image is searched for particular objects given their templates. Both the matching approaches are prone to performance deterioration due to varying illumination condition, and the presence of dust and dirt [22], which are expected in an industrial environment. Therefore, instead of applying the above recognition techniques over the whole image, if matching can be applied only at the potential object locations, it will boost the performance and accuracy of the whole system. Here, humans' strategy of selective visual attention implemented through proper salient region detection and localization would be effective in achieving both speed and accuracy for industrial robotic manipulation. A brief review of recognition methods used for robotic manipulation is given in Section 2.2.

In this paper, we propose an approach to detect and localize multiple salient regions in images particularly to perform intelligent robotic manipulation, which is inspired by the selective visual attention strategy of humans. We then design a system to guide a robotic arm by looking for objects quickly with uncompromising accuracy. The selective visual attention of humans is implemented using the proposed approach to detect multiple salient regions. Although a plethora of saliency computation and salient object detection algorithms are available, they are highly sensitive to the large number of parameters involved [23]. While the saliency computation approaches are good at detecting many important areas, they do not provide accurate object boundaries [23]. On the other hand, salient region detection approaches are tuned to detect only a few highly salient objects (in general only one) in the scene with proper boundaries [23]. For a vision-guided robot to perform tasks such as pick and place, the foremost requirement is the detection of all objects to be manipulated by the robot as salient regions with proper boundaries, without the necessity to change a set of initially supplied parameters.

We propose such a multiple salient region detection and localization approach that focuses on detecting multiple objects. In our approach, first, we obtain local image structure through proto-object partitioning based on the mean-shift procedure considering color consistency and spatial proximity analysis. Next, center-surround operator masks adapted to the already obtained local image structure are generated. Such local structure adapted masks are used at multiple scales to perform center-surround operations. Multiple maps are generated from these operations, which are then combined to produce a single map with delineating region signatures that is directly used for detecting multiple salient regions with precise boundaries.

The proposed multiple salient region detection approach is compared with state-of-the-art salient region detection approaches including those based on deep learning. Observation of qualitative results and quantitative analysis through F-measure, shuffled AUC, MAE and our newly designed shuffled F-measure considering multiple standard databases with images having multiple salient objects demonstrate that the proposed approach in general outperforms the state-of-the-art in detecting multiple salient regions. This indicates that the proposed approach is effective as a generic multiple salient region detection method.

After detecting multiple salient regions with precise boundaries, the subsequent aim is to localize the salient regions for manipulation by a robot. The object localization approach is designed to be robust in the presence of shadows, avoiding related errors. To achieve this robustness, we propose a fast shadow detection technique based on hue values of the detected salient regions. Contrary to existing shadow removal techniques [24–26], our primary goal is to quickly detect (but not remove) only the shadow areas and avoid them during subsequent processing.

After objects are detected and localized using our approach, any recognition technique can be employed in our system. For our industrial object manipulation experiments, recognition through two popular matching techniques, namely, feature point matching and shape matching are chosen. The combined area of the detected and localized objects regions is substantially smaller when compared to the entire image. Therefore, we achieve much faster recognition by applying the recognition approaches only at the detected and localized object regions. Moreover, as the matching processes are restricted only to the relevant areas, the recognition performance also improves. Therefore, our system is designed to detect, localize and recognize multiple industrial objects simultaneously in a single image at different scales and orientations, and also to estimate their poses accurately. To demonstrate the utility and efficiency of our entire system (that includes multiple salient region detection followed by shadow detection and recognition) in performing robotic object manipulation, experiments are carried out to compare the proposed robotic manipulation procedure with the standard procedure in terms of localization, pose estimation accuracy, positional and angular error, detection and recognition rates, and processing speed. For this, we deploy a vision-guided industrial robot manipulator for performing autonomous pick and place operations.

To summarize, the main contributions of the paper are:

- We propose a proto-object partitioning driven structure-aware multiple salient region detection approach, which is specifically designed to detect multiple salient objects, and the proposed approach is validated against the state-of-the-art

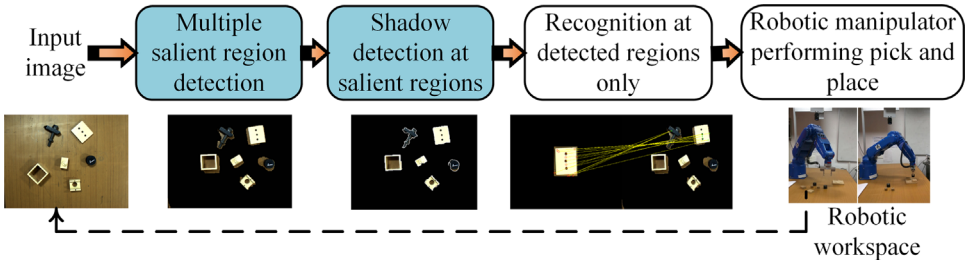

FIGURE 1 The proposed framework for vision guided manipulation by industrial robots with the novel modules contributed by this paper highlighted in blue

salient object detection through extensive experiments that includes the use of our newly designed shuffled F-measure.
- The proposed multiple salient region detection approach uses the local structure information from the proto-object image partitions to design masks that capture boundaries of multiple salient objects properly, which are at a standard that can be used for localizing the objects for robotic manipulation.
- We propose a simple and fast hue based shadow detection algorithm to improve the localization accuracy of the detected salient objects for autonomous robotic manipulation, and the proposed algorithm is validated for the task at hand.
- We show the utility of the proposed multiple salient region detection and localization approach in robotic pick and place operations, for which a well-defined strategy for collision-free robotic operation is designed and a CAD model is provided to that effect.

A concise graphical representation of our entire system is shown in Figure 1. This work is a substantial improvement over our preliminary work reported in [27]. The rest of the paper is organised as follows. In Section 2, existing works related to our proposal are discussed. Our proposed approaches for detecting multiple salient regions and shadows are presented in Sections 3 and 4, respectively. In Section 5, other parts of the proposed system for autonomous robotic manipulation of objects are described. The comprehensive experimental results presented in Section 6 demonstrate the effectiveness our novel proposals. Section 7 concludes the paper.

2 | RELATED LITERATURE

In this section, we briefly present the most relevant existing literature corresponding to the different modules of our entire system. Hence, we present the literature under three categories: (1) salient object /region detection, (2) object detection and recognition techniques for robotic manipulation, and (3) shadow detection techniques.

Over the past couple of decades, a plethora of approaches have been proposed for visual attention modeling, saliency computation, and salient region detection. A detailed review of salient region detection can be found in [23, 28]. We briefly discuss a few salient region detection approaches in Section 2.1, particularly focusing on those aimed at detecting multiple objects.

A vast literature is available on object detection and recognition for robotic manipulation in an unstructured environment. Among them, approaches based on feature point matching and shape matching have been the popular ones for robotic manipulation [2, 17, 29–31]. Here in Section 2.2, we review quite a few such recent object recognition approaches.

Although the presence of shadows can degrade object localization performance, shadow detection is not considered to be an inherent part of an object localization system. Shadow detection followed by its removal and then inpainting are often clubbed together as a single separate system [32]. As shadow removal and inpainting are generally not required in robotic manipulation, we suggest that only the shadow detection component be considered and proper object boundaries be obtained during localization by avoiding the detected shadow areas. Hence, we also discuss in Section 2.3, a few shadow detection approaches which are parts of existing shadow removal systems.

2.1 | Multiple salient object/region detection

Salient object detection approaches in general aim to detect and segment the most (or a few highly) salient object/s in a scene with proper boundaries, and only a few have explicitly attempted to detect multiple salient objects [23]. Most works on salient object detection are based on capturing the uniqueness, rarity, local and global contrasts etc., at different locations in a scene.

The methods of [33–37] and [38] combine contrast cues with higher-level guidance like center prior, spatial distribution prior, region uniformity prior etc., to detect the most salient object in the scene. [39] and [40] detected a salient object by combining backgroundness prior with contrast to detect salient objects. As often image boundary pixels are part of the image background, backgroundness prior is computed from the image boundary. As models using such backgroundness prior may fail if an object is connected to a boundary, to overcome this issue, [33] and [41] proposed boundary connectivity priors for salient object detection. For the same, [42] measured a pixel's connectivity to the image boundary by taking the minimum barrier distance. A couple of works like [43, 44] presented models for the detection of single as well as multiple salient objects in an image. [43] presented a model where spatial saliency clues for salient

object region detection are obtained from multiple-level clustering of regional features and recursive processing of the clustering results. [44] performed salient object detection using a graph-based optimization framework, where they used multiple graphs instead of one to describe different image properties. A multiple-instance learning based saliency detection framework that combines low, mid and high level features for detection is proposed by [45]. [46] employed visual saliency to detect and segment objects lying in a plane for robot navigation.

Recently, deep learning-based models have been employed for salient object detection. [47] proposed a convolutional neural network model which uses a multi resolution 4 × 5 grid structure to combine local and global information. [6] proposed a boundary-aware salient object detection architecture consisting of a densely supervised encoder-decoder network for saliency prediction and a residual refinement module for salient object map refinement. [48] proposed a framework for salient object detection where larger resolution features of shallower layers are discarded by using a partial decoder and features of deeper layers are integrated to obtain a precise salient object map. In [49], first, a coarse global prediction is performed using various global saliency cues and then, the details of salient object maps are refined hierarchically by integrating local context information. In [50], a more advanced feature representation is obtained by directly integrating multi-level features. [51] proposed a salient object detection model where context-aware multi-scale features are extracted and a bidirectional structure is used to pass a message between them. The framework of [52] integrates multi-level feature maps into multiple resolutions, which simultaneously incorporate coarse semantics information and fine details. These feature maps are combined at each resolution to predict different maps, which are fused to obtain salient objects. Recently, a multi-scale feature fusion framework is proposed by [53] that fuses multi-scale features using a search cell and a search space containing relevant information only. [54] proposed a single round training approach for weakly supervised salient object detection and the proposed aggregation module fuses features from multiple levels to estimate the saliency map. [55] proposed a multiview clustering method for detecting coherent groups by a structural context descriptor designed based on structural properties of individuals. Feature points are clustered based on orientation and context similarities. To demonstrate better detection in uneven indoor lighting and complex indoor environment,[56] proposed a joint target detection using RGB-D image based on faster R-CNN algorithm.

2.2 | Recognition approaches used in robotic manipulation

Most of the literature [2, 29, 30, 57, 58] on pose estimation of an object for robotic manipulation from a single image are based on finding the best fit correspondences between input image features and the features of stored database images. Harris corner detector in combination with SIFT descriptor is used as a feature for recognition of objects in [58]. Similarly, [57] and [2] proposed Iterative Clustering Estimation (ICE) algorithm for recognizing objects in complex scenes. Through ICE, feature clustering for object correspondences, and pose estimation are performed iteratively. While [59] performed object recognition by combining SURF features with color histogram, [60] used SIFT for deformable object recognition. Similarly, [30] presented Hough transform-based clustering of SURF features and [29] used a modified SIFT for object recognition. Instance recognition system of [61] is based on SIFT, color, and shape based features. [14] proposed a method for category-level object manipulation where an object is represented by using 3D keypoints. A method of object recognition and pose estimation is proposed in [13], where features are extracted from the colored point cloud and feature descriptor is built using local texture and shape information. Correspondence between a scene and point cloud is established by performing matching using the obtained descriptors and the final pose is estimated by applying Hough transform and RANdom Sample Consensus. [62] proposed an instance recognition and object localization approach where a sparse feature model for training is built by structure key points obtained from shape and texture cues. For each object, feature descriptors are obtained by using Signature of Histograms of Orientations and SIFT. Though feature point matching performs well for textured objects, recognition of textureless objects by it usually turns out to be unsatisfactory as the number of key points extracted is often not sufficient for a proper description [22, 63].

Shape-based object detection and recognition techniques consider precise object boundaries and match object shapes or an edge descriptor at each discrete point in object boundary to the discrete point in a given template where a suitable cost minimizes. [17] proposed a technique where the correspondence between the shape primitives obtained from the captured scene and the boundary representation of the computer-aided design (CAD) model is established by matching feature vectors obtained from geometric properties and their relationship. While [31] used shape (2D contour) and appearance (RGB histogram) to represent 2D models of objects, [16] represented an object by taking shape and contour primitives. Similarly, object's contour representation is used in [18], based on which robot vision system detects generic classes of objects in a cluttered environment. [15] identified graspable objects by first segmenting the scene and then performing shape matching between a segmented object and known object models. Similarly, [22] used Chamfer matching along with the CAD model for object pose estimation. [19] proposed an algorithm for detection and manipulation of cylindrical objects by detection of elliptic shape primitives. Elliptic shape primitive is detected by using elliptic edge curvature and by splitting complex curves into arcs. [20] performed 3D shape matching using the local reference frame (LRF), which is constructed on the local surface and used as a local feature descriptor. Although shape-based recognition is in general reliable, the underlying algorithms are usually of higher complexity compared to the feature-based ones, as the latter works only on extracted key points [22, 63].

Deep learning-based object recognition has been considered lately for robotic manipulation. [64] used R-CNN for

recognition of multiple objects for bin-picking operation. Different grasp poses prediction for multiple objects using deep convolutional neural networks are presented in [65]. [66] proposed a deep learning-based method for indoor object recognition using color knowledge and scene knowledge as deep features. Although deep learning methods have shown reasonable success in industrial robotic applications, their performance often heavily depends on the availability of appropriate vast labeled data from the environment [11]. As industrial environments vary substantially, the availability of the required data can not be assumed.

[67] proposed a target grab position detection technique using a candidate region suggestion network. The advantage of the proposed model is that it can handle external environmental interference. To demonstrate better grasping of space debris, [68] proposed a dual arm space robot with hollow shaped end-effector pairs and caging pair method to capture complex objects.

2.3 | Shadow detection

Most of the state-of-the-art methods [25, 32, 69–71] on shadow detection are based on combined use of shadow variant and invariant image features. Chromaticity based approaches [25, 32, 69, 70] assume that chromatic components are shadow invariant while intensity varies in the region underlying shadow. Similarly, texture based models [25, 70] assume that textures are shadow invariant. Hence, shadowed and non-shadowed regions can be distinguished by measuring texture similarity and intensity variation. [32] considered illumination invariant sensing that produced only non-shadow edges, and compared it to corresponding RGB image to detect shadow. For improved performance, some authors proposed a learning-based approach [25, 70–72] to detect shadow regions. [71] identified shadow regions by considering texture, illumination, and gradient statistics features and the system is trained using CRF to classify shadow and non-shadow regions. A region-based approach is proposed in [25] where a pairwise classification of segmented regions is done based on their color and textural information. Shadow detection by using a convolutional neural network is proposed in [72] where feature learning is done at a super-pixel level and along image boundaries. Though it performs well in most cases, it fails at thin shadow regions [72]. [73] and [24] took user assistance for shadow detection. [74] proposed a distraction-aware shadow detection network (DSDNet) where input image features are augmented with learned distraction features for detection of shadows. [75] proposed a scGAN model and introduced a sensitivity parameter to the generator which controls the sensitivity of the shadow detector. [26] presented the so-called ST-CGAN framework composed of two stacked CGAN to jointly learn shadow detection and removal. [76] proposed a method where image context is analyzed in a direction-dependent manner to detect shadows and to learn spatial context in four directions for which, an RNN module is used. A method of shadow detection and removal by integrating feature fusion and dictionary learning is presented by [77]. Although most shadow detection approaches mentioned above could be used for robotic manipulation, preference must be for the techniques with a lower computational complexity which will enable the robot to operate faster.

3 | PROPOSED MULTIPLE SALIENT REGION DETECTION APPROACH

Our novel approach's framework for detecting multiple salient regions, which is application-oriented and adapts to local structure, is presented in Figure 2. First, we carry out proto-object partitioning by employing the mean shift procedure [78]. We select the mean shift procedure as it is one of the most popular methods to appropriately partition an image preserving the boundaries of objects in it. This proto-object partitioning, which considers homogeneity by taking into account similarity in color (RGB vector) between different image locations, makes our proposed approach local structure-aware. Next, operator masks adapted to the local structure in the neighborhood of pixels are generated based on the proto-object partitions and a predefined set of scale parameters. By using the corresponding generated masks around each image pixel, center-surround operations are executed at multiple scales to generate maps of salient regions. Then, we take into account all the individual maps produced by considering all the scales and generate a combined map that corresponds to salient regions in an image by normalized addition of the maps. Finally, we detect multiple salient regions where object boundaries are properly preserved. These are subsequently used in object localization for robotic manipulation.

3.1 | Operator mask generation for detection of multiple salient objects

In this section, we elaborate on the generation method of our local image structure-aware center-surround operator masks, which is application-oriented and is based on color similarity and spatial proximity analysis. First, multiple circular Gaussian functions with different standard deviations (σ) are generated. To make the ensuing process application-oriented, these standard deviations can be considered as one-third of the expected object scales (diameter/side length) in a given task. Such a consideration allows better quantification of the distinctness of the objects as a whole (3σ) against their backgrounds. If the object scales are not known, a predefined set of scales can be used. Further, a couple of standard deviations, one less than a single pixel width and the other more than one-third of the largest image dimension should be used. These will allow the capture of local pixel-level details and a global image gist during multiple salient region detection. On the other hand, for a generic process, the multiple standard deviations can be taken as multiples /factors of each other as considered in [3].

We generate a mask from a circular Gaussian function (application-oriented) in the following manner [27]:

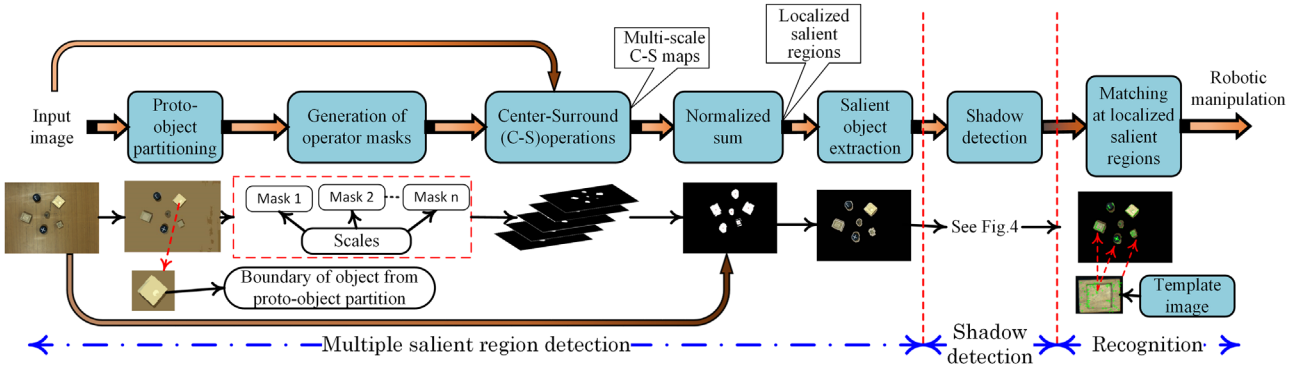

FIGURE 2 Proposed multiple salient region detection approach used before shadow detection [27]

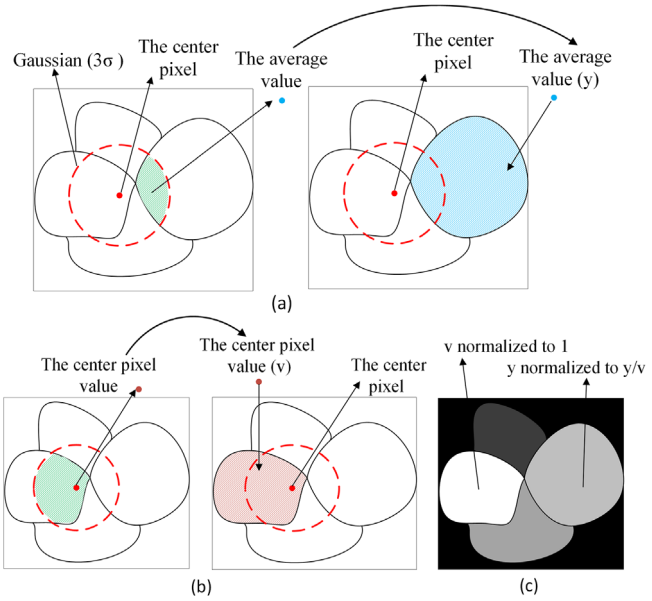

FIGURE 3 (a) Mask value assignment to any proto-object partitions except the center partition (partition having the center of the Gaussian function). The average (y) of all pixel values from the intersection of a partition and the 3σ extent of the Gaussian function is computed. That computed value y is allotted as the mask value to all the in pixels in that partition. (b) Mask value assignment to the center partition. The center value (v) of the Gaussian function lying on a pixel inside a partition (center partition) having intersection with the 3σ extent of the Gaussian function is considered. That value v is allotted as the mask value to all the in pixels in that partition. (c) The obtained mask with values $\in$ [0,1]. The obtained mask values for the partitions having intersections with the 3σ extent of the Gaussian function are normalized by v to get the mask [27]

- Let us consider a particular pixel of the image, referred to as the center pixel, around which such a circular Gaussian function is centered (refer Figure 3a). Now, take into account all the pixels from the proto-object partitions lying within the range of 3σ of the Gaussian function. So we mathematically put $I(x,y)$ as the image, where $(x,y) \in L$ is the set having all the image pixels, and $P_1, P_2, \ldots p, P_N$ as the N proto-object partitions such that:

$$P_i \cap P_j = \emptyset \quad \forall\, i,j \in \{1,2,\ldots,N\}, i \neq j \text{ and}$$

$$\bigcup_{j=1}^{N} P_j = L. \tag{1}$$

Consider the center pixel as (x_c,y_c). A 2D circular Gaussian function with a standard deviation σ and centered at (x_c,y_c) is given by

$$G(x,y) = \frac{1}{2\pi\sigma^2} e^{\frac{-\left((x-x_c)^2+(y-y_c)^2\right)}{2\sigma^2}}, \text{ where } (x,y) \in L. \tag{2}$$

The set containing pixels (as elements of the partitions) occurring within the range of 3σ of the Gaussian function are:

$$P(x_c,y_c) = \{P_i, P_i \cap E(x_c,y_c) \neq \emptyset\} \forall i \in \{1,2,\ldots,N\}, \tag{3}$$

$$\text{where } E(x_c,y_c) = \{(x,y), \forall (x,y) \in L \text{ and } (x-x_c)^2 + (y-y_c)^2 \leq 9\sigma^2\}.$$

- To obtain the mask, consider such a partition having intersection (within the range of 3σ) with the Gaussian function (Refer Figure 3a). The set having pixels in the intersecting portion of the Gaussian function centered at (x_c,y_c) and a proto-object partition P_i is

$$A_i(x_c,y_c) = \{(x,y), (x,y) \in \{P_i \cap E(x_c,y_c)\}\} \forall\, i. \tag{4}$$

The mean value computed from the intersecting portion of the Gaussian function is allotted to all the pixels in the partition. The average value computed is

$$V_{avg(x_c,y_c)}(i) = \frac{1}{|A_i(x_c,y_c)|} \sum_{\forall (x,y) \in A_i(x_c,y_c)} G(x,y). \tag{5}$$

The same (average) value allotment to every pixel in the partition is in line with the well-accepted understanding that pixels present in a partition may not be differentiated from each other. Then, the mask value assigned is:

$$Mask_{i,(x_c,y_c)}(x,y) = V_{avg(x_c,y_c)}(i), \forall (x,y) \in P_i,$$

$$\forall P_i \in P(x_c, y_c) \text{ and } i \neq i0. \quad (6)$$

However, if the center pixel is present in such a partition (center partition), all the pixels in that center partition are assigned the Gaussian function's value at the center pixel [Refer Figure 3b]. Assuming the center partition to be P_{i0}, that is, $P_{i0} \in P(x_c, y_c)$ and $(x_c, y_c) \in P_{i0}$, the mask value assigned is:

$$Mask_{i0,(x_c,y_c)}(x,y) = G(x_c, y_c), \forall (x,y) \in P_{i0}. \quad (7)$$

Here, the value at the center pixel is allotted instead of the average to create a greater difference in values between the neighboring partitions and the center one, resulting in a vivid local structure representation. For a partition having no intersection with the Gaussian function, a zero mask value is assigned to that partition.

$$Mask_{i,(x_c,y_c)}(x,y) = 0, \ \forall (x,y) \in P_i, \ \forall P_i \notin P(x_c, y_c). \quad (8)$$

- An example mask obtained from the above-mentioned procedure is represented in Figure 3c. The mask obtained as illustrated in Figure 3c is by a procedure that emphasizes color indiscernibility /similarity among the elements in a partition, without considering a partition's spatial extent. However, to determine the informativeness or saliency of a pixel, the spatial distance of the pixel from the center pixel is also important [34]. Larger the spatial spread of a neighboring partition farther would be the most distant pixels in it from the center pixel. Thus, a smaller weight is multiplied to a neighboring partition's mask value when the spatial extent of that partition is farther from the center pixel. Hence, this inverse relation for spatial proximity analysis is implemented as:
Let for a pixel (x, y), the rightmost and leftmost values of a partition P_i in terms of x and y be $x_r^{P_i}, x_l^{P_i}, y_r^{P_i}, y_l^{P_i}$.
With respect to the center pixel, let a partition's extent measure be

$$D_i = 1 - \frac{T_i}{(M+N)} \ \forall \ i \ \& \ i \neq \ i0, \quad (9)$$

$$\text{where } T_i = \max[\left|x_c - x_l^{P_i}\right|, \left|x_c - x_r^{P_i}\right|] +$$

$$\max[\left|y_c - y_l^{P_i}\right|, \left|y_c - y_r^{P_i}\right|], \ \forall \ i \text{ and } i \neq i0.$$

Here, M+N (the maximum possible extent) is considered for normalization to ensure that $D_i, \forall i$ lie in [0,1]. The weighted value of the mask is:

$$Mask_{i,(x_c,y_c)}(x,y) = V_{avg}(i) \times D_i,$$

$$\forall (x,y) \in P_i, \ \forall \ P_i \in P(x_c, y_c) \text{ and } \ i \neq i0. \quad (10)$$

Combining (6), (8) and (10), the mask obtained at (x_c, y_c) for a Gaussian function having a σ is denoted as:

$$Mask^{\sigma}_{(x_c,y_c)}(x,y) = Mask_{i,(x_c,y_c)}(x,y),$$

$$\forall (x,y) \in P_i \ \forall \ i \in \{1, 2, \ldots, N\}. \quad (11)$$

Finally, considering multiple masks obtained using Gaussian functions with different standard deviations, at location (x_c, y_c), the center-surround operator mask is obtained as follows:

$$CSMask^{\sigma_k,\sigma_t}_{(x_c,y_c)}(x,y) = \frac{Mask^{\sigma_k}_{(x_c,y_c)}(x,y)}{\sum_{(x,y)\in L} Mask^{\sigma_k}_{(x_c,y_c)}(x,y)} -$$

$$\frac{Mask^{\sigma_t}_{(x_c,y_c)}(x,y)}{\sum_{(x,y)\in L} Mask^{\sigma_t}_{(x_c,y_c)}(x,y)}, \ \sigma_t > \sigma_k. \quad (12)$$

At (x_c, y_c) location, the obtained saliency is:

$$S_{\sigma_k,\sigma_t}(x_c, y_c) = \left| \sum_{x=1}^{M} \sum_{y=1}^{N} CSMask^{\sigma_k,\sigma_t}_{(x_c,y_c)}(x,y) \times I(x,y) \right|. \quad (13)$$

Now, considering the saliencies calculated for all the different values of the pair (σ_k, σ_t) (all scales), we compute the region saliency at all image pixels as

$$S(x_c, y_c) = \sum_{\sigma_k,\sigma_t} \frac{S_{\sigma_k,\sigma_t}(x_c, y_c)}{\max_{(x_c,y_c)} S_{\sigma_k,\sigma_t}(x_c, y_c)}. \quad (14)$$

The above computed saliency can be referred to as the spatio-range saliency, as it is based on both spatial analysis and color (range) similarity. In (14), we see there is a normalization operation applied to the saliency computed at each scale. This normalization approach ensures that the computed saliency at a pixel is relative to the saliencies at all the other pixels. This normalization also makes the saliency quantities being added in (14) to be in the same range [0, 1], making the addition across scales suitable.
An image pixel having a spatio-range saliency value S>0 is taken as salient, and therefore, forms a part of the salient region. Thus multiple salient region detection is achieved. We are able to work with the trivial threshold operation of S>0 because our locally adaptive center-surround mask precisely and uniformly highlights only multiple object regions preserving their boundaries. Some noteworthy characteristics of our operator mask generation to get multiple salient objects are:

- Mask values at pixels in a partition are identical, and hence, such values in partitions capture the structure of local homogeneous regions, making our masks adaptive to the image's local structure.
- The Gaussian functions taken to generate the masks can be related to the approximate scales of the objects which should be detected as salient. Thus the mask can be adapted such that it is application-oriented ensuring appropriate performance in specific multiple objects scenarios.

- The proto-object partitioning of an image results in precise boundaries of detected salient regions aiding accurate object localization. As the results in Figure 8 show later, availability of precise boundaries after salient region detection as good as given by our proposed approach has not been achieved before.

Comprehensive quantitative and qualitative evaluation of our multiple salient region detection approach is presented in Section 6.1, which is compared to the state-of-the-art in salient region detection. In that while applying our approach on images from the standard databases and our industrial object image set, we do not consider application-oriented masks to uphold fairness.

We also demonstrate the applicability of our multiple salient region detection approach as a module in a robotic manipulation system performing pick and place operation through an example, where we consider application-oriented masks with known object scales. To further improve the detection accuracy during the robotic manipulation, we consider the detection and separation of shadow partitions from the detected regions which is described in the next section.

4 | AUTONOMOUS ROBOTIC MANIPULATION: A NEW SHADOW DETECTION METHOD

This is the first among the two sections that describes our autonomous robotic manipulation system driven by the multiple salient region detection approach proposed in Section 3. Here, we describe our novel fast shadow detection method, which follows after the multiple salient region detection module (see Figure 1). Salient regions detected by our local structure-aware approach may contain partitions of object shadow on the background around objects, which tend to act as object partitions. The presence of such shadow partitions may introduce errors during object localization. Thus, apart from accurate boundary detection through proto-object partitioning, shadow detection is also required to ensure proper object localization.

We draw our motivation from the vast studies available related to shadow detection. According to literature, [69, 79], chromaticity based methods are the fastest among all, which motivated us to propose a simple hue-based shadow detection technique. Chromaticity based methods considering pixel-level comparisons are prone to noise [80]. Hence, we consider a comparison of hue distributions of the regions (proto-object partitions) for the same.

4.1 | Proposed method of shadow detection

We propose a hue histogram-based approach to detect shadow partitions within salient regions obtained from our multiple salient region detection method. Our shadow detection approach is obviously applied to the object, shadow, and background partitions that may be present only within the salient regions (not the entire image). Object and background partitions usually have different hue content, which can be used to separate them.

The separation of the shadow partitions from the object partitions using our novel approach is depicted in Figure 4. Object partitions are expected to have similar hue content. A shadow partition would stand out from them, as the shadow of the object falling on the background is most likely to have hue content (but not intensity value) similar to the background. Note that, if a shadow (of something else) falls on an object, then the relevant partition is most likely to have hue content similar to that object. Therefore, this partition would not be separated, and rightly so, as it is indeed a partition within the object. Once we separate the shadow partitions within a salient region from the object partitions, only the latter shall be used to get an improved estimation of object location. Note that, this does not require shadow removal, but only separation of the partitions.

In our approach, first, image parts corresponding to each detected salient region are extracted from the input RGB image. These image parts may contain objects only, objects along with their shadows, or very rarely, objects, their shadows, and background. RGB to HSV conversion is performed in these image parts to decouple chromaticity components from intensity (value). For each partition in such an image part, a histogram of the hue component (bins: $[0°, 1°, \ldots, 359°]/360°$) is obtained as shown in Figure 5. The similarity between each pair of partitions is obtained by taking Bhattacharyya distance [81] between their hue distributions. We choose Bhattacharyya distance as it is symmetric and has been popularly used [82]. The Bhattacharyya distance D_B between two hue distributions p, q over same domain X is:

$$D_B(p,q) = -\ln B,\ B = \sum_{x \in X} \sqrt{p(x)q(x)}, \tag{15}$$

where B is the Bhattacharyya coeficient.

A fully connected undirected graph $\mathfrak{G} = (\mathcal{N}, \mathcal{E})$ is constructed by taking each partition in a salient region as a node $\mathcal{N}_k$ after getting the distance D_B between each pair (k, l) of partitions. The distance represents the weight at the edge $\mathcal{E}_{kl}$ between a pair of nodes. Then, graph partitioning using the normalized cut method [83] is applied to divide all the partitions present in a detected salient region into two sets: one is the object set containing object partitions only, and another one is the non-object set containing shadow and background partitions (as the hue of object's shadow and background is similar). Among the two obtained sets, the one where the average saliency value of the pixels is higher is considered the object set. For the object set, the average saliency value would be higher as compared to the shadow and background set mainly due to our spatial proximity analysis.

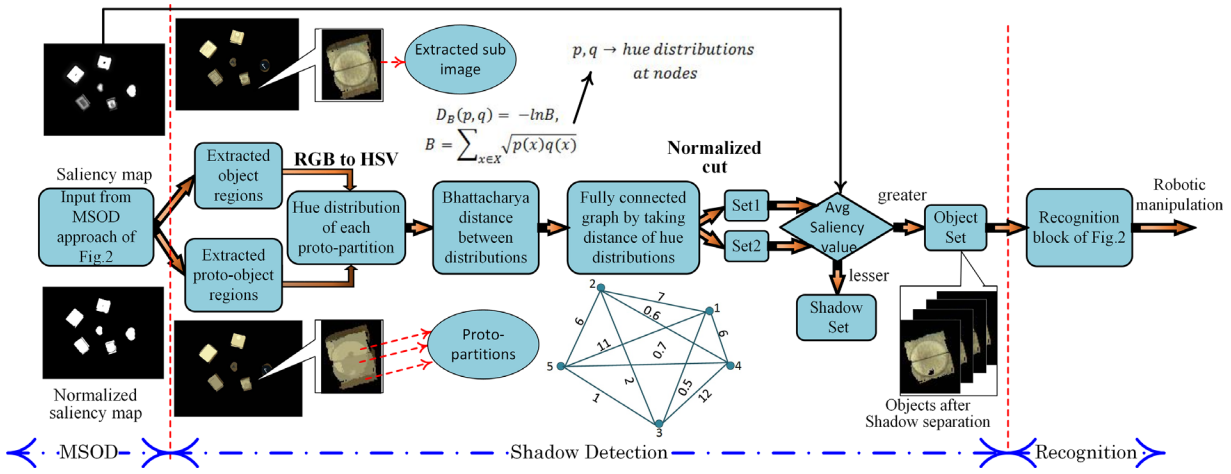

FIGURE 4 Proposed approach for detecting shadow partitions from salient regions

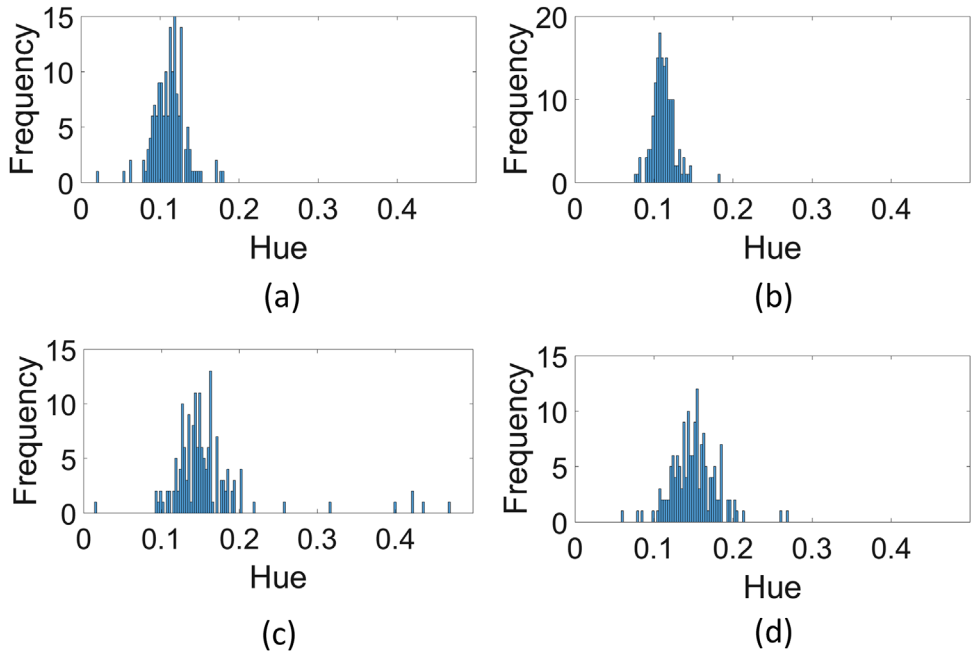

FIGURE 5 An example of hue histograms of (a) detected shadow partition. (b) Background partition. (c) and (d) Detected object partitions

5 | AUTONOMOUS ROBOTIC MANIPULATION: OTHER OPERATIONS

This is the second among the two sections that describes our autonomous robotic manipulation system. Here, we describe the other operations performed based on techniques existing in literature after the proposed multiple salient object detection and shadow detection approaches (see Figure 1).

Our proposed system (salient region and shadow detection) extracts a set of salient regions corresponding to multiple objects present in a scene, preserving their accurate boundaries, making it suitable for use in autonomous robotic manipulation. Although our system is applicable for a wide variety of robotic manipulation operations, to demonstrate such use of our approach, we consider the application of picking and placing industrial components by a robot. For automatic robotic pick and place, the orientation and location of all objects must be obtained in addition to their recognition after successful detection. After salient region detection, our system can be integrated with any widely used recognition techniques. We apply two well-known matching based recognition techniques, namely, shape and feature point matching. Here, the matching techniques are used since we have prior knowledge of all the objects to be manipulated in an application. Template matching has been used widely for recognition [63] and to obtain orientation details. But, as stated earlier, applying matching over the whole image for recognition is computationally expensive. In our system, matching is performed only at the regions detected as salient using our approach, which is designed to detect all objects that shall be manipulated as important/salient.

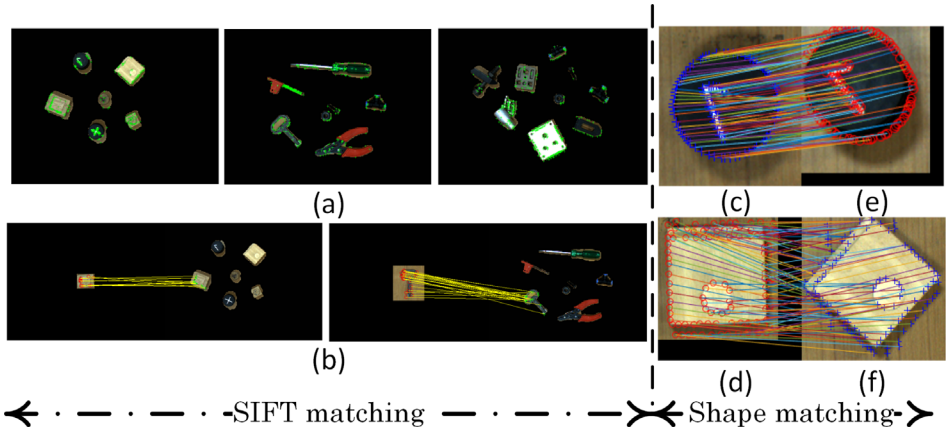

FIGURE 6 (a) Extraction of SIFT key points at the regions detected as salient, (b) Matches of SIFT key points between an image template and an object detected as salient, (c-f) Matching of the shape context between templates of objects (c) and (d), and the obtained salient regions (e) and (f) at optimum matching cost. The estimated angles at optimum matching cost are 7° and 44° respectively [27]

Our system, makes recognition more efficient/faster despite the added "burden" of saliency detection. Additionally, after separating shadow, performing matching for recognition only at the regions detected as salient increases the rate of recognition by reducing false positives in background and shadow regions. Upon successful detection and recognition of multiple salient objects, the robotic manipulator can proceed to manipulate the objects to perform specific tasks.

5.1 | Recognition of objects by feature point matching

As one of the two ways of object recognition, we use scale-invariant feature transform (SIFT) for the recognition of objects to be manipulated by a robot. Many state-of-the-art methods [2, 29] have used SIFT-based object recognition for robotic manipulation. In their methods, after extracting feature points along with their feature descriptors from an input image, nearest neighbor matching is done with that of the feature descriptors of stored images to find the best match. As this feature matching is done across the whole image, the possibility of false positive matches arising due to background feature points increases. Contrary to that, in our proposed method there is the least possibility of such false positive matches as the extraction of the feature point, and the matching is performed only at regions detected as salient (see Figure 6a). Further improvement in initial matching is achieved by Hough transform clustering [84] (see Figure 6b) and maxima among the feature points in the Hough space is considered for object pose estimation using RANSAC homography [85].

5.2 | Recognition of objects by shape matching

We considered the shape matching of [86] as the second method of template matching in our system. As our proposed system can preserve proper boundaries (shape) of multiple detected salient regions, shape information can be exploited for recognition of the objects. In this method of matching, a log polar histogram is measured by considering all the extracted sampled edge points' coordinates with reference to a specific edge point, which is known as origin. This log polar histogram is denoted as the shape context of the considered origin point. After taking the shape contexts of detected salient objects and the template, the optimum matching cost which is the smallest chi-square test statistic between the obtained shape contexts is considered for the matching.

All edge points X in a detected salient region are rotated in-plane by a same angle θ using the 2D rotation matrix.

The best transformation parameter $\hat{\theta}$ between the edge points of the template and the object in the detected region is given by

$$\hat{\theta} = \arg\min_{\theta} \mathcal{TC}(U, F(X;\theta)), \quad 0° \leq \theta < 360°, \tag{16}$$

where U is the set of all sampled edge points of the template image, $\mathcal{TC}$ represents the total cost of shape context matching between sampled edge points of a template image and a detected region, and $F(X;\theta)$ is obtained using the 2D rotation matrix. Figure 6c–f depicts the matching of candidate points between a template and a region detected as salient corresponding to minimum matching cost.

5.3 | Automated mechanical pick and place

During the aforesaid recognition of an object through template matching, the object's orientation, as an inherent part of matching, and the object's location, as the centroid of the recognized object within the salient region, are obtained. After detection, recognition, and localization of all objects, which also yields their orientations and locations, the height of each object is extracted from the already available computer-aided design

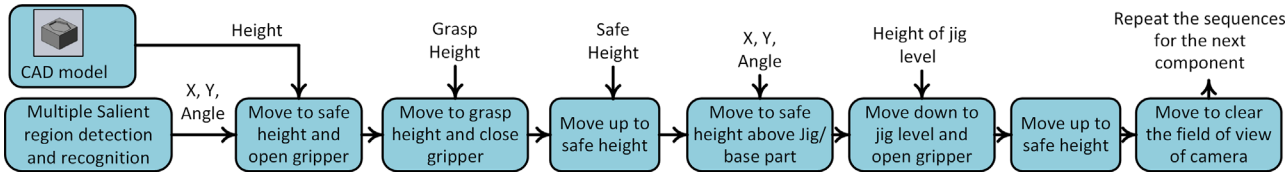

FIGURE 7 Flowchart of the strategy for robotic pick and place

(CAD) model. Now robotic manipulators can pick up objects one by one from the obtained locations at the estimated orientation. As in the case of any automated operation, our robotic manipulator does the various operations through a well-defined strategy as shown in Figure 7, which especially helps to ensure a collision free pick and place operation or motion planning. The robotic manipulator follows a pre-defined motion sequence for proper handling of each part without collision [87]. The manipulator moves from the initial/ home position to the estimated location of the objects to be picked up. However, it maintains a safe height while in motion to avoid a collision. Then, it moves down to the grasp height of the object, then moves up to the safe height after grasping which is followed by a movement towards the jig/ base part (Figure 12a). All these manipulator movements are controlled using a robot controller interface called digimatrix operated through NI LabVIEW graphical programming environment.

6 | RESULTS AND DISCUSSION

In this section, the evaluation of our proposed robotic manipulation system is presented in three folds. Section 6.1 presents the evaluation of our multiple salient region detection algorithm and its effectiveness through comparisons with the state-of-the-art salient object detection approaches. This is followed by the proposed shadow detection algorithm's evaluation in Section 6.2.1, and in Section 6.2.2, we demonstrate the efficiency and capability of our entire system for vision guided robotic pick and place operation in terms of localization, angular error, detection and recognition rates, and processing speed.

6.1 | Evaluation of multiple salient region detection

We evaluate our multiple salient region detection method by using images with multiple salient objects. As this evaluation is regarding generic salient region detection, just like [3], we consider a fixed set of standard deviations (object scales are not known) of the multiple Gaussian functions involved in our approach, which are $\{3, 5, 7\}$ pixel widths for the center Gaussian function and $\{19, 23, 29\}$ pixel widths for the surround Gaussian function. These parameters are stated for the image size 256×256 and are to be scaled linearly for larger or smaller sizes of images. An analysis of the sensitivity of these parameters is given in Appendix A.3. For the evaluation, we choose the SED2 database [88], GIT database [89], and SalMoN database [90] as these standard databases predominantly contain images with more than one salient object. Along with this, we also consider images of the YCB benchmark [91] and on our collection of images, which are related to robotic manipulation containing multiple objects. For comparison, several existing salient object detection algorithms like HS [34], FT [92], MB [42], SEG [93], LPS [94], SMD [95], MIL [96], NLDF[47], BASNet[6], CPD[48], SCR[97], SPD[98], LDF[99], MSFNet[53] and SCWSSOD[54] with publicly available implementation codes producing state-of-the-art results are considered. Among these, NLDF, BASNet, CPD, SCR, SPD and LDF, MSFNet and SCWSSOD are latest deep learning-based models. The deep learning-based algorithms are implemented by using GPU NVIDIA 2080 Ti (11GB) in a machine with a RAM of 64 GB. The other methods not involving deep learning are implemented in a machine with an Intel i5 processor (CPU) with 8 GB RAM. Along with the qualitative and quantitative evaluation, the computation time of different methods is also considered. For the deep learning-based methods, the pre-trained models provided by the corresponding authors have been used.

6.1.1 | Qualitative evaluation and discussion

The salient regions detected by the different methods in a few images from the standard databases and of industrial objects are shown in Figures 8 and 9, respectively. Images of Figure 8 are from the SED2 database (first five rows), GIT database (sixth and seventh row), and SalMoN database (the rest). The last four images of Figure 9 are from the YCB benchmark and the first five images of Figure 9 are that of our collection of industrial objects. As can be seen, our method consistently detects multiple salient regions with accurate boundaries (Figures 8 and 9), and generates accurate salient region maps even when very small objects are present (third and fourth images of Figure 9) and objects have low contrast with the background (first two images of Figure 9). No other approach, including the ones based on deep learning, is as consistent as ours in collectively detecting multiple salient image regions with accurate boundaries. To check the robustness of the proposed approach in terms of multiple salient object detection in images with low illumination, with objects having similar shape, size, and color, and with objects viewed from an oblique angle, results on such images are shown in Figure A.3 in Appendix A.2.

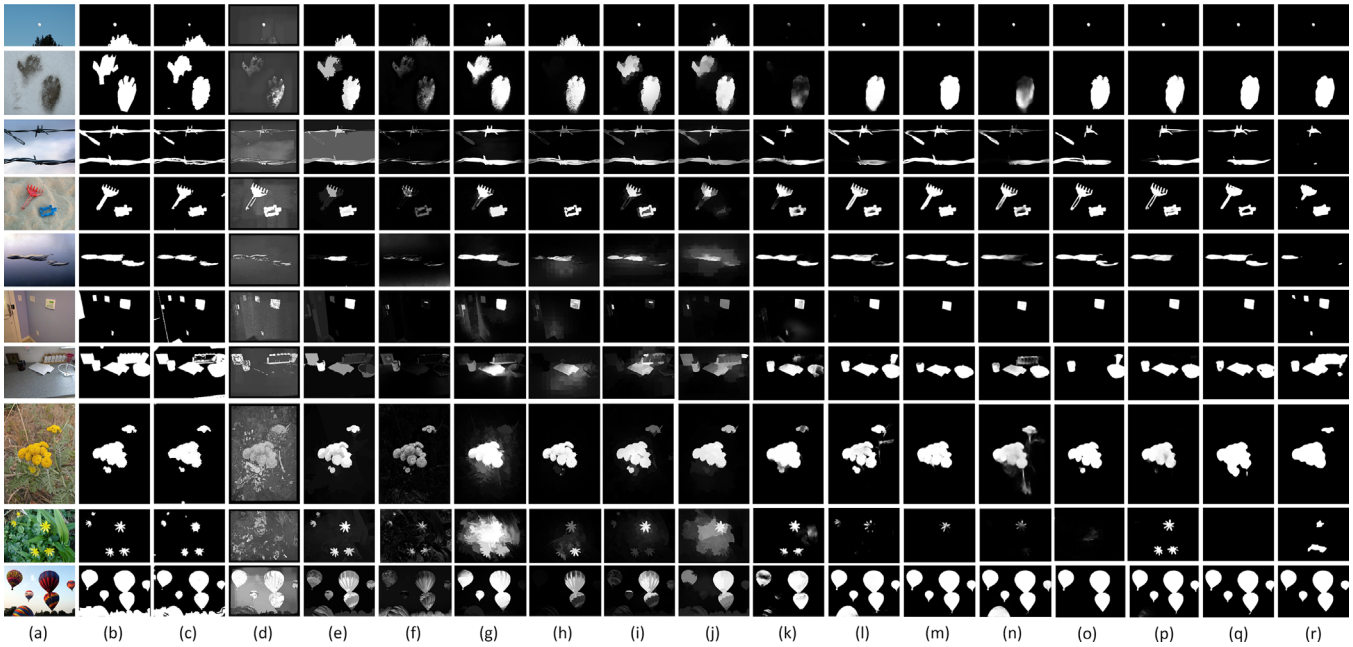

FIGURE 8 Visual comparison of detected salient regions obtained by different state-of-the-art algorithms on images from SED2 database [88] (First five rows), GIT database [89] (6th and 7th row) and SalMoN database [90] (the rest). (a) Input images; (b) ground truth; (c) our method; (d) SEG; (e) HS; (f) FT; (g) MB; (h) LPS; (i) SMD; (j) MIL; (k) NLDF; (l) SPD; (m) CPD; (n) SCR; (o) BASNet; (p) LDF; (q) MSFNet; (r) SCWSSOD

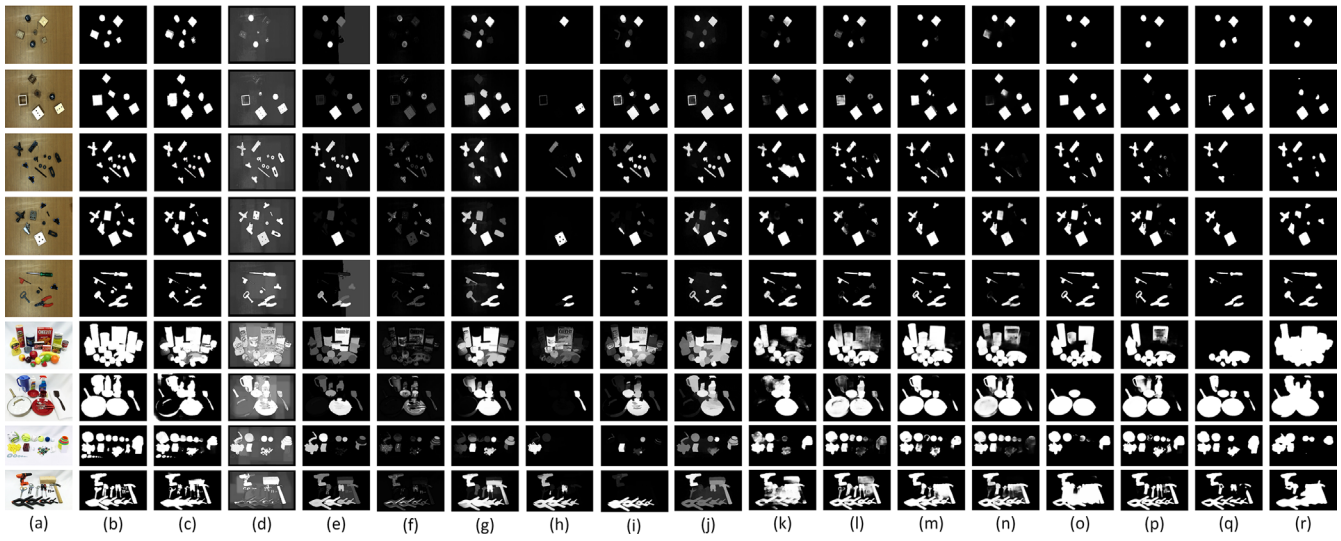

FIGURE 9 Visual comparison of salient regions obtained by different state-of-the-art algorithms. (a) Input images; (b) ground truth; (c) our method; (d) SEG; (e) HS; (f) FT; (g) MB; (h) LPS; (i) SMD (j) MIL (k) NLDF (l) SPD (m) CPD (n) SCR (o) BASNet (p) LDF (q) MSFNet (r) SCWSSOD. First five images with multiple objects are our collection of industrial objects taken from our robotic workspace and rest four are from YCB benchmark database[91]

6.1.2 | Quantitative evaluation and discussion

We evaluate the performance of different salient object detection algorithms utilizing four evaluation metrics, F-measure, mean absolute error (MAE), shuffled area under receiver operating characteristics (ROC) curve (sAUC), and shuffled F-measure. F-measure is a quantitative measure that is often used to evaluate the agreement between the detected regions as binary maps and binary ground truth [23]. The salient object detection ground truth is already binary and we binarize a computed salient object map using the image dependent adaptive thresholding proposed in [92]. By using the binary salient object map thus obtained against corresponding binary ground truth, we calculate precision and recall values, which are used to get a F-measure as:

$$F = \frac{2(\text{Precision})(\text{Recall})}{\text{Precision} + \text{Recall}}. \tag{17}$$

The shuffling process suggested in [100], which ensures that a better performance is not due to any prior bias, is applied to get

the shuffled F-measure and shuffled AUC. In the shuffling process, for an image, its binary salient object ground truth is taken as the positive reference and a negative reference is formed by taking the union of the ground truths of a set of other images. For sAUC computation, multiple thresholds are used to generate multiple binary maps from an algorithm generated saliency map. Comparing such a binary salient object map obtained for an image to its positive reference, the true positive rate TPR is obtained as

$$TPR = \frac{TP}{P}, \tag{18}$$

where TP is the true positive and $P = TP + FN$, with FN as false negative, represent the number of salient pixels in the positive reference. Then, comparing the binary salient object map to the negative reference, the false positive rate FPR is obtained as

$$FPR = \frac{TP_n}{P_n}, \tag{19}$$

where P_n is the number of salient pixels in the negative reference and TP_n is the true positive for the negative reference. sAUC is then obtained from the ROC calculated using the TPR and FPR values for the multiple binary salient object maps. For shuffled F-measure computation, a threshold, as suggested in [92], is used to compute a single binary salient object map. Using this map against the positive and the negative references, we compute $Recall = TPR = TP/(TP + FN)$ and $Precision = TP/(TP + FP)$, where $FP = FPR \times P_n$, from the quantities in (18) and (19).

Further, in the computation of F-measure or shuffled F-measure, true negatives, which quantifies pixels correctly marked as non-salient, are not considered. As pointed in [23], they may favor methods that successfully assign high saliency to salient regions but fail to reject non-salient regions over methods that do well in both but comparatively do not assign high saliency value to salient regions. Taking this into account, we also consider the mean absolute error (MAE) between the obtained salient object map (S) normalized to [0,1] and the binary ground truth (GT) for evaluation. The mean absolute error is given by:

$$MAE = \frac{1}{W \times H} \sum_{x=1}^{W} \sum_{y=1}^{H} | S(x,y) - GT(x,y) |, \tag{20}$$

where W and H are height and width of the salient object map and binary ground truth images. Table 1 shows the shuffled F-measure, sAUC, F-measure and MAE averaged over all images in a database for all compared methods considering the standard SED2, GIT, SalMoN, YCB benchmark, and our industrial object image databases. As can be seen, among the 20 different cases (database and measure combinations), our approach gives the best results in 9, which is better than any other. The approaches LDF and SPD, which are the latest deep learning-based methods, come distant second by producing the best results in 3 cases. The proposed method ranks within the top three in almost all the cases and it outperforms all the others in terms of all the evaluation metrics on our industrial object image set. It should be noted that while the deep learning-based approaches such as SPD, CPD, LDF, BASNet and MSFNet do well for a database or two, they do not outperform the proposed approach's overall performance across the 5 databases. The superior performance of our multiple salient region detection approach is seen considering databases with all images containing multiple salient objects and databases predominantly having images with single salient objects are not considered. Unlike many approaches in the literature, our approach not only detects objects with high saliency but also objects with moderate or lower degree of saliency in the given image at hand. Our proposed approach is designed to detect multiple objects, and the design does not depend on the number of objects present in an image. On the other hand, it is difficult to obtain databases (with ground truth) to train deep learning-based approaches that have equal representation of images with different numbers of objects. This may create a class-imbalance kind of an issue in the learning that may hinder the performance of deep learning-based approaches for multiple salient region detection. The performance of the learned deep models may not scale well with change (increase) in the number of salient objects in images. This is evident from the Figure 8 (first, second, sixth, and ninth row images), where only a single object is detected by most of the deep learning-based methods. Similarly, for our industrial objects and YCB object database as shown in Figure 9 and A.3, all objects are not detected as salient by the deep models.

Further, we compare the performance of the proposed multiple salient region detection approach with recent generic object detection techniques, YOLO and mask R-CNN [101–104], which can detect all objects in an image. The qualitative and quantitative results are shown in Figure 10 and in Table 3. For YOLO, bounding boxes show the objects detected, and when no object is detected, the output is shown as a black image in Figure 10. Generic object detectors try to detect every object irrespective of saliency. Although generic object detectors are good at detecting multiple objects compared to many state-of-the-art salient object detection approaches, their performance in saliency detection is expected to be low as demonstrated in Table 3.

Processing speed: The average processing time per pixel (in milliseconds) taken to detect salient regions by the different approaches are given in Table 2, in which only our approach is explicitly targeted towards detecting multiple salient regions. The number of parameters to be tuned and the simulation platform for the different approaches are also presented in Table 2. As can be seen, the different salient object detection approaches are implemented in different platforms using different tools, and hence, most of them are not comparable one-to-one. However, it is clearly evident that the proposed approach is not computationally expensive.

TABLE 1 Comparison of salient object detection methods using shuffled F-measure, sAUC, F-measure and MAE on SED2, GIT, SalMoN, YCB and our industrial object image databases containing images with multiple salient objects. The top three results for each of the 20 cases are shown in magenta (best), green (second best) and blue (third best)

Method classification →		**Methods not involving machine learning**						**Machine learning**		**Deep learning**							
Database	**Parameter**	**Our method**	**HS 2013**	**FT 2009**	**MB 2015**	**LPS 2015**	**SMD 2016**	**SEG 2010**	**MIL 2017**	**NLDF 2017**	**SPD 2019**	**CPD 2019**	**SCR 2019**	**BASNet 2019**	**LDF 2020**	**MSFNet 2021**	**SCWSSOD 2021**
SED2	shuffled F-measure	0.7505	0.6453	0.5623	0.6798	0.5515	0.6618	0.5065	0.6553	0.6247	0.6783	0.6753	0.7102	0.7472	0.7414	0.7489	0.7159
	sAUC	0.7778	0.7084	0.5707	0.7130	0.6697	0.7169	0.6410	0.7191	0.7083	0.7587	0.7437	0.7478	0.7764	0.7631	0.7819	0.7570
	F-measure	0.7987	0.6580	0.5928	0.7235	0.6658	0.7240	0.3929	0.7470	0.7428	0.7832	0.7546	0.7859	0.7998	0.7945	0.8256	0.7871
	MAE	0.0380	0.1594	0.2056	0.1385	0.1428	0.1334	0.3131	0.1314	0.1053	0.0876	0.0785	0.0860	0.0655	0.0688	0.0648	0.0685
GIT	shuffled F-measure	0.6088	0.4752	0.4120	0.5205	0.4053	0.4162	0.4154	0.4673	0.4625	0.5230	0.3102	0.2985	0.4854	0.4769	0.5212	0.5455
	sAUC	0.6836	0.5984	0.5115	0.6171	0.5743	0.5904	0.5999	0.5833	0.6097	0.6121	0.5330	0.5326	0.6471	0.6299	0.6824	0.7006
	F-measure	0.5380	0.4130	0.3570	0.4864	0.3955	0.4462	0.2738	0.4768	0.4823	0.5321	0.2554	0.2455	0.4975	0.5190	0.5043	0.4925
	MAE	0.1721	0.2998	0.2279	0.2439	0.2344	0.2340	0.3555	0.2407	0.2011	0.1851	0.2531	0.2648	0.2328	0.2013	0.2392	0.2231
SalMon	shuffled F-measure	0.8054	0.5645	0.4123	0.5585	0.4597	0.6036	0.5032	0.5545	0.7634	0.8171	0.7521	0.7665	0.7440	0.7956	0.7958	0.8011
	sAUC	0.8215	0.6856	0.5609	0.7451	0.6256	0.6930	0.7646	0.6918	0.8049	0.8414	0.8172	0.8113	0.8194	0.8283	0.8166	0.8108
	F-measure	0.7881	0.5808	0.4237	0.5779	0.4614	0.5899	0.4998	0.5773	0.7878	0.8303	0.7919	0.7856	0.7870	0.8206	0.7923	0.7960
	MAE	0.0449	0.1947	0.1627	0.1444	0.1367	0.1230	0.3059	0.1434	0.0471	0.0417	0.0419	0.0452	0.0473	0.0395	0.0415	0.0444
YCB	shuffled F-measure	0.6275	0.4206	0.1742	0.5077	0.2462	0.4026	0.2380	0.4537	0.5536	0.6060	0.6460	0.6112	0.5998	0.6368	0.5580	0.6232
	sAUC	0.6209	0.5642	0.5083	0.5811	0.5360	0.5575	0.5667	0.5612	0.5867	0.6140	0.6273	0.6193	0.6103	0.6339	0.6159	0.6132
	F-measure	0.8195	0.6658	0.6260	0.7260	0.4300	0.5782	0.4914	0.7413	0.6948	0.7857	0.8335	0.8008	0.7654	0.8377	0.7435	0.8131
	MAE	0.1257	0.2222	0.3009	0.1775	0.2784	0.2191	0.2982	0.2026	0.1946	0.1312	0.1070	0.1278	0.1536	0.1056	0.1567	0.1306
Our industrial object image	shuffled F-measure	0.7737	0.4821	0.1877	0.6765	0.3660	0.6082	0.7004	0.6702	0.5883	0.7085	0.6928	0.6712	0.7333	0.6072	0.6792	0.7207
	sAUC	0.7868	0.6251	0.5197	0.7328	0.5977	0.6942	0.7506	0.7257	0.6827	0.7494	0.7353	0.7278	0.7632	0.6905	0.7331	0.7525
	F-measure	0.8339	0.5293	0.8298	0.8266	0.4329	0.7565	0.7239	0.8333	0.7349	0.8312	0.7807	0.8276	0.8045	0.6433	0.7285	0.7658
	MAE	0.0350	0.1257	0.0996	0.0670	0.0887	0.0627	0.2350	0.0667	0.0662	0.0433	0.0423	0.0501	0.0415	0.0713	0.0475	0.0470

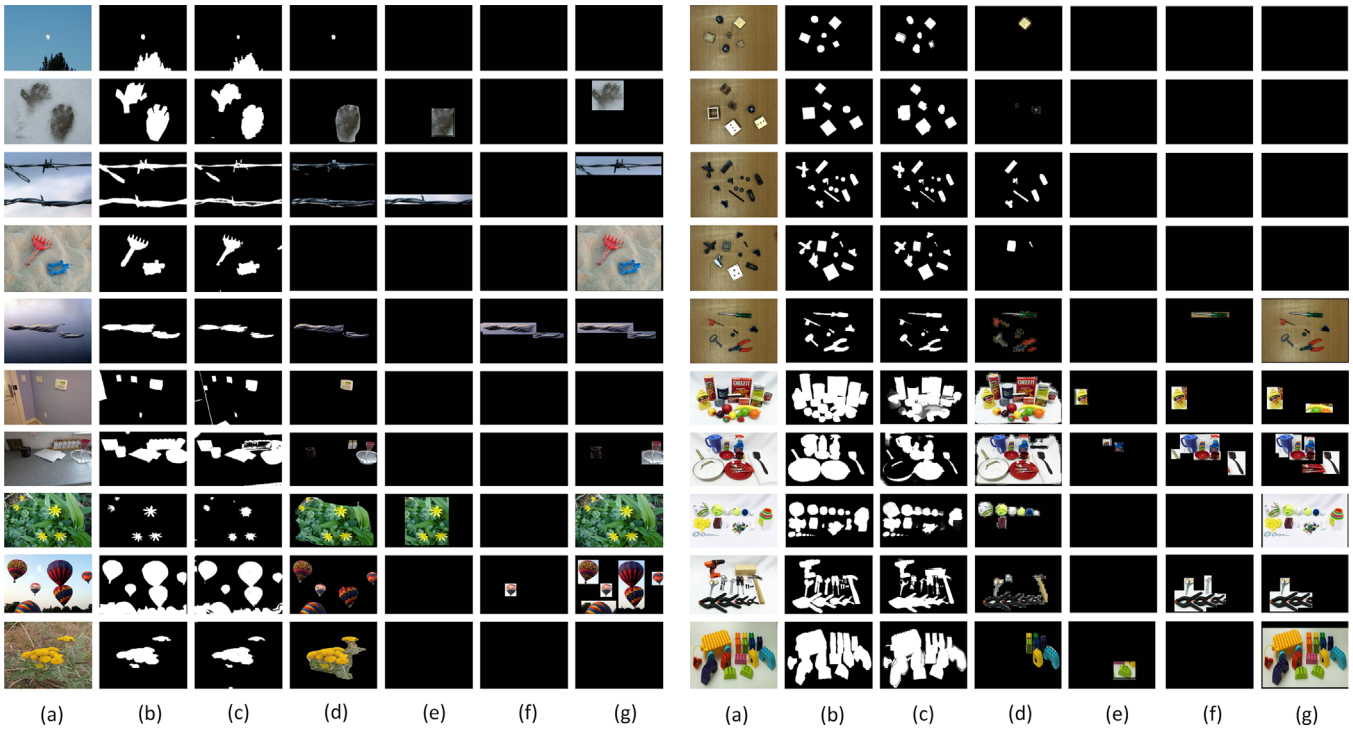

FIGURE 10 Visual comparison of detected salient regions obtained by different state-of-the-art algorithms target detection methods on images from different databases. (a) Input images; (b) ground truth; (c) our method; (d) mask R-CNN; (e) YOLO v2; (f) YOLO v3; (g) YOLO v4

TABLE 2 Computation time (millisecond/pixel) [considering multiple images] and number of parameters of the different salient region detection methods implemented in different platforms. Only testing times (GPU) are shown for the machine/deep learning-based methods

	Device	Ours	SEG	HS	FT	MB	LPS	SMD	MIL	NLDF	SPD	CPD	SCR	BASNet	LDF	MSFNet	SCWSSOD
Time (mSec)	CPU	0.07	0.193	0.013	0.0003	0.0005	0.005	0.006	0.375	-	-	-	-	-	-	-	-
	GPU	-	-	-	-	-	-	-	-	0.0004	0.0002	0.0002	0.0004	0.0003	0.0002	0.0003	0.0002
Number of parameters		<10	<10	<10	<10	<10	<20	<10	<15	4×10^7	7×10^7	3×10^7	8×10^7	2×10^7	8×10^7	3×10^7	6×10^7
Code		matlab	matlab	.exe	.exe	.exe	matlab	matlab	matlab	tensorflow	pytorch	pytorch	pytorch	pytorch	pytorch	pytorch	pytorch

6.2 | Evaluation considering automatic robotic manipulation

6.2.1 | Proposed shadow detection approach performance

Although the boundaries of the salient regions detected by our technique are updated using our proposed shadow detection method (see Section 4), we consider the proposed shadow detection approach separately here to evaluate it. We consider both qualitative and quantitative evaluations of the proposed shadow detection technique along with its utility in our case. Figures 11a and 11b show detected salient regions with shadow and regions obtained after shadow separation, respectively for objects to be manipulated by an industrial robot. The main objective of the proposed shadow detection technique is to improve localization accuracy. Thus, for quantitative evaluation, we consider a reduction in localization error (as a quantitative measure) after applying the proposed shadow detection technique to the regions detected as salient and the results are presented in Tables 4 and 7, and Figures 13 and 14, which we discuss later in the following Section 6.2.2.

6.2.2 | Automated robotic pick and place performance

Our saliency based robotic pick and place manipulation is demonstrated by a system comprising of a 6-axis industrial robot manipulator (Yaskawa Motoman MH5) which has a two finger pneumatic gripper, a robot control interface called Digimetrix, and a robot controller having open software architecture. The vision system is equipped with an overhead camera that is used to capture images of objects in the robotic workspace along with a computer with a data acquisition board. A 1.3 MP resolution Basler acA1300-22gc GigE camera

TABLE 3 Comparison of the proposed salient object detection method with advanced target detection methods using shuffled F-measure, sAUC, F-measure and MAE on SED2, GIT, SalMoN, YCB and our industrial object image databases containing images with multiple salient objects. The best result for each of the 20 cases are shown in magenta

Database	Parameter	Our method	mask R-CNN 2017	YOLO v2 2017	YOLO v3 2018	YOLO v4 2020
SED2	Shuffled F-measure	0.7505	0.4897	0.3762	0.3864	0.4197
	sAUC	0.7778	0.6595	0.5981	0.6105	0.6315
	F-measure	0.7987	0.5554	0.4800	0.3983	0.4211
	MAE	0.0380	0.1535	0.4179	1.3916	1.5517
GIT	Shuffled F-measure	0.6088	0.4887	0.1842	0.2716	0.3723
	sAUC	0.6836	0.5691	0.5280	0.4988	0.5564
	F-measure	0.5380	0.4376	0.4229	0.2149	0.3658
	MAE	0.1721	0.3121	7.2514	6.7277	7.4512
SalMon	Shuffled F-measure	0.8054	0.7031	0.4923	0.4629	0.6278
	sAUC	0.8215	0.7727	0.6535	0.5214	0.7203
	F-measure	0.7881	0.7203	0.4532	0.3834	0.4957
	MAE	0.0449	0.1105	8.5325	5.8812	5.5746
YCB	Shuffled F-measure	0.6275	0.5888	0.3475	0.3017	0.4905
	sAUC	0.6209	0.5700	0.5023	0.5224	0.5265
	F-measure	0.8195	0.7384	0.5720	0.4417	0.3718
	MAE	0.1257	0.2376	2.3209	2.2512	3.9795
Our industrial object image	Shuffled F-measure	0.7737	0.3980	0.1416	0.1524	0.2220
	sAUC	0.7868	0.6223	0.5131	0.5053	0.6118
	F-measure	0.8339	0.4758	0.3145	0.2115	0.2428
	MAE	0.0350	0.0765	8.2259	1.3754	8.8698

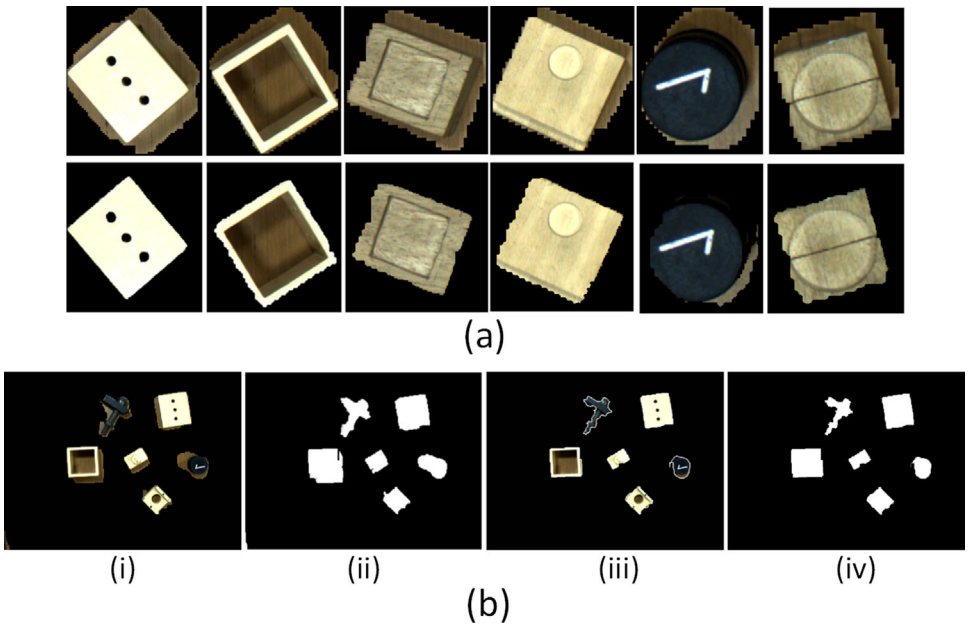

FIGURE 11 (a) First row: Objects present in regions detected salient, Second row: Objects after separating detected shadow partitions (b) (i) and (iii) Objects present in regions detected as salient before and after shadow detection, (ii) and (iv) Regions detected as salient before and after shadow detection

TABLE 4 Evaluation of detection performance using the proposed method for all set of objects from Figure 9 (images 2–5). Each image corresponds to an object set

	Salient region detection			Salient region + Shadow detection		
	Positional error (cm)			Positional error (cm)		
Objects from Figure 9	X	Y	Detection rate (%)	X	Y	Detection rate (%)
Set2	0.23±0.03	0.18±0.05	100	0.09±0.06	0.08±0.05	100
Set3	0.13±0.03	0.21±0.06	100	0.08±0.05	0.11±0.06	100
Set4	0.12±0.02	0.14±0.03	100	0.10±0.03	0.09±0.04	100
Set5	0.11±0.04	0.14±0.06	100	0.11±0.04	0.13±0.06	100

TABLE 5 Evaluation of recognition performance using the proposed method for all set of objects from Figure 9 (images 2-5). Each image corresponds to an object set [27]

	Recognition by SIFT		Recognition by Shape matching	
Objects from Figure 9	Angular error (%)	Recognition rate (%)	Angular error (%)	Recognition rate (%)
Set2	2.3±0.5	100	1.0±0.4	100
Set3	3.2±0.3	70	0.9±0.3	100
Set4	5.1±0.7	41	2.1±0.2	100
Set5	3.4±0.7	82	1.4±0.6	100

TABLE 6 Evaluation of performance of the conventional approach for all set of objects from Figure 9 (image 2-5). Each image corresponds to an object set. [NA-Not Applicable] [27]

	Detection and Recognition by SIFT					Detection and Recognition by Shape matching				
Objects from Figure 9	Positional error (cm) X	Positional error (cm) Y	Detection rate (%)	Angular error (deg)	Recognition rate (%)	Positional error (cm) X	Positional error (cm) Y	Detection rate (%)	Angular error (deg)	Recognition rate (%)
Set2	0.48±0.06	0.35±0.03	NA	4.8±0.3	100	0.19±0.06	0.21±0.08	NA	1.8±0.4	100
Set3	0.62±0.02	0.74±0.06	NA	6.4±0.5	41	0.21±0.02	0.27±0.06	NA	2.1±0.6	100
Set4	0.57±0.04	0.46±0.05	NA	3.7±0.7	60	0.19±0.08	0.16±0.03	NA	1.9±0.3	100
Set5	0.33±0.07	0.41±0.08	NA	5.4±0.7	70	0.14±0.03	0.31±0.03	NA	2.1±0.6	100

TABLE 7 Detailed evaluation of pick and place performance using the proposed methods for specific objects /parts (parts in image1 of Figure 9, Figure 12a). [NA-Not Applicable]

Proposed method (saliency +shadow+ recognition)									
	Saliency detection			Saliency + shadow detection			Shape matching		SIFT matching	
Assembly parts	Positional error (cm) X	Positional error (cm) Y	Detection rate (%)	Positional error (cm) X	Positional error (cm) Y	Detection rate (%)	Angular error (deg)	Recognition rate (%)	Angular error (deg)	Recognition rate (%)
Object1	0.21±0.05	0.18±0.045	100	0.06±0.02	0.04±0.03	100	1.2±0.5	100	3.2±0.4	100
Object2	0.16±0.035	0.21±0.05	100	0.09±0.03	0.07±0.02	100	0.9±0.3	100	2.1±0.6	100
Object3	0.15±0.03	0.16±0.035	100	0.09±0.04	0.11±0.04	100	0.6±0.2	100	3.6±0.2	100
Object4	0.14±0.049	0.23±0.055	100	0.10±0.05	0.12±0.06	100	0.8±0.4	100	2.5±0.5	100
Object5	0.24±0.06	0.21±0.052	100	0.08±0.03	0.09±0.05	100	1.1±0.3	100	1.9±0.7	100
Object6	0.16±0.035	0.19±0.04	100	0.12±0.04	0.15±0.06	100	1.2±0.6	100	2.2±0.5	100

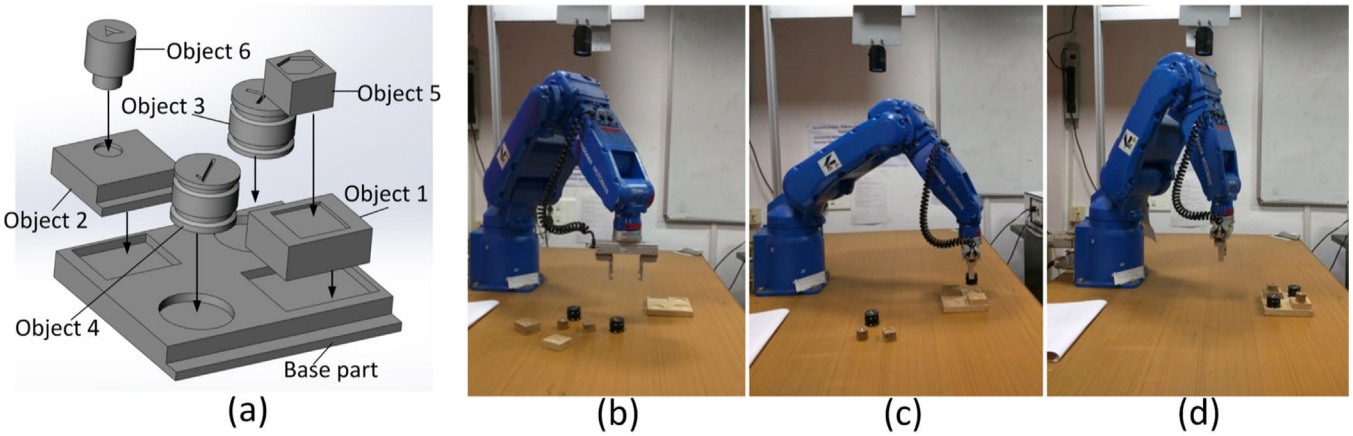

FIGURE 12 (a) Parts to be assembled. (b)–(d) Images of operation[27]. The demo videos can be found at https://github.com/sudiptabhuyan1/SAMSOD

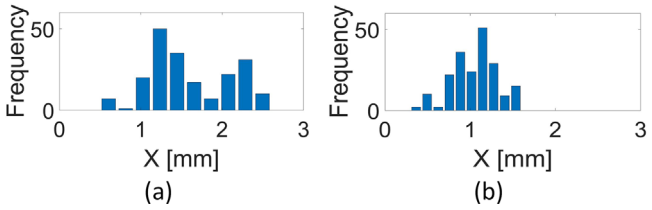

FIGURE 13 Histogram of positional error along X-axis (a) before shadow detection, and (b) after applying shadow detection at detected salient regions

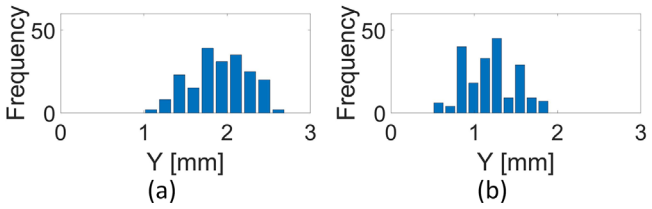

FIGURE 14 Histogram of positional error along Y-axis (a) before shadow detection, and (b) after applying shadow detection at detected salient regions

powered by Sony ICX445 CCD sensor is used along with a 6mm focal length Edmund Optics lens. The mapping between image coordinates and the robot's real world coordinates is done by calibrating the camera. All computations have been performed with Matlab R2016b running on an 8GB RAM, intel core i5 processor system clocked at 3.30GHz and the obtained parameters (X, Y, Z, angle) are exported to the robot control interface (Digimetrix) through LabVIEW. The image processed is of size 1078×958.

In our pick and place experiment, we consider six components as shown in Figure 12a. Here, we have considered regular sized objects as they can be easily manipulated by our two finger gripper. The robot performing various stages of pick and place operation based on our proposed method is demonstrated in Figures 12b, c, and d.

Some specifics of our approach applied to the pick and place operation are:

- The center-surround operations based detection of multiple salient objects considers Gaussian functions having five different standard deviations (σ), which are chosen according to the approximate object sizes.
- The order in which the objects are to be placed on the jig /base part decides the sequence of considering the different templates for recognition.
- A few parameters of SIFT algorithm [105] are modified to get sufficient key points after experimentation. The threshold value for accurate key point localization is taken as 0.02. Similarly, the distance ratio which is the ratio between the distance of the closest neighbor to the second closest one is taken as 1.5.

In unstructured robotic workspaces, our algorithm is designed to deliver accurate and fast object localization. The first five images of Figure 9 captured in our environment having multiple objects validate the said capability of our algorithm. The figure also shows the multiple salient object detection accuracies. As can be seen, we perform the test by taking largely distinct mechanical components with varying shapes and sizes. In our system, we consider the use of the two recognition techniques (see Section 5) to validate their suitability at the detected multiple salient regions. To demonstrate the accuracy of angle and position obtained, 25 observations of distinctly shaped and positioned objects in the robotic workspace are recorded and illustrated in the first five images of the figure. These observations are taken from five different images /sets of objects and five observations per image /set changing object positions. The values pertaining to correctly detected and recognized objects with the estimated orientation (Angular error) and localization (X, Y) error are demonstrated in Table 4. The values provided in the table correspond to each set of objects. The first column of Table 4 denotes the localization error at the detected multiple salient regions only. The second column of Table 4 shows the improvement in localization error after applying our shadow detection algorithm at the detected multiple salient regions, which we had mentioned earlier. The errors are calculated by taking the absolute difference between the measured values and actual values.

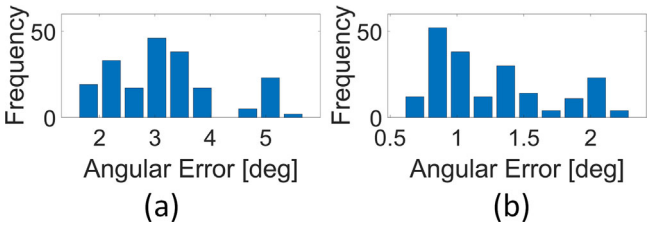

FIGURE 15 Histogram of (a) angular error after applying SIFT matching at detected regions, and (b) angular error after applying shape matching at detected regions

In addition, the localization and orientation error histograms for all objects (objects in first five images of Figure 9 kept at varying positions) are shown in Figures 13, 14 and 15. Figure 13a, b shows the localization errors along the x-axis before and after applying shadow detection technique at the regions detected as salient and Figure 14a, b shows them along Y-axis before and after applying the shadow detection technique. Here, only localization error histograms are shown, as our primary reason for separating shadow partitions is to improve localization accuracy. Figure 15 represents angular errors obtained after applying SIFT and shape matching for recognition only at the regions detected as salient, respectively.

To emphasize the utility of our saliency detection approach for robotic manipulation, we do a performance comparison between our system and a system where multiple salient region detection is not used (termed as conventional method). The basic difference in the systems is that in our system recognition is executed only at the regions detected as salient, whereas in the conventional method the recognition techniques are applied over the entire image. Table 6 summarizes localization and angular errors, and recognition rates when feature point and shape based recognition techniques are applied over the whole image. Similarly, the error histograms for recognition by SIFT and by shape matching without considering saliency (conventional method), are shown in Figures 16 and 17, respectively.

Additionally, we have taken 20 observations where we have placed the objects at various orientations and positions in the robotic work envelope. All the above discussed evaluation parameters are computed considering the 20 observations for each object in Figure 12a (also the shown first image of Figure 9) that is shown being manipulated by the robotic arm. Table 7 summarizes the localization error for the detected salient regions with and without considering shadow detection. Angular errors obtained by SIFT and shape based matching techniques are also summarized in Table 7. Table 8 represents the error values obtained using conventional method for objects being manipulated by the robotic arm.

Processing speed: Consider the computation times given in Table 9. It lists the time taken to perform each step in our approach for robotic pick and place operation, and the total time taken leading to the recognition as well. As can be seen, among the various components of our approach, shadow detection takes only a fraction of time and the most time is consumed by the shape matching at the salient region. The gain in efficiency (total time to recognition) are in the ratio of 0.36:1 for ours (saliency detection + shadow detection + SIFT matching) to conventional SIFT matching and in the ratio of 0.0038:1 for ours (saliency detection + shadow detection + shape matching) to conventional shape matching.

6.2.3 | Discussion

The advantage of employing our proposed multiple salient region detection method is the extraction of multiple objects with exact proper boundaries, which is crucial for their robotic manipulation. Fast location and shape information extraction from objects becomes much easier due to the use of our approach. Summing up the results of the robotic manipulation experiments in Tables 4–8, and Figures 13–17, we conclude that multiple salient region detection can be effectively and successfully used in robotic pick and place by making the existing object recognition techniques more efficient without compromising on performance. As salient regions refer to probable object regions, any recognition technique can be used only at the regions detected as salient instead of on the whole image making the overall system faster and more accurate. This paper shows that the concept of utilizing salient region detection for quick robotic manipulation in an unstructured environment is very pertinent and viable. We observed a few issues that are outlined below, which can be addressed to enhance the performance further.

- Feature matching based recognition was unable to distinguish between a few objects which were very similar in texture /feature, although they were detected correctly by our approach. Use of a recognition approach with higher inter-object discrimination capability post our salient region detection may help.
- A workspace with varying illumination at different areas results in intra-object color variations, which in turn reduces recognition rate. Proper illumination or application of an illumination invariance method can help.
- In case of occlusion, salient region detection and shadow detection can be done successfully. However, object localization and recognition through template matching can be erroneous depending upon the percentage of overlapping of the objects. The use of multiple cameras and/or depth sensors can prove beneficial in such cases.

Another example of robotic pick and place operation using the proposed multiple salient region detection and shadow detection approaches is shown in Appendix A.1, where the camera is at a different point of view.

Further, the algorithm of implementing the proposed multiple salient region detection and shadow detection approaches for the pick and place operation are shown in Appendix A.4 along with the algorithm to perform the related robotic manipulation.

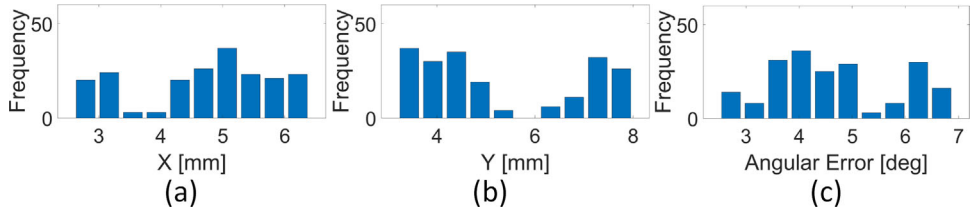

FIGURE 16 Histogram of (a) positional error along X-axis, (b) positional error along Y-axis, and (c) angular error by conventional SIFT matching

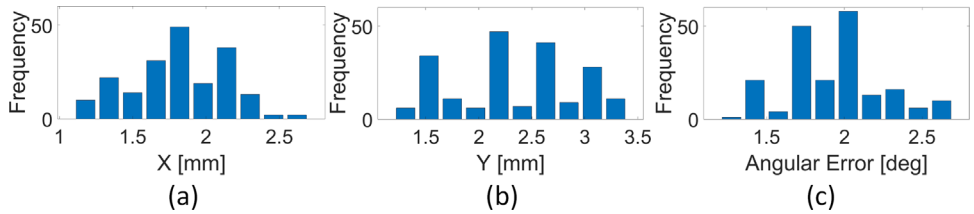

FIGURE 17 Histogram of (a) positional error along X-axis, (b) positional error along Y-axis, and (c) angular error by conventional shape matching

TABLE 8 Detailed evaluation of pick and place performance using the conventional method for specific objects /parts (parts in image1 of Figure 9, Figure 12a). [NA-Not Applicable] [27]

Conventional methods										
	Shape matching					SIFT matching				
	Positional error (cm)					Positional error (cm)				
Assembly parts	X	Y	Detection rate (%)	Angular error (deg)	Recognition rate (%)	X	Y	Detection rate (%)	Angular error (deg)	Recog. rate(%)
Object1	0.25±0.06	0.27±0.05	NA	2.4±0.6	100	0.383±0.03	0.283±0.03	NA	3.5±0.5	100
Object2	0.18±0.04	0.23±0.05	NA	1.6±0.4	100	0.29±0.05	0.31±0.07	NA	2.7±0.2	100
Object3	0.29±0.06	0.24±0.05	NA	1.8±0.5	100	0.28±0.05	0.35±0.04	NA	3.3±0.6	100
Object4	0.22±0.03	0.29±0.02	NA	2.3±0.6	100	0.31±0.04	0.23±0.06	NA	3.6±0.4	100
Object5	0.31±0.05	0.38±0.06	NA	1.9±0.4	100	0.32±0.03	0.41±0.05	NA	2.9±0.5	100
Object6	0.20±0.06	0.32±0.07	NA	3.4±0.3	100	0.26±0.05	0.27±0.04	NA	2.9±0.3	100

6.3 | Summary

This subsection highlights the goals of our proposal and the results achieved, based on the qualitative and quantitative analyses performed earlier in this section.

Consider the following summary of our approach with respect to its goal and the achieved results:

- Our first goal is to detect multiple salient objects. The proposed method is tested on databases containing images with multiple salient objects and the achieved results are presented in Table 1 and in Figures 8, 9, 10, A.3, and A.4. The results demonstrate the effectiveness of the proposed approach in detecting all salient objects in images.

TABLE 9 Computation time (milliseconds/pixel) of the proposed pick and place system, and the conventional matching based techniques

	Proposed system (Total time: 0.176 / 0.266)				Conventional method (total time)	
			Recognition			
Time (mSec)	Saliency detection	Shadow detection	SIFT	Shape matching	Shape matching	SIFT matching
	0.07	0.006	0.10	0.19	69.8	0.5

- Our second goal is to preserve object boundaries accurately for proper localization during robotic manipulation. The qualitative results shown in Figures 8, 9, 10, A.3, and A.4 depict that we have achieved the same.
- Our third goal is to eliminate the effect of shadow in the detection and localization of objects in a robotic environment. This is achieved by proposing a fast and simple hue based shadow detection and the effectiveness of the technique is shown in Figures 11, 13, and 14 and Tables 4 and 7. In the mentioned tables, the quantitative values demonstrate their effect in increasing the localization accuracy of the objects to be manipulated by the robot.
- Our fourth goal is to deploy a robotic manipulation system for pick and place operation based on our proposed multiple salient region detection and shadow detection approaches. Figure 7 shows the collision free strategy for safe execution of the pick and place operation by the robotic manipulator. Tables 4–8 and Figures 13–15 show the accuracy in performance achieved and Table 9 depicts the execution time of the different modules of our proposed system.

Further, a few aspects of our proposed multiple salient region detection technique that stands out from that exist are:

- Our proposed multiple salient region detection algorithm performs well for any number of salient objects present in the scene, and therefore, can be interpreted to scale well with a change in the number of salient objects in images.
- Our approach uses proto-object partitioning that helps in preserving the precise boundaries of the detected salient objects.
- In real robotic environments, our approach achieves good performance in spite of a change in lighting conditions, presence of objects of same shape and size, presence of a large number of salient objects and clutter etc. (see Figure A.3).

7 | CONCLUSION

A novel method for detecting multiple salient regions for application in autonomous robotic manipulation is presented. The proposed multiple salient region detection technique is found to be effective in comparison with the relevant state-of-the-art when evaluated on multiple databases both qualitatively and quantitatively in terms of F-measure, shuffled F-measure, sAUC, and MAE. For improving localization accuracy, the effect of the shadow is reduced by employing a new fast shadow detection algorithm. In vision guided robotic pick and place experiments, our proposed system has been observed to be significantly more efficient compared to when salient region detection is not employed. The efficiency with respect to the rate of detection and recognition, and angular and positional errors was also found to be better while using the proposed method. In an unstructured environment equipped with an industrial robot, we have established the viability and effectiveness of our proposed multiple salient region detection based concept.

ACKNOWLEDGEMENTS

This work was financially supported in part by Tata Steel Pvt. Ltd., India.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ORCID

Debashis Sen https://orcid.org/0000-0002-9756-1191

REFERENCES

1. Herakovic, N.: Robot vision in industrial assembly and quality control processes. In: Robot Vision. InTech, London (2010)
2. Collet, A., Martinez, M., Srinivasa, S.S.: The MOPED framework: Object recognition and pose estimation for manipulation. Int. J. Rob. Res. 30(10), 1284–1306 (2011)
3. Itti, L., Koch, C., Niebur, E.: A model of saliency-based visual attention for rapid scene analysis. IEEE Trans. Pattern Anal. Mach. Intell. 20(11), 1254–1259 (1998)
4. Koch, C., Ullman, S.: Shifts in selective visual attention: Towards the underlying neural circuitry. In: Matters of Intelligence, pp. 115–141. Springer, Dordrecht (1987)
5. Klein, D.A., Illing, B., Gaspers, B., Schulz, D., Cremers, A.B.: Hierarchical salient object detection for assisted grasping. In: IEEE International Conference on Robotics and Automation (ICRA), pp. 2230–2237. IEEE, Piscataway (2017)
6. Qin, X., Zhang, Z., Huang, C., Gao, C., Dehghan, M., Jagersand, M.: Basnet: Boundary-aware salient object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7479–7489. IEEE, Piscataway (2019)
7. Sun, J., Wang, P., Luo, Y.-K., Hao, G.-M., Qiao, H.: Precision workpiece detection and measurement combining top-down and bottom-up saliency. Int. J. Autom. Comput. 15(4), 417–430 (2018)
8. Židek, K., Lazorík, P., Pitel', J., Hošovskỳ, A.: An automated training of deep learning networks by 3D virtual models for object recognition. Symmetry 11(4), 496 (2019)
9. Hossain, D., Capi, G., Jindai, M., Kaneko, S.-i.: Pick-place of dynamic objects by robot manipulator based on deep learning and easy user interface teaching systems. Ind. Rob.: Int. J. (2017)
10. Lin, Y., Tang, C., Chu, F.-J., Vela, P.A.: Using synthetic data and deep networks to recognize primitive shapes for object grasping. In: IEEE International Conference on Robotics and Automation (ICRA), pp. 10 494–10 501. IEEE, Piscataway (2020)
11. Zhang, D., Han, J., Zhang, Y., Xu, D.: Synthesizing supervision for learning deep saliency network without human annotation. IEEE Trans. Pattern Anal. Mach. Intell. 42(7), 1755–1769 (2019)
12. Loncomilla, P., Ruiz-del Solar, J., Martínez, L.: Object recognition using local invariant features for robotic applications: A survey. Pattern Recognit. 60, 499–514 (2016)
13. Tsai, C.-Y., Tsai, S.-H.: Simultaneous 3D object recognition and pose estimation based on RGB-D images. IEEE Access 6, 28 859–28 869 (2018)
14. Manuelli, L., Gao, W., Florence, P., Tedrake, R.: kpam: Keypoint affordances for category-level robotic manipulation. arXiv preprint, arXiv:1903.06684 (2019)
15. Ciocarlie, M., Hsiao, K., Jones, E.G., Chitta, S., Rusu, R.B., Şucan, I.A.: Towards reliable grasping and manipulation in household environments. In: Experimental Robotics, pp. 241–252. Springer, London (2014)
16. Holz, D., Nieuwenhuisen, M., Droeschel, D., Stückler, J., Berner, A., Li, J., Klein, R., Behnke, S.: Active recognition and manipulation for mobile robot bin picking. In: Gearing Up and Accelerating Cross-fertilization between Academic and Industrial Robotics Research in Europe, pp. 133–153. Springer, Cham (2014)
17. Somani, N., Perzylo, A., Cai, C., Rickert, M., Knoll, A.: Object detection using boundary representations of primitive shapes. In: IEEE International Conference on Robotics and Biomimetics (ROBIO), pp. 108–113. IEEE, Piscataway (2015)

18. Teo, C.L., Fermüller, C., Aloimonos, Y.: A gestaltist approach to contour-based object recognition: Combining bottom-up and top-down cues. Int. J. Rob. Res. 34(4-5), 627–652 (2015)
19. Dong, H., Asadi, E., Sun, G., Prasad, D.K., Chen, I.-M.: Real-time robotic manipulation of cylindrical objects in dynamic scenarios through elliptic shape primitives. IEEE Trans. Rob. 35(1), 95–113 (2018)
20. Yang, J., Xiao, Y., Cao, Z.: Toward the repeatability and robustness of the local reference frame for 3D shape matching: An evaluation. IEEE Trans. Image Process. 27(8), 3766–3781 (2018)
21. Li, D., Liu, N., Guo, Y., Wang, X., Xu, J.: 3D object recognition and pose estimation for random bin-picking using partition viewpoint feature histograms. Pattern Recognit. Lett. 128, 148–154 (2019)
22. Liu, M.-Y., Tuzel, O., Veeraraghavan, A., Taguchi, Y., Marks, T.K., Chellappa, R.: Fast object localization and pose estimation in heavy clutter for robotic bin picking. Int. J. Rob. Res. 31(8), 951–973 (2012)
23. Borji, A., Cheng, M.-M., Jiang, H., Li, J.: Salient object detection: A benchmark. IEEE Trans. Image Process. 24(12), 5706–5722 (2015)
24. Zhang, L., Zhang, Q., Xiao, C.: Shadow remover: Image shadow removal based on illumination recovering optimization. IEEE Trans. Image Process. 24(11), 4623–4636 (2015)
25. Guo, R., Dai, Q., Hoiem, D.: Single-image shadow detection and removal using paired regions. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2033–2040. IEEE, Piscataway (2011)
26. Wang, J., Li, X., Yang, J.: Stacked conditional generative adversarial networks for jointly learning shadow detection and shadow removal. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1788–1797. IEEE, Piscataway (2018)
27. Bhuyan, S., Sen, D., Deb, S.: Saliency based fast object localization and recognition for mechanical assembly. In: 15th IEEE India Council International Conference (INDICON), pp. 1–6. IEEE, Piscataway (2018)
28. Borji, A., Cheng, M.-M., Hou, Q., Jiang, H., Li, J.: Salient object detection: A survey. Comput. Vis. Media 5(2), 117–150 (2019)
29. Nie, H., Long, K., Ma, J., Yue, D., Liu, J.: Using an improved SIFT algorithm and fuzzy closed-loop control strategy for object recognition in cluttered scenes. PLOS One 10(2), e0116323 (2015)
30. Seib, V., Kusenbach, M., Thierfelder, S., Paulus, D.: Object recognition using Hough-transform clustering of Surf features. In: Workshops on Electrical and Computer Engineering Subfields, pp. 169–176. Koc University, Istanbul (2014)
31. Ma, L., Ghafarianzadeh, M., Coleman, D., Correll, N., Sibley, G.: Simultaneous localization, mapping, and manipulation for unsupervised object discovery. In: IEEE International Conference on Robotics and Automation (ICRA), pp. 1344–1351. IEEE, Piscataway (2015)
32. Finlayson, G.D., Hordley, S.D., Lu, C., Drew, M.S.: On the removal of shadows from images. IEEE Trans. Pattern Anal. Mach. Intell. 28(1), 59–68 (2006)
33. Cheng, M.-M., Mitra, N.J., Huang, X., Torr, P.H., Hu, S.-M.: Global contrast based salient region detection. IEEE Trans. Pattern Anal. Mach. Intell. 37(3), 569–582 (2014)
34. Yan, Q., Xu, L., Shi, J., Jia, J.: Hierarchical saliency detection. In: IEEE Conference on Computer Vision and Pattern Recognition, (CVPR), pp. 1155–1162. IEEE, (2013)
35. Perazzi, F., Krähenbühl, P., Pritch, Y., Hornung, A.: Saliency filters: Contrast based filtering for salient region detection. In: IEEE Conference on Computer Vision and Pattern Recognition, (CVPR), pp. 733–740. IEEE, Piscataway (2012)
36. Margolin, R., Tal, A., Zelnik-Manor, L.: What makes a patch distinct? In: IEEE Conference on Computer Vision and Pattern Recognition, (CVPR), pp. 1139–1146. IEEE, Piscataway (2013)
37. Liu, T., Yuan, Z., Sun, J., Wang, J., Zheng, N., Tang, X., Shum, H.-Y.: Learning to detect a salient object. IEEE Trans. Pattern Anal. Mach. Intell. 33(2), 353–367 (2010)
38. Shi, K., Wang, K., Lu, J., Lin, L.: Pisa: Pixelwise image saliency by aggregating complementary appearance contrast measures with spatial priors. In: IEEE Conference on Computer Vision and Pattern Recognition, (CVPR) pp. 2115–2122. IEEE, Piscataway (2013)
39. Wei, Y., Wen, F., Zhu, W., Sun, J.: Geodesic saliency using background priors. In: European Conference on Computer Vision, (ECCV), pp. 29–42. Springer, Berlin (2012)
40. Jiang, B., Zhang, L., Lu, H., Yang, C., Yang, M.-H.: Saliency detection via absorbing markov chain. In: IEEE International Conference on Computer Vision, (ICCV), pp. 1665–1672. IEEE, Piscataway (2013)
41. Zhu, W., Liang, S., Wei, Y., Sun, J.: Saliency optimization from robust background detection. In: IEEE Conference on Computer Vision and Pattern Recognition, (CVPR), pp. 2814–2821. IEEE, Piscataway (2014)
42. Zhang, J., Sclaroff, S., Lin, Z., Shen, X., Price, B., Mech, R.: Minimum barrier salient object detection at 80 fps. In: IEEE International Conference on Computer Vision (ICCV), pp. 1404–1412. IEEE, Piscataway (2015)
43. Oh, K., Lee, M., Lee, Y., Kim, S.: Salient object detection using recursive regional feature clustering. Inf. Sci. 387, 1–18, (2017)
44. Zhang, J., Ehinger, K.A., Wei, H., Zhang, K., Yang, J.: A novel graph-based optimization framework for salient object detection. Pattern Recognit. 64, 39–50 (2017)
45. Wang, Q., Yuan, Y., Yan, P., Li, X.: Saliency detection by multiple-instance learning. IEEE Trans. Cybern. 432, 660–672 (2013)
46. Craye, C., Filliat, D., Goudou, J.-F.: Environment exploration for object-based visual saliency learning. In: IEEE International Conference on Robotics and Automation (ICRA), pp. 2303–2309. IEEE, Piscataway (2016)
47. Luo, Z., Mishra, A., Achkar, A., Eichel, J., Li, S., Jodoin, P.-M.: Non-local deep features for salient object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6609–6617. IEEE, Piscataway (2017)
48. Wu, Z., Su, L., Huang, Q.: Cascaded partial decoder for fast and accurate salient object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3907–3916. IEEE, Piscataway (2019)
49. Liu, N., Han, J.: Dhsnet: Deep hierarchical saliency network for salient object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 678–686. IEEE, Piscataway (2016)
50. Li, G., Yu, Y.: Deep contrast learning for salient object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 478–487. IEEE, Piscataway (2016)
51. Zhang, L., Dai, J., Lu, H., He, Y., Wang, G.: A bi-directional message passing model for salient object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1741–1750. IEEE, Piscataway (2018)
52. Zhang, P., Wang, D., Lu, H., Wang, H., Ruan, X.: Amulet: Aggregating multi-level convolutional features for salient object detection. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 202–211. IEEE, Piscataway (2017)
53. Zhang, M., Liu, T., Piao, Y., Yao, S., Lu, H.: Auto-msfnet: Search multi-scale fusion network for salient object detection. In: Proceedings of the 29th International Conference on Multimedia, pp. 667–676. ACM, New York (2021)
54. Yu, S., Zhang, B., Xiao, J., Lim, E.G.: Structure-consistent weakly supervised salient object detection with local saliency coherence. In: Proceedings of the Conference on Artificial Intelligence. AAAI Press, Palo Alto (2021)
55. Wang, Q., Chen, M., Nie, F., Li, X.: Detecting coherent groups in crowd scenes by multiview clustering. IEEE Trans. Pattern Anal. Mach. Intell. 42(1), 46–58 (2020)
56. Jiang, D., Li, G., Tan, C., Huang, L., Sun, Y., Kong, J.: Semantic segmentation for multiscale target based on object recognition using the improved faster-rcnn model. Future Gener. Comput. Syst. 123, 94–104 (2021)
57. Cleveland, J., Thakur, D., Dames, P., Phillips, C., Kientz, T., Daniilidis, K., Bergstrom, J., Kumar, V.: Automated system for semantic object labeling with soft-object recognition and dynamic programming segmentation. IEEE Trans. Autom. Sci. Eng. 14(2), 820–833 (2016)
58. Azad, P., Asfour, T., Dillmann, R.: Combining Harris interest points and the SIFT descriptor for fast scale-invariant object recognition. In: IEEE/RSJ International Conference on Intelligent Robots and Systems, (IROS), pp. 4275–4280. IEEE, Piscataway (2009)
59. Stückler, J., Behnke, S.: Integrating indoor mobility, object manipulation, and intuitive interaction for domestic service tasks. In: 9th IEEE-RAS International Conference on Humanoid Robots, pp. 506–513. IEEE, Piscataway (2009)

60. Li, Y., Chen, C.-F., Allen, P.K.: Recognition of deformable object category and pose. In: IEEE International Conference on Robotics and Automation (ICRA), pp. 5558–5564. IEEE, Piscataway (2014)
61. Xie, Z., Singh, A., Uang, J., Narayan, K.S., Abbeel, P.: Multimodal blending for high-accuracy instance recognition. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 2214–2221. IEEE, Piscataway (2013)
62. Zheng, L., Wang, H., Chen, W.: A fast 3D object recognition pipeline in cluttered and occluded scenes. In: International Conference on Intelligent Robotics and Applications, pp. 588–598. Springer, Berlin, Heidelberg (2017)
63. Cao, Z., Sheikh, Y., Banerjee, N.K.: Real-time scalable 6DOF pose estimation for textureless objects. In: IEEE International Conference on Robotics and Automation (ICRA), pp. 2441–2448. IEEE, Piscataway (2016)
64. Shin, H., Hwang, H., Yoon, H., Lee, S.: Integration of deep learning-based object recognition and robot manipulator for grasping objects. in 16th International Conference on Ubiquitous Robots (UR), pp. 174–178. IEEE, Piscataway (2019)
65. Bergamini, L., Sposato, M., Pellicciari, M., Peruzzini, M., Calderara, S., Schmidt, J.: Deep learning-based method for vision-guided robotic grasping of unknown objects. Adv. Eng. Inf. 44, 101052 (2020)
66. Ding, X., Luo, Y., Li, Q., Cheng, Y., Cai, G., Munnoch, R., Xue, D., Yu, Q., Zheng, X., Wang, B.: Prior knowledge-based deep learning method for indoor object recognition and application. Syst. Sci. Control Eng. 6(1), 249–257 (2018)
67. Jiang, D., Li, G., Sun, Y., Hu, J., Yun, J., Liu, Y.: Manipulator grabbing position detection with information fusion of color image and depth image using deep learning. J. Ambient Intell. Humanized Comput. 12, 10809–10822 (2021)
68. Zhang, X., Liu, J., Feng, J., Liu, Y., Ju, Z.: Effective capture of nongraspable objects for space robots using geometric cage pairs. IEEE/ASME Trans. Mech. 25(1), 95–107 (2019)
69. Salvador, E., Cavallaro, A., Ebrahimi, T.: Shadow identification and classification using invariant color models. In: IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pp. 1545–1548. IEEE, Piscataway (2001)
70. Lalonde, J.-F., Efros, A.A., Narasimhan, S.G.: Detecting ground shadows in outdoor consumer photographs. In: European Conference on Computer Vision (ECCV), pp. 322–335. Springer, Berlin (2010)
71. Zhu, J., Samuel, K.G., Masood, S.Z., Tappen, M.F.: Learning to recognize shadows in monochromatic natural images. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 223–230. IEEE, Piscataway (2010)
72. Khan, S.H., Bennamoun, M., Sohel, F., Togneri, R.: Automatic shadow detection and removal from a single image. IEEE Trans. Pattern Anal. Mach. Intell. (3), 431–446 (2016)
73. Bousseau, A., Paris, S., Durand, F.: User-assisted intrinsic images. ACM Trans. Graphics (TOG) 28(5), 130 (2009)
74. Zheng, Q., Qiao, X., Cao, Y., Lau, R.W.: Distraction-aware shadow detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5167–5176. IEEE, Piscataway (2019)
75. Nguyen, V., Yago Vicente, T.F., Zhao, M., Hoai, M., Samaras, D.: Shadow detection with conditional generative adversarial networks. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 4510–4518. IEEE, Piscataway (2017)
76. Hu, X., Zhu, L., Fu, C.-W., Qin, J., Heng, P.-A.: Direction-aware spatial context features for shadow detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7454–7462. IEEE, Piscataway (2018)
77. Chen, Q., Zhang, G., Yang, X., Li, S., Li, Y., Wang, H.H.: Single image shadow detection and removal based on feature fusion and multiple dictionary learning. Multimedia Tools Appl. 77(14), 18 601–18 624 (2018)
78. Comaniciu, D., Meer, P.: Mean shift: A robust approach toward feature space analysis. IEEE Trans. Pattern Anal. Mach. Intell. 24(5), 603–619 (2002)
79. Sanin, A., Sanderson, C., Lovell, B.C.: Shadow detection: A survey and comparative evaluation of recent methods. Pattern Recognit. 45(4), 1684–1695 (2012)
80. Prati, A., Mikic, I., Trivedi, M.M., Cucchiara, R.: Detecting moving shadows: Algorithms and evaluation. IEEE Trans. Pattern Anal. Mach. Intell. 25(7), 918–923 (2003)
81. Bhattacharyya, A.: On a measure of divergence between two multinomial populations. Sankhyā: Indian J. Stat. 7(4), 401–406 (1946)
82. Michailovich, O., Rathi, Y., Tannenbaum, A.: Image segmentation using active contours driven by the Battacharyya gradient flow. IEEE Trans. Image Process. 16(11), 2787–2801 (2007)
83. Shi, J., Malik, J.: Normalized cuts and image segmentation. IEEE Trans. Pattern Anal. Mach. Intell. 22(8), 888–905 (2000)
84. Ballard, D.H.: Generalizing the Hough transform to detect arbitrary shapes. Pattern Recognit. 13(2), 111–122 (1981)
85. Fischler, M.A., Bolles, R.C.: Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. Commun. ACM 24(6), 381–395 (1981)
86. Belongie, S., Malik, J., Puzicha, J.: Shape matching and object recognition using shape contexts. IEEE Trans. Pattern Anal. Mach. Intell. 24(4), 509–522 (2002)
87. Mishra, A., Sainul, I., Bhuyan, S., Deb, S., Sen, D., Deb, A.: Development of a flexible assembly system using industrial robot with machine vision guidance and dexterous multi-finger gripper. In: Precision Product-Process Design and Optimization, pp. 31–71. Springer, Singapore (2018)
88. Alpert, S., Galun, M., Basri, R., Brandt, A.: Image segmentation by probabilistic bottom-up aggregation and cue integration. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1–8. IEEE, Piscataway (2007). Dataset available at https://www.wisdom.weizmann.ac.il/~vision/Seg_Evaluation_DB/dl.html
89. Ciptadi, A., Hermans, T., Rehg, J.M.: An in depth view of saliency. In: British Machine Vision Conference (BMVC). BMVA Press, London (2013). Dataset available at https://www.cc.gatech.edu/cpl/projects/depth_saliency/
90. Yildirim, G., Sen, D., Kankanhalli, M., Süsstrunk, S.: Evaluating salient object detection in natural images with multiple objects having multi-level saliency. IET Image Process. 14(10), 2249–2262 (2020). Dataset available at https://github.com/gokyildirim/salmon_dataset
91. Calli, B., Singh, A., Bruce, J., Walsman, A., Konolige, K., Srinivasa, S., Abbeel, P., Dollar, A.M.: Yale-CMU-Berkeley dataset for robotic manipulation research. Int. J. Rob. Res. 36(3), 261–268 (2017). Dataset available at https://www.ycbbenchmarks.com/object-set/
92. Achanta, R., Hemami, S., Estrada, F., Susstrunk, S.: Frequency-tuned salient region detection. In: IEEE Conference on Computer Vision and Pattern Recognition, (CVPR), pp. 1597–1604. IEEE, Piscataway (2009)
93. Rahtu, E., Kannala, J., Salo, M., Heikkilä, J.: Segmenting salient objects from images and videos. In: European Conference on Computer Vision (ECCV), pp. 366–379. Springer, Berlin (2010)
94. Li, H., Lu, H., Lin, Z., Shen, X., Price, B.: Inner and inter label propagation: Salient object detection in the wild. IEEE Trans. Image Process. 2410, 3176–3186 (2015)
95. Peng, H., Li, B., Ling, H., Hu, W., Xiong, W., Maybank, S.J.: Salient object detection via structured matrix decomposition. IEEE Trans. Pattern Anal. Mach. Intell. 39(4), 818–832 (2016)
96. Huang, F., Qi, J., Lu, H., Zhang, L., Ruan, X.: Salient object detection via multiple instance learning. IEEE Trans. Image Process. 26(4), 1911–1922 (2017)
97. Wu, Z., Su, L., Huang, Q.: Stacked cross refinement network for edge-aware salient object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 7264–7273. IEEE, Piscataway (2019)
98. Liu, J.-J., Hou, Q., Cheng, M.-M., Feng, J., Jiang, J.: A simple pooling-based design for real-time salient object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3917–3926. IEEE, Piscataway (2019)
99. Wei, J., Wang, S., Wu, Z., Su, C., Huang, Q., Tian, Q.: Label decoupling framework for salient object detection. In: Proceedings of the

IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 13025–13034. IEEE, Piscataway (2020)
100. Bylinskii, Z., Judd, T., Oliva, A., Torralba, A., Durand, F.: What do different evaluation metrics tell us about saliency models? IEEE Trans. Pattern Anal. Mach. Intell. 413, 740–757 (2018)
101. Bochkovskiy, A., Wang, C.-Y., Liao, H.-Y.M.: Yolov4: Optimal speed and accuracy of object detection. arXiv preprint, arXiv:2004.10934 (2020)
102. Farhadi, A., Redmon, J.: Yolov3: An incremental improvement. In: Computer Vision and Pattern Recognition, pp. 1804–2767. Springer, Berlin Heidelberg (2018)
103. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask R-CNN. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2961–2969. IEEE, Piscataway (2017)
104. Redmon, J., Farhadi, A.: Yolo9000: better, faster, stronger. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7263–7271. IEEE, Piscataway (2017)
105. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vision 60(2), 91–110 (2004)

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Bhuyan, S., Sen, D., Deb, S.: Structure-aware multiple salient region detection and localization for autonomous robotic manipulation. IET Image Process. 1–27 (2022). https://doi.org/10.1049/ipr2.12399

APPENDICES

A.1 | Pick and place experiment with changed camera view

To check the robustness, another example of multiple salient object detection and recognition for robotic pick and place operation is presented with a change in camera point of view. First, the camera is calibrated by taking multiple images of a calibration pattern. Few example images of the calibration pattern are shown in Figure A.1a. The distances between the detected corner points in the image, and the corresponding world points projected into the image is called the reprojection error. The accuracy of the calibration is estimated by taking the mean reprojection error for the calibration images which is shown in Figure A.1b. To improve the accuracy in world co-ordinate estimation, lens distortion from the image is removed using the camera parameters. The undistorted image after removing lens distortion is shown in Figure A.2b. In the next step, the rotation and translation of the camera are estimated and using those parameters, image co-ordinates are mapped to real world co-ordinate. In the last step, the obtained real world co-ordinates are mapped with respect to robot co-ordinates. The experiment is carried out using a camera (Logitech c 920, 5MP) and calibration is done using the MATLAB module.

Figure A.2 demonstrates the results of our entire proposed system for the image containing salient objects to be manipulated by the robot. Unlike the previous robotic pick and place operation where localization of detected objects is done taking centroid of the detected regions, in this experiment the location and orientation information is obtained by SIFT matching. The quantitative evaluation results for the five objects placed as shown in Figure A.2a is presented in Table A.1. Table A.1 represents the average positional and angular errors obtained by placing the objects shown in Figure A.2 at 15 different positions. Table A.2 represents the recognition results for conventional SIFT matching technique.

A.2 | Additional qualitative results

In Figure A.3, results are shown in images from our collection of industrial objects and the YCB database, where challenging

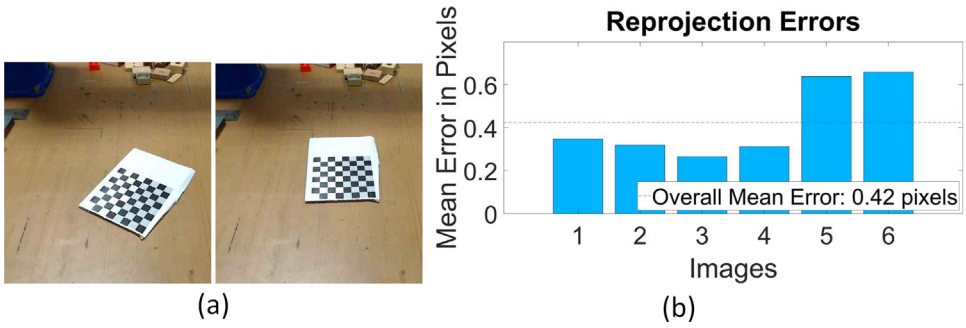

FIGURE A.1 (a) Images for camera calibration. (b) Graph showing mean projection error for the calibration images

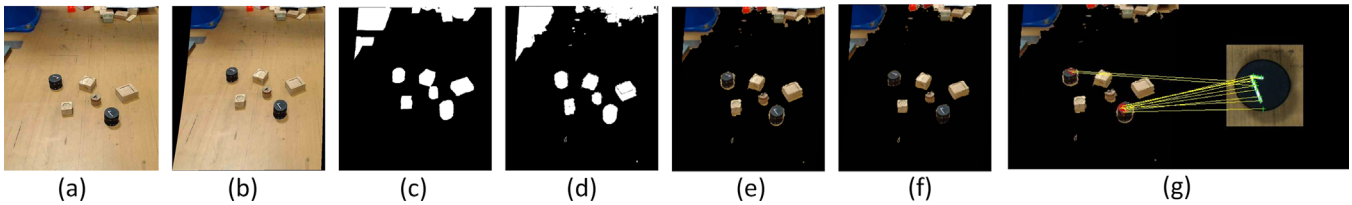

FIGURE A.2 (a) Original image. (b) Undistorted original image after correcting reprojection error. (c) Ground truth. (d) Detected multiple salient regions. (e) Objects present at detected regions. (f) Objects after shadow detection. (g) Recognition by SIFT matching

TABLE A.1 Evaluation of pick and place performance using the proposed method for all set of objects from Figure A.2

Our proposed method

Objects from Figure A.2	Saliency detection			Saliency + shadow detection			Recognition by SIFT	
	Positional error (mm)		Detection rate (%)	Positional error (mm)		Detection rate(%)	Angular error (%)	Recognition rate(%)
	X	Y		X	Y			
Object1	2.9	3.3	100	2.6	3.1	100	2.7	100
Object2	3.3	3.5	100	3.3	3.7	100	3.1	100
Object3	2.7	2.4	100	2.4	2.1	100	2.4	100
Object4	3.1	2.3	100	2.7	2.1	100	3.6	100
Object5	2.7	2.2	100	2.7	2.2	100	3.4	86
Object6	4.0	4.4	100	3.9	4.3	100	2.9	73

TABLE A.2 Evaluation of pick and place performance using the conventional method for all set of objects from Figure A.2.[NA-Not applicable]

Objects from Figure A.2	Detection and recognition by SIFT				
	Positional error (mm)		Detection rate (%)	Angular error (%)	Recognition rate (%)
	X	Y			
Object1	3.2	3.8	NA	4.5	86
Object2	3.7	4.2	NA	3.5	93
Object3	2.9	2.6	NA	4.7	100
Object4	3.9	2.7	NA	3.2	93
Object5	3.2	2.9	NA	3.1	73
Object6	3.7	4.0	NA	5.2	66

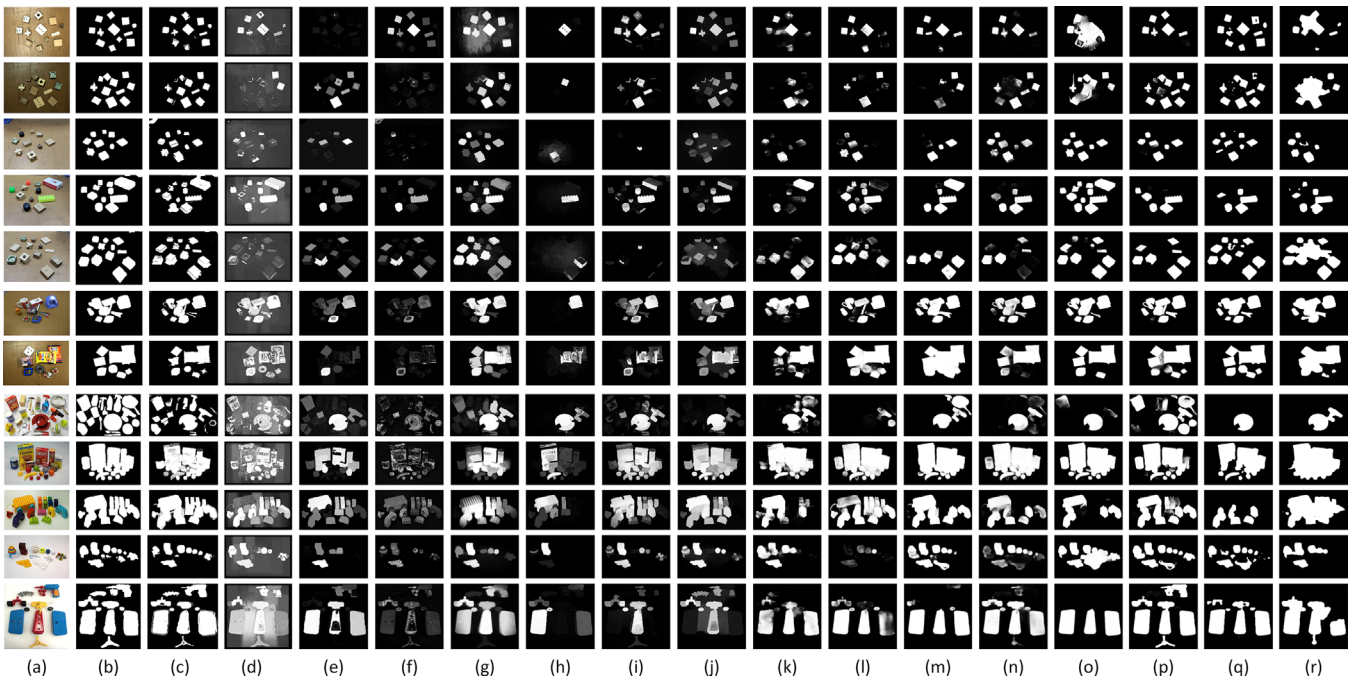

FIGURE A.3 Visual comparison of salient regions obtained by different state-of-the-art algorithms. (a) Input images; (b) ground truth; (c) our method; (d) SEG; (e) HS; (f) FT; (g) MB; (h) LPS; (i) SMD; (j) MIL; (k) NLDF; (l) SPD; (m) CPD; (n) SCR; (o) BASNet; (p) LDF; (q) MSFNet; (r) SCWSSOD. First seven images with multiple objects are our collection of industrial objects taken from our robotic workspace and rest five are from YCB benchmark database[91]

TABLE A.3 Performance evaluation of the proposed multiple salient region detection approach for SED2 and GIT databases with different set of (center/surround) standard deviations. Image size considered here is 256×256

Database	Parameters	Our method sigma set 1 (original) {3, 5,7}/{19,23,29}	Our method sigma set 2 (varied) {3, 5,7}/{21,25,29}	Our method sigma set 3 (varied) {3, 7,9}/{15,19,29}
SED2	shuffled F-measure	0.7505	0.7495	0.7482
	sAUC	0.7778	0.7770	0.7781
	F-measure	0.7987	0.7989	0.7976
	MAE	0.0380	0.0399	0.0387
GIT	shuffled F-measure	0.6088	0.6071	0.5935
	sAUC	0.6836	0.6842	0.6811
	F-measure	0.5380	0.5284	0.5377
	MAE	0.1721	0.1769	0.1736

ALGORITHM 1 Proposed multiple salient region and shadow detection algorithm

Result: $SMap$=saliency map
Data: I=original image, C=center scales, S=surround scales
$MS = \Gamma(I)$ (Estimate proto-object partitions);
while $n <$*length*(C) **do**
  $G = \Lambda(C(n))$ (Generate Gaussian filter);
  $C_{Mask} = \alpha(MS, G)$ (Estimate center operator masks);
end
while $n <$*length*(S) **do**
  $G = \Lambda(S(n))$ (Generate Gaussian filter);
  $S_{Mask} = \alpha(MS, G)$ (Estimate surround operator masks);
end
$CS_{Map} = \beta(C_{Mask}, S_{Mask}, I)$ (Estimate multi-scale center surround map);
$SMap = \gamma(CS_{Map})$ (Estimate saliency map);
Extract salient objects from detected regions;
Extract proto-object partitions from detected regions;
Obtain hue distribution (H) of each proto-object partition;
Obtain fully connected graph using H;
Perform normalised cut and estimate shadow partitions;
Remove shadow partitions to obtain final saliency map;

ALGORITHM 2 Proposed robotic manipulation

Data: I=Acquired image, $\bar{Z}$=safe height, T=template image, CAD= CAD model, $\bar{H}$=distance of supporting plane of objects from robot TCP
$S = \Gamma(I)$ (Detect multiple salient regions);
$\bar{S} = \beta(S, I)$ (Perform shadow detection);
$X, Y, \theta = \alpha(\bar{S}, I, T)$ (Perform recognition);
$H = \zeta(CAD, T)$ (Extract object height);
Move robot in joint mode to $(X, Y, \bar{Z})$, orient the gripper at θ and open gripper;
Move in linear mode to the grasp location X, Y, Z, where $Z = (\bar{H} - H/2)$ and close gripper;
Move in linear mode to the safe height $\bar{Z}$;
Extract jig′s pose information where object to be placed;
Move the robot in joint mode above the jig maintaining a safe height $\bar{Z}$;
Move the robot in linear mode to jig level and open gripper;
Move up in linear mode to $\bar{Z}$;
Repeat steps 3 to 11 for all objects to be picked and placed;

scenarios like low illumination, oblique viewing (camera) angle, objects of similar shape, size and color, and objects cluttered very close to each other exist. Similar to the observations made in Section 6.1.1, we see that our approach is the most consistent in detecting the multiple salient objects present.

In Figure A.4, we show some additional qualitative results on images from the SED2, GIT and SalMoN databases. The observations in the figure leads to the similar deductions about the superiority of the proposed approach in multiple salient region detection as done from Figure 8 of Section 6.1.1.

A.3 | Analysis of the parameters of the proposed approach

In Section 6.1, it is stated that following [3], the standard deviation parameters of $\{3, 5, 7\}$ pixel widths for the center Gaussian function and $\{19, 23, 29\}$ pixel widths for the surround Gaussian function are chosen to perform the proposed multiple salient region detection. Here, we vary these parameters by a reasonable amount and view the change in performance that depicts the sensitivity of these parameters for generic multiple salient region detection. Table A.3 shows the different quantitative performance measures obtained for the images in 2 standard database, when three different standard deviation parameter sets are applied. As can be seen, the variation in performance is almost insignificant, indicating that the proposed multiple salient region detection approach is not substantially sensitive to reasonable change in its parameter values.

A.4 | Algorithms involved in the proposed system

The algorithm of the proposed system is presented in this section. Algorithm 1 describes the proposed multiple salient region detection and shadow detection techniques. The robotic pick and place operation is presented in Algorithm 2.

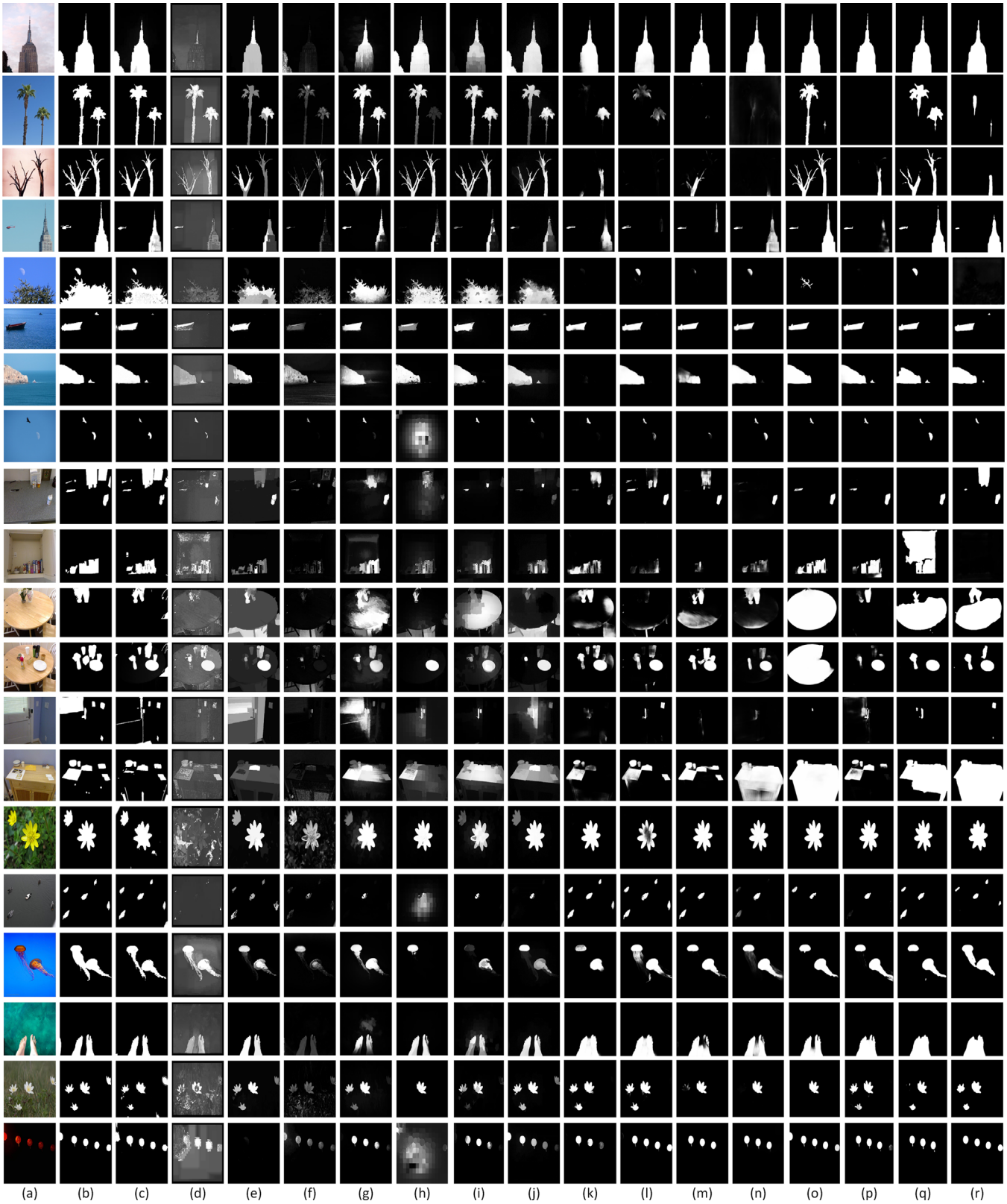

FIGURE A.4 Visual comparison of detected salient regions obtained by different state-of-the-art algorithms on images from SED2 database [88] (First eight rows), GIT database [89] (9th to 14th row) and SalMoN database [90] (the rest). (a) Input images; (b) ground truth; (c) our method; (d) SEG; (e) HS; (f) FT; (g) MB; (h) LPS; (i) SMD; (j) MIL; (k) NLDF; (l) SPD; (m) CPD; (n) SCR; (o) BASNet; (p) LDF; (q) MSFNet; (r) SCWSSOD