PC/BC-DIM Network Based Hybrid Saliency Visual Perception Model for Humanoid Robots

Talha Rehman[1], Wasif Muhammad[1,*], Anum Naveed[1], Muhammad Jehanzeb Irshad[1], Zubair Mehmood[1], Nazam Siddique[1], Sajjad Manzoor[2]

[1]Intelligent Systems Laboratory, Department of Electrical Engineering, University of Gujrat, Gujrat, 50400, Pakistan
[2]Mirpur University of Science and Technology (MUST), Mirpur, Azad Jammu and Kashmir 10250, Pakistan
* Correspondence: syed.wasif@uog.edu.pk

Highlights

- Predictive Coding Biased Competition Divisive Input Modulation Network (PC/BC-DIM)
- Humanoid Robots

Received date: 2024-08-23
Accepted date: 2024-10-30
Published date: 2024-11-13

Abstract: Research on developing visual attention models and saliency detection for humanoid robots has exploded in recent years. A hybrid visual attention model for human-robot interaction can be created by combining the top-down and bottom-up visual saliency detection methods. Due to their high computational cost and complexity, most hybrid visual saliency models are not computationally viable for real-world deployment on humanoid robots. The primary flaw in most visual attention models is that while they can identify the important object in natural photos with a simple background, they struggle to function properly in images with a chaotic or textured background. When there are several prominent items in an image, most global contrast-based techniques do not yield effective results. The problem with hybrid models based on local and global contrast-based techniques is that they tend to predict background regions as salient regions. This study presents a hybrid stereo saliency model that effectively identifies salient objects in backdrop pictures that are simple, crowded, and textured. The suggested paradigm is ideal for implementation on humanoid robots because of its additional benefits, which include simplicity, robustness, and CPU-based execution. Multiple salient items can be detected using the suggested saliency detection model, which computes saliency maps using a Decisive Input Modulation (DIM) neural network, predictive coding (PC), and biased computation (BC). To reduce the complexity of scene analysis preprocessing has been performed using double opponent colors, intensity, and orientation features in the hybrid saliency model. Laplacian of Gaussian (LOG) filter plays a crucial role in processing features such as intensity and orientation features. The top-down factor enhances the saliency of a salient region. The PC/BC-DIM network computes the saliency of preprocessed images after passing through the network. The stereo visual attention model performs preprocessing and saliency map computation separately for each eye and it used depth information as a cue for stereo saliency detection. At the end, binocular saliency maps are combined using the disparity map calculation technique for the extraction of the stereo saliency map. The mean absolute error (MAE) score for the monocular hybrid saliency model was 0.22 and for stereo saliency model MAE score was 0.375. Both monocular and binocular models are computationally efficient and cost-effective for implementation on humanoid robots.

Keywords: Hybrid Visual Saliency, Predictive Coding Biased Competition Divisive Input Modulation Network (PC/BC-DIM), Bottom-up Model, Laplacian of Gaussian (LoG), Top-down Model.

1. Introduction

Due to its use in region of interest detection (ROI), forecasting human image browsing, visual attention, and editing, image quality assessment, and numerous other applications, saliency detection has emerged and grown in prominence in recent years. One of the fundamental characteristics of the human visual system is its ability to provide prompt, appropriate attention to any object. Instead of concentrating on video-based saliency detection, researchers have concentrated saliency detection on still images. According to its characteristics, attention can be classified as "overt" or "covert" [6]. While covert attention involves consciously focusing on one of the many potential sensory inputs, overt attention involves orienting the fovea toward a stimulus [6]. While the covert visual attention model depends on sensory stimuli without altering gaze attention, the overt visual attention model requires the camera maneuvering mechanism. Color, orientation, size, motion, and depth are examples of attention-driven low-level visual features; on the other hand, face detection, text, or people are examples of high-level features that are specifically designed to identify prominent items. Bottom-up and top-down factors are two techniques that can simulate human attention to the provided input images. According to [25], the bottom-up method depends on aspects of the input image, including color, contrast, orientation, and intensity. The top-down method depends on characteristics of the internal history of sensory data, including the human's motivation, goals, and goal-directed attention [15].

For textured or chaotic visuals, the majority of visual attention models do not produce appropriate results. The primary disadvantage of color contrast-based techniques is that they don't work well for pictures with almost identical backgrounds and foregrounds. When there are several prominent items in an image, the majority of global contrast-based techniques do not yield effective results. Similar to the human visual system, the hybrid model accurately detects salient objects by utilizing both low-level and high-level feature information. Due to their high computational cost and model complexity, hybrid visual saliency models are typically not cost-effective for use in object identification and on humanoid robots. The hybrid models based on local and global contrast-based methods have limitation that mostly background regions are predicted as salient regions.

Stereoscopic vision refers to the attribute of the human visual system to view objects from both eyes and the brain’s capability to compute the spatial distance from the difference between two pictures on the retina and to create a combined overall image. Human beings can avail stereo vision to see the object in all dimensions and observe its width, height, and depth. Depth information is pertinent information necessary for the pursuit of saliency detection in stereoscopic images. The main problem in most of the saliency detection models is the extraction of depth information from a stereo scene. Most of the saliency detection models did not use depth information which is pertinent information for the stereo saliency detection purpose.

In this work a hybrid stereo saliency model is presented for the visual attention which can efficiently detect single or multiple salient objects from the simple, cluttered, and textured background images. A novel hybrid stereo saliency model is presented which is simplest model, robust, and based on a CPU system, due to which this model is economical for humanoid robotic development purposes and indoor mobile robots. The advantage of the proposed hybrid visual attention model is its biological plausibility and enhanced saliency of a salient object. The other benefit of the hybrid visual attention model is that a defined boundary of the salient object and maximum region of the salient objects are obtained even in cluttered and textured images. Using depth information as a cue for stereo saliency detection, the suggested novel stereo model imitates the HVS for both left and right eye stereo vision. Using the PC/BC-DIM network, this study model considers low-level variables such as hue, intensity, and orientation for both eyes to create a low-level (bottom-up) saliency map. A new disparity technique was used to fuse the bottom-up saliency maps of the left and right eyes. Using new disparity techniques, this model generates results for textured and crowded images that are biologically reasonable.

Visual perception [4] is essential for effective interaction between humans and robots. Camera sensors and their ability to detect diverse sensory inputs enhance visual perception capabilities. O-D thermal view expands the robots' line-of-sight and enhances their perception in low-light conditions. Experimental results demonstrate high accuracy in human target identification. Visual perception is used to enable robots to comprehend their surroundings, identify objects in the environment, and recognize human targets. Human-robot interaction is the most significant one for autonomous systems, as it requires precise information about their behavior, shape, and movement. Traditional sensors often lack a sufficient field of view (FOV) and ample space to accommodate gesture commands. Robots must assist the target of interest and meet the needs of humans in their environment to enhance the human-robot interaction. Robotic perception is essential for autonomous systems, security, safety, behavior analysis, and human-robot interaction. Studies have explored visual perception in robotic

systems, using thermal and visible sensors, gait recognition, and deep convolutional neural networks. Visual perception is essential for human-robot interaction and can be improved through the use of diverse camera sensors. O-D thermal views expand the robot's line-of-sight, enabling better perception even in low-light conditions. Experimental results demonstrate high accuracy in human target identification, enhancing human-robot and robot-robot interactions.

As advanced intelligent manufacturing replaces traditional manufacturing, robots are becoming more and more significant. A major challenge in attaining high-performance robot manipulation is satisfying the demands of contemporary manufacturing, which call for a high degree of precision, complexity, and diversity. Robots must possess more system precision—including repeatable accuracy and sensory precision—than the task itself to perform precise manipulation tasks. For instance, a high-precision visual sensor is utilized to determine the exact location of the hole in the peg-in-hole assembly process. The robot then determines the position coordinates of the peg using a high-precision encoder and computes the assembly trajectory. To guarantee that the peg's location matches the intended assembly trajectory for successful assembly, the robot's control torques are computed based on the error between the current and goal coordinates. This allows for modifications to the movement of the robot arm and removes posture mistakes. Visual perception is used to enable robots to comprehend their surroundings, identify objects in the environment, and recognize human targets. Human-robot interaction is the most significant one for autonomous systems, as it requires precise information about their behavior, shape, and movement.

This research paper's primary contributions are as follows:

- The application of a computationally effective hybrid PC/BC-DIM-based model for the detection of prominent objects in textured and crowded images.
- A new stereo vision model for detecting salient objects in stereo pictures via disparity estimation, based on the PC/BC-DIM network.
- Both hybrid and stereo visual attention models are economical and computationally efficient models that can be used with humanoid robots.

1.1 Related Work

The Bayesian based visual saliency model was presented by Butko [6], which can comfortably provide saliency maps in about 10ms per video frame on a modern low-end computer and thus being particularly suitable for robotic applications [7]. The algorithm proposed in [6] provided a useful front-end for the robotic cameras by effectively using foveal information to orient the camera towards likely regions of interest. The intensity channels of images were used, and the difference of box filter was used. The filter impulse response distribution was modeled as a Laplacian distribution with unit variance. The model was empirically evaluated in the domain of controlling saccades of a camera in social robotics situations [6]. The goal was to orient a camera as quickly as possible toward human faces [6]. The advantages of model include less computation cost, fast and light-weighted due to which it is suitable for robotic applications. The robot presented in this research work can operate in real-time and leaves enough cycles to perform other operations. The proposed model is also robust to the lighting conditions. The model can perform efficiently in situations when there are plenty of potential tasks of interest. The limitations of model are that it is hard to implement its camera control conditions. Ali Borji introduced a hybrid visual attention model that uses regression, SVM, and AdaBoost classifiers to learn a direct mapping from low-level features like orientation, color, intensity, and saliency maps of the best bottom-up models with top-down cognitive visual features (e.g., faces, humans, cars, etc.) to eye fixations [8].

The advantages of model include accurate salient object detection, competitive performance for salient object detection but cannot perform detection of high-level features instead of face, car and human. Itti had proposed saliency model based on low level knowledge (Color, intensity, motion and orientation) and high-level knowledge from training of low-level features to different relevant classes based on gaze pattern for human [21]. The advantages of proposed model include computation of eye movement of people while playing games. The limitation of model is that it can perform detection for low level features but unable to predict eye movement. Cerf had proposed a visual saliency model using low-level features (color, intensity, and orientation) and high-level face features. This model has less computational cost and is biologically plausible [9].

The different regions of salient objects overlapped in the output image. Pour had proposed a saliency model based on low-level knowledge of color feature RGB and LAB for different image rarities. The high-level feature knowledge was obtained from the learning of dictionaries from image patches of target objects like cars, humans, etc. [27]. The proposed model can be used for object recognition tasks, but the drawback of the model

is that it works for the only trained object for humans and cars etc. and has high computation cost. Both object detection and the human eye-fixation probability map served as the foundation for the presentation of the baby humanoid robot in [23]. The hybrid model algorithm includes picture segmentation, eye fixation prediction map computation, salient object extraction, and saliency map computation. Both global and local saliency maps were calculated and in global saliency map is obtained through non-linear fusion process maps which belongs to chromaticity and luminance. The local saliency map was obtained from comparison of luminance/chromaticity histogram of area inside and outside some border [23]. The humanoid robot can perform salient object detection, autonomous object detection, show navigation whenever it sees attractive features, can also detect smoke and fire in surrounding environment.

The limitation of robot is that whenever the density of objects is high in environment it is necessary to keep value of distraction angle low and this parameter is again dependent on other parameters such as initial position of robot, density of objects, and empirical value for eye fixation algorithm than the importance of value itself. In [14] Contourlet Transform and the second is a Hybrid model were presented which combines Imamoglu's model and Le Moans's model. The Contourlet transform based model was presented to address issue of the lack of geometrical structure in the two-dimensional separable Wavelet Transform. This model consists of sub-band decomposition and directional transformation. The Laplacian filter and directional filters bank have been used for this purpose. The second step consists of feature map extraction from Contourlet Transform and to obtain final saliency map from local and global saliency map. The other method for saliency-based object detection presented was hybrid model which used LeMoan&al (SSmodel) [1] and Imamoglu model. The RGB image is passed through LeMoan&al (SSmodel) [1] and intermediate saliency map was obtained. The intermediate saliency map was converted into grayscale image and passed to Imamoglu model to obtain final saliency map. The major advantage of hybrid model is that it improves results and is much better than SSmodel [1].

The execution time of hybrid model was also less due to which it was more efficient as compared to SSmodel [1]. The contourlet transform based model can extract contour in any orientation but wavelet transform can extract edge information only in specific direction due to which it was better approach to use contourlet transform as compared to wavelet transform. The indoor mobile robot was presented in [26] to address issues of other models which can detect salient object in natural images but cannot provide efficient results in complex indoor environments. The proposed model consists of opposing a new method comprised of graph-based RGB-D segmentation, primary saliency measure, background distribution measure, and combination [26]. Both background distribution map and saliency map were calculated using the color, depth and spatial layout information of each region for final saliency map calculation. The background distribution measure was obtained from region roundness, boundary connectivity and spatial layout between image center and region. The indoor mobile robot can detect objects using visual saliency model even in different conditions such as different viewpoints, illumination variations and partial occlusions.

The indoor mobile robot presented in [26] can perform saliency-based object detection and address issues of models which can detect salient objects in natural images but cannot provide good results in case of indoor environment having complex background, multiple salient objects and illumination variations. The model is suitable for saliency detection but not good enough for mobile robot. The illumination variations such as shadows degrade segmented salient regions due to which they are prone to rough edges, and also have influence on final saliency map. The computing model for saliency-based detection for service robot was presented in [10]. Both static and dynamic features were used for attention selection purpose. Information from sensor network is transformed and incorporated into the model. The focus of attention (FOA) is selected based on a winner-take-all (WTA) network and rotated by inhibition of return (IOR) principle [10]. A computing model of visual attention based on data fusion using intelligent space was presented [10]. The information was mixed and collected from different cameras which had improved the accuracy and robustness of the model. The intensity, color, depth, orientation, and optical flow features have been used to calculate the saliency map which then passed to the WTA network for choosing FOA. The selective visual attention model presented was useful for object detection in complex environments even in partial occlusions, scale change illumination and variation. Achanta et al. had proposed low level saliency model based on low level features of luminance and color [3]. The proposed model is robust to noise and computationally efficient but shows poor results for same background and foreground color.

This research paper [5] presents a comprehensive survey on sensory equipment that aids in human detection and action recognition in industrial settings. This research paper examines various sensors and

perception techniques used in applications involving human-robot interaction (HRI), robot guidance, and collision avoidance across different types of robotic systems commonly employed in industrial settings. It reviews several applications that heavily rely on human-robot perception (HRP) to achieve human-robot collaboration (HRC). Vision sensors, such as monocular RGB, stereo, RGB-D, and event-based cameras, are prominently featured among the different types of sensors used for human perception in various robotic systems. Two proofs of concept developed by the authors demonstrate potential collaborative robotic applications based on enhanced human perception and interaction capabilities. A summary of the results and possible future trends are presented in the paper's conclusion. It examines the best sensors and techniques for detecting and reacting to human operators in collaborative and cooperative industrial settings. In order to prepare for future collaborative robotic applications that capitalize on enhanced human perception and interaction capabilities, it also presents two proofs of concept that the authors have built.

The ability to process RGB and depth images independently enables a versatile monitoring technique for dynamic obstacle avoidance, reducing the volume of information. Multiple sensors are necessary to ensure full coverage of the robot's 3D workspace and eliminate blind spots. The proposed approach in [4] uses an omnidirectional (O-D) thermal imager to identify human targets and recognize their gestures. It demonstrates a high level of accuracy in identifying human targets and enhances their vision capabilities, overcoming the limitations of traditional methods.

Traditional sensors often lack a sufficient field of view (FOV) for simultaneous tracking and analysis of multiple targets' body features and motion behaviors. Robots must assist the target of interest and meet the needs of humans in their environment to enhance the human-robot interaction. Our proposed approach uses visual perception to enable human-robot and robot-robot interactions based on command cognition.

We employ multiple sensors to identify targets and aid, and an omnidirectional (O-D) robot directs the other robots. Future plans involve utilizing gesture recognition to predict the targets' next movements. Robotic systems [29] are limited in their ability to react quickly to dynamic stimuli due to low temporal resolution and time required to process data. To capture high-speed dynamics, the sampling rate needs to be increased, and the available bandwidth and computing capabilities limit the speed at which data can be processed. Vision neuromorphic chips have demonstrated impressive performance in tasks such as tracking, ball goalkeeping, and pencil balancing.

In contrast to a state-of-the-art artificial attention system, this work presents a biologically inspired attention system for the humanoid robot iCub that is noticeably faster at choosing and focusing on novel stimuli. Because they only collect, send, and process data when an input change is recognized, event-driven vision systems are more efficient. Activities include object segmentation, tracking, recognition, scene comprehension, action selection, visual tracking, object manipulation, navigation, self-localization, and simultaneous localization and mapping (SLAM) all require selective attention. EVA is a real-time selective attention implementation that is based on the bio-inspired paradigm that Itti and Koch (2001) developed. To enhance robotic visual perception in autonomous systems, an omnidirectional (O-D) vision system is employed. High precision manipulation can be accomplished with less dependence on precise sensing and control thanks to the ARIE approach. High degrees of precision are achievable by robots, but they also need to be dependable, flexible, capable of learning quickly, and able to balance speed and accuracy when making decisions.

This review [28] examines the latest advancements in human-inspired intelligent robots, focusing on decision-making, cognition, motion control, and system design. By integrating robotics, artificial intelligence, brain science, and neuroscience, researchers can gain insights from the internal mechanisms to external structures of humans. In the future, human-inspired intelligent robots will continue to receive increased attention and developmental opportunities due to their similarity to humans in intrinsic mechanisms and external structures. Human-inspired intelligent robots are expected to receive increased attention and development opportunities due to their similarity to humans in terms of intrinsic mechanisms and external structures. This will enable a better understanding and anticipation of human collaborator requirements, efficient experience sharing, and friendly responses, leading to seamless human-robot collaboration in unstructured environments.

In a large lab room, the primary target was to be able to communicate with the robot and control its movement with 96.99% accuracy. The suggested approach made it possible to understand gesture commands in human-robot interaction over a large region (20.26 to 122.88 m2). For left, right, and the primary target, the accuracy of gesture command decisions was 98.43%, 92.28%, and 90.49%, respectively. The technique recognized gesture commands with 93.75% accuracy without requiring training data or a training procedure.

Robotic systems have used visual perception [4] to comprehend their surroundings, detect human targets, and distinguish items in their environment. This study makes use of the heat signatures released by surrounding objects and human bodies by using an omnidirectional (O-D) thermal imager to cover a broad horizontal field of view (FOV) of 360°. In order to navigate around other humans and keep them aware of their surroundings, the thermal view of humans is evaluated to comprehend their trajectory behavior and movement kinematics during mobility. The suggested approach successfully overcomes the difficulties of target identification and gesture recognition that come with conventional techniques by exhibiting great accuracy in identifying human targets and improving their restricted visual skills. By comparing the final vector with each sensor's estimation, the error of the final target direction was determined. From each sensor, human targets were identified based on their kinematics and beginning trajectories.

When trajectory and kinematic studies were integrated, the stereo sensor obtained the lowest identification error of 9.38%. In a large lab room, the primary target was able to communicate with the robot and control its movement with 96.99% accuracy. The suggested approach made it possible to understand gesture commands in human-robot interaction over a large region (20.26 to 122.88 m2). For left, right, and the primary target, the accuracy of gesture command decisions was 98.43%, 92.28%, and 90.49%, respectively. The technique recognized gesture commands with 93.75% accuracy without requiring training data or a training procedure. For visual perception-based human-robot and robot-robot interaction, a command cognition system was suggested. Using a variety of sensors, we were able to identify targets and provide assistance through gesture commands. From every angle, the O-D robot made it easier for several robots to work together. The future involves using gesture recognition and multiple human target identification to anticipate their next actions. The process is repeated from a new perspective utilizing a separate robotic visual perception, and human targets are tagged based on their movement trends. The labeled targets are then identified using the transferred kinematics, and a relationship between the same targets in the two distinct thermal views is established.

The main target communicates commands through gesture signals, which are recognized by the O-D robot. A 3D map shows the positions of the robots and targets, and the gesture command designates the main target. The robots collaborate to follow the humans. The most important details are that thermal images captured by the O-D sensor, stereo sensor, and perspective single sensor were converted to binary images using a threshold pixel value corresponding to human body temperature. PCA was applied to each part of the target region to determine the orientation of the head, arm, and leg regions. The accuracy of these labels was compared, with the O-D sensor showing 22.23% error, while the stereo and single perspective sensors had 11.42% and 11.12% error, respectively. The smaller size of targets in the O-D sensor images and sudden changes contribute to its higher error rate. We tracked the target regions by fitting boxes to them and monitored their changes in consecutive image frames.

2. Methodology

The hierarchical neural network PC/BC-DIM is a reliable model with a consistent model incorporating a range of neurophysiological and psychophysical data is PC/BC-DIM, a hierarchical neural network. Its foundations include Divisive Input Modulation (DIM) [34], predictive coding (PC) [20], and biased competition (BC) [32]. The PC/BC-DIM network is predicated on the idea of perceptual inference carried out in the cortex, claims Spratling [33]. The error between actual sensor data and projected sensor input is reduced iteratively using this network. Prediction error in the V1 region of the primary visual cortex determines the response of the prediction neuron [32]. The free-energy theory is put into practice in predictive coding. This idea states that an activity that will reduce error is created from the sensory prediction mistakes. W.M. Sparling claims that the PC/BC-DIM paradigm is predicated on prediction mistakes in the primary visual cortex's V1 region. The dynamic attribute of this network model is steady.

$$E_o = X_o \oslash (w_2 + \sum_{n=1}^{p} W_{ok} * Y_k) \quad (1)$$

$$Y_k \leftarrow (w_1 + Y_k) \otimes \sum_j W_{ok} \star E_j \quad (2)$$

$$Y_k \leftarrow Y_k \otimes (1 + \eta A_k) \quad (3)$$

The synaptic weights for neuron type I's RF in input channel j are represented as a 2-dimensional matrix (a convolution mask) for each element {i,j} of the cell array. The output prediction neuron, represented by the

Yk sign, is a cell array. Each element of the Yk cell array is a 2-dimensional matrix of size{A}, where A is the number of different types of neurons. The predicted activation for type I neurons is defined by each component of the cell array. Each element of the Ej cell array is a 2-dimensional matrix of size {B}, where B is the number of input channels. The Ej symbol stands for the prediction error neuron, which is a cell array. Each cell array element defines the mistake in the reconstruction of input channel j. The network determines saliency via residual error, also known as prediction error or θ[ON; OFF]. The reconstruction neuron prediction, represented by the Rj sign, is a cell array. Each member of Rj is a 2-dimensional matrix of size{B} that describes the reconstruction of input channel j. Each 2-dimensional matrix element of the Xo cell array specifies the input to channel j of the current processing phase. The Xo sign stands for the input image, which is a cell array of size{B}. XON (central on/off surround) and XOFF (center off/surround) are parts of Xo. The element-wise division operation is represented by the symbol ⊘.

An operation involving element-wise multiplication is represented by the symbol ⊗. Cross-correlation, which is equivalent to convolution but in which the kernel is not rotated by 180 degrees, is denoted by the symbol ⋆. 2-D convolution is represented by the sign ∗. Eo stands for residual error or prediction error neurons, and it is equivalent to the size of the input image. The size of the input image is represented by Yk, the prediction neurons. Wok represents 2-D Gabor [22] (Kong et al., 2013) [24] [16] weighted filter of size 32 by 2 kernel that gives synaptic weights, normalized with maximum value ψ =5000. The wok is also a weighted kernel, having normalized maximum value ψ =5000, P represents that total no of Gabor [24] [16] weighted kernels. W2 is set to 250, w1 is equal to 0.0001, and η is equal to 1. The weighted total of top-down prediction neurons from the extrastriata cortical region is represented by Ak, a two-dimensional array of the same size as the input image. XON (central on/off surround) and XOFF (center off/on surround) are two preprocessed pictures for intensity. For each feature, there are four XON (central on/off surround) and XOFF (center off/on surround) pre-processed photos for orientation.

Double opponent colors [36] are not pre-processed, they are produced from the center surrounded by the contrast between color channels and submission of single opponent colors as in. All the pre-processed images are stored to Xo, where Xo [ON; OFF]: Xo is given to the V1 network including on and off the channel. According to equation (2), prediction neuron response Yk was initialized with zero value at each index, equal to the size of the input image. Equation (2) applies the division operator to compute the desired residual error or difference of prediction neuron and pre-processed image for a certain feature. Each index of Yk is convolved with normalized weighted 2-D Gabor [22] [24] (Greenspan et al., 1994) filter Wok. Each synaptic weight of Wok is convolved with the corresponding index of Yk. Residual error Eo in equation (1) represents how far corresponding Yk (prediction neuron response) is different as compared to pre-processed Xo output. w2 will avoid infinite response in first iteration because in first iteration the Yk is initialized with zero matrices equal to the size of the input image. Equation (1) computes Eo for the ON and OFF channel of the 2D Gabor [22] [24] [16] weighted filter. In equation (2) new predicted response neurons are computed and updated into equation (1). In equation (2) residual error is cross-correlated with synaptic weights defined by 2-D Gabor [24] [16] filter for ON and OFF channels. After correlation, element-wise multiplication is performed with previous Yk (prediction neuron response) in addition to a very small value of w1. New perceived, prediction response neurons are again forwarded in equation (1). The small value of w1 will help to regain prediction neuron response. Equations (1) and (2) include feed-forwarded processing, in these equations, there is no cortical feedback. Ak is cortical top-down feedback, which is set to zero. Bottom-up processing does not include equation (3). In other words, this equation will be ignored. Comparison of stimulus-driven neurons response such as Xo with sensory-driven prediction neurons response Yk, produces the residual error. The residual error further defines the updated prediction neurons to give saliency map in V1 region of the primary visual cortex. Prediction neurons that defines bottom-up saliency map, depend on residual error. It has been tested and evidenced that the prediction of neurons will be very high whenever the strength of error neurons is high. Both with and without the influence of cortical feedback (bottom-up model) have been incorporated into the hybrid PC/BC model. Ak is a top-down forecast that originates in the brain's V1 area. The response of prediction neurons is highly influenced by this top-down cortical feedback. At this stage of processing, cortical feedback via Ak will boost the response of the prediction neuron. Ak's value is set to 1. For every value of Yk, the value of Ak is set to 1. Equation 4 provides a more detailed representation of the top-down scenario in terms of cortical feedback.

$$Y_k \leftarrow Y_k + Y_k \otimes \eta A_k) \tag{4}$$

The response of cortical feedback on prediction neuron response is provided by the equation above. This cerebral feedback has made it possible to use it to any particular region's receptive fields. In this case, the entire image receives feedback, but only the areas where the response of the prediction neuron is not zero receive it. To provide a saliency map of the V1 region, the output is iteratively sent to equation (1) and then equation (2) for each iteration. The V1 region of the human visual system served as the model's inspiration. Unlike V2 and V4 processing, this paradigm is so efficient that it does not require top-down processing. The process of creating a bottom-up saliency map following input image preprocessing is shown in Fig. 1.

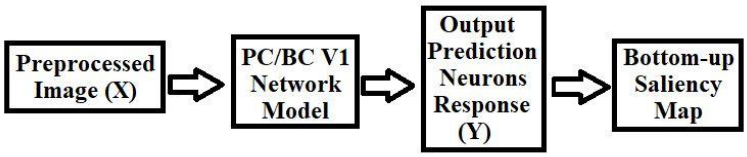


Figure 1. PC/BC-DIM model for hybrid and stereo saliency

1.1 Proposed Hybrid Saliency Model

To extract important properties like intensity, color, and orientation, the 250*250 input image is linearly filtered. Following the division of color channels into red, green, and blue hues, the intensity of an input image is determined. The intensity image is derived from the average values of these colors. According to [22], [24], and [16], Gabor Filter is biologically realistic. Four orientations are employed in Gabor Filter to extract orientation-related characteristics. To provide orientation features, intensity images are additionally convolved with a Gabor filter for each orientation. The input image is utilized to generate biologically believable double opponent colors for color characteristics [36]. Double opponent colors are characteristics that are inspired by biology [36]. A 250 by 250-pixel input image is used for calculations. Jun Zhang's model is used to calculate double opponent colors. RG, YB, RC, and white are double opponent colors [36]. Following convolving with the Gabor [24]; Greenspan et al., 1994) [16] filter at θ[0;45;90;135], four orientation pictures are calculated. In order to detect edges and orientation, pre-processing (LNG/Retina) is performed on intensity and orientation images using a LOG filter [31][24]; [35]. Additionally, the output is gathered in terms of center onoff-surround and center offon-surround. XON (intensity), XOFF (intensity), XON (0), XOFF (0), XON (45), XOFF (45), XON (90), XOFF (90), XON (135), and XOFF (135) are images for intensity and orientation in terms of center onoff-surround and center offon-surround regions.

ALGORITHM 1: Algorithm for Separation of Color Channels and Intensity Features

```
1. image = double('image')
2. result = scalenormalize(image, range)
if isempty(range) == 1
range = [0,1]
end
3. maximum = max(image(:)); minimum = min(image(:)); average = sum(range) / 2
4. oldvariance = maximum - minimum; newvariance = abs(range(2) - range(1))
5. if maximum == minimum
result = image + average - maximum
else
result = (image - minimum) / oldvariance * newvariance + min(range)
end
```

6. redchannel = scalenormalize(image(:,:,1),[0,1])

. function call for extraction of redchannel

7.greenchannel = scalenormalize(image(:,:,2), [0,1])

8.bluechannel = scalenormalize(image(:,:,3), [0,1])

9. I_0 = redchannel/3 + greenchannel/3 + bluechannel/3

ALGORITHM 2: Algorithm for Gabor Filter Convolution with Intensity Features

1. Map.garborientations = [0,45,90,135]

2 for i = 1:1

3. Map.intensity[i] = I0

4. Map.orientation[i,1] = imfifilter(I0, Map.garborfifilter1)

5. Map.orientation[i,2] = imfifilter(I0, Map.garborfifilter2)

6. Map.orientation[i,3] = imfifilter(I0, Map.garborfifilter3)

7.Map.orientation[i,4] = imfifilter(I0, Map.garborfifilter4)

8.Ozero=Map.orientation[i,1]

9. Ofourtyfifive=Map.orientation[i,2]

10. Oninety=Map.orientation[i,3]

11. Oonethreefifive=Map.orientation[i,4]

12. **end**

ALGORITHM 3: Algorithm for The Separation of Double Opponent Color Descriptor

1. Image = double(image)/255

2. x = points(j) * cos(theta) - points(i) * sin(theta)

3. y = points(j) * sin(theta) + points(i) * cos(theta)

4. if sqrt(x * x + y * y) <= rfCount / 2

5. e = exp(-(x * x + aspectRatio * aspectRatio * y * y) / (2 * sigma * sigma))

e = e * cos(2 * pi * x / lambda + alpha)

. e is a fifilter

else

e = 0

5. conv2padded(im(:,:,kk), squeeze(fifilter) // computation of single opponent colors

6. computation of double opponent colors from single opponent colors

The PC/BC-DIM network model for V1 receives all of the preprocessed outputs, including the center on off-surround and center off-surround images. The PC/BCV1 model additionally incorporates the double opponent colors [36] outputs: RG, YBRC, and white. Prediction error neurons, denoted as E, are so anticipated. Everywhere in the image where there is a high value of prediction error neurons, prediction neurons respond with Y. The response of prediction neurons is a gauge of saliency at every visual position. Prediction neurons Y receive additional cortical feedback to increase the salient object's intensity. The LOG filter was used for pre-processing [31] [24] [35]. The Laplacian filter is a derivative function that detects edges, whereas the Gaussian filter conducts orientation contrast at each image position. Therefore, for orientation feature, LOG [31] [24]

[35] will provide orientation contrast for four feature images at orientation [0,45,90,135], while LOG filter [31] [24] [35] will provide intensity images with edge detection regions of all objects residing in an image and orientation contrast of each image location. The retina is used to collect information, and it has a high resolution in a few numbers of concentric field angles. Giving a scene complete foveal representation, such as a high-capacity visual buffer, is necessary to give it the full attention it deserves. The center of the fovea has the highest foveal attention, while the center and a few visual angles have the lowest foveal attention. This idea is applied by simulating concentric receptive fields (RFs) in the retina and LNG. Gaussian and LOG filters are applied with a standard deviation of two pixels for intensity and orientation features [31][24]; Wu, 2016) [35]. After LOG, multiplicative gain is added to strengthen the gain [31][24]; Wu, [35].

$$X = \tanh\{k(I * l)\} \tag{5}$$

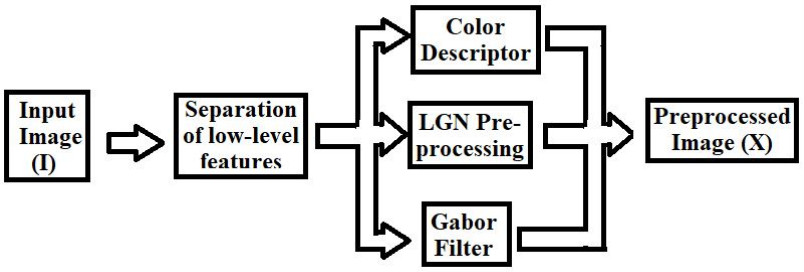


Figure 2. PC/BC-DIM model for hybrid and stereo saliency

The multiplicative gain, or parameter K, responds to nonlinear saturation. The relationship is displayed in equation (5), where l is a LOG filter [31][24]; Wu, [35] and X is a pre-processed image for each orientation and intensity attribute. For every pre-processed image X, I is taken as an intensity image and an orientation image at zero, forty-five, ninety, and thirteenth degrees. For all experiments, K=2 is utilized. XON and XOFF represent X's positive and negative replies. The retina and LNG are represented by the receptive field responses X ON (central on/off surround) and X OF (center off/on surround). For predictive coding, the primary visual cortex's (PC/BC-DIM) input neural network will receive X ON and XOFF channels of intensity and each direction. There is no pre-processing of the double opponent color (J. Zhang et al., 2012) [36] attributes. The pre-processing processes required to obtain a preprocessed image are shown in Fig. 2.

1.2 Proposed Stereo Saliency Model

The bottom-up saliency map fusion of the left and right eyes is included in this section. Cortical feedback from Ak is not considered by each eye's saliency when combining the left and right images. Two methods are available: (1) the sum of absolute differences (SAD) and (2) the stereo disparity computation. While SAD includes noise from camera movement, the first methodology is used in this study since it is more biologically plausible. The number of rows and columns is determined when the normalized saliency maps for the left and right eyes are saved into the variables for the left and right eyes, respectively. Since the number of rows for stereo images is equal, the disparity of the two maps was calculated by dividing the number of columns by the number of rows. The disparity of large images was then added to the images with fewer columns. The final saliency map for images was determined using this stereo saliency model. The procedure for obtaining a stereo saliency map using the PC/BC-DIM network's left bottom-up saliency map and right bottom-up saliency map is shown in Fig. 3.

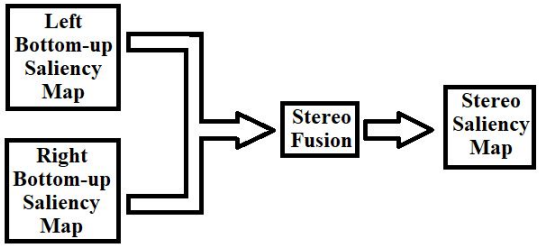


Figure 3. PC/BC-DIM based stereo saliency model

ALGORITHM 4: Algorithm For PC/BC-DIM Based Saliency Detection
1. Read input image
2. Separation of color channels and intensity features
3. Convolve Gabor orientation filter with intensity feature at four orientations
4. [RG,RC,YB,white] = D-Descriptor(image)
5. Define Gabor filter
6. [X]=preprocess(I, O_0,O_{45},O_{90},O_{135})
X = tanh{k(I ∗ l)}
7. A[k]=initialize feedback(image, [a,b])
8. [Y,E]=network model(w,X,iterations,A)
9. **for** k = 1to N
10. **while** k < N
11. E_o=X_o⊘(w_2+$\sum_{n=1}^{p} W_{ok}$ ∗Y_k)
12. Y_k ← (w_1+Y_k)⊗ $\sum_j W_{ok} \star E_j$
13. Y_k ←Y_k ⊗(1+ ηA_k)
14. **End**
15. **End**

3. Results and Discussion

3.1. Demonstration of Hybrid Saliency Model

The output was a saliency image derived from the PC/BC-DIM model, while the input image was 250 by 250 pixels in size. The PC/BC-DIM network received the pre-processed pictures Xo to calculate the Y saliency map. The conspicuous object is present in the 250*250 output image that the suggested model produces. The hybrid model's outcomes and a comparison with other cutting-edge saliency detection methods are presented in this section. The suggested model outperforms the other five cutting-edge saliency models in terms of results. For photos with minimal backdrop orientation and intensity contrast, the suggested model performs better. Whether the model's color contrast is strong or low is irrelevant. Images with high or low color contrast can yield better results using the suggested methodology. Nonetheless, items with contours imbued with strong orientation, color, and intensity contrast must be present in the foreground.

Original RGB input photos are shown in Fig. 4's first row (a). The outputs from the hybrid model without feedback from the extrastriata cortical region are shown in Fig. 4's second row (b). The result of the hybrid model with feedback is shown in the third row. The extrastriata cortical areas provided feedback to the third row's output response neuron. The application of this input resulted in an improved and enhanced saliency output image. The output prediction neuron received 100% feedback. Even with textured and chaotic photos, the results in Fig. 4's third row (c) are superior. The image's prominent objects are represented by well-defined shapes and bounds. The suggested model's outcomes are likewise believable from a biological standpoint. Subheadings may be used to split this section. It ought to offer a succinct and accurate explanation of the experimental findings, their interpretation, and any inferences that may be made from them.

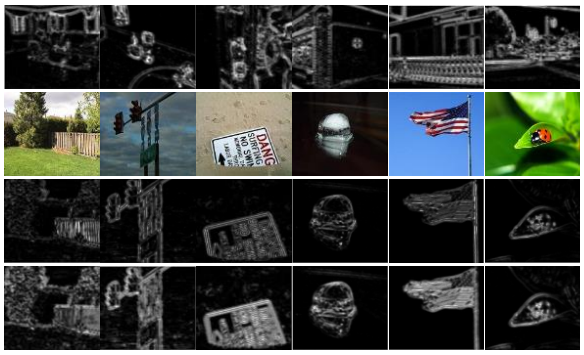
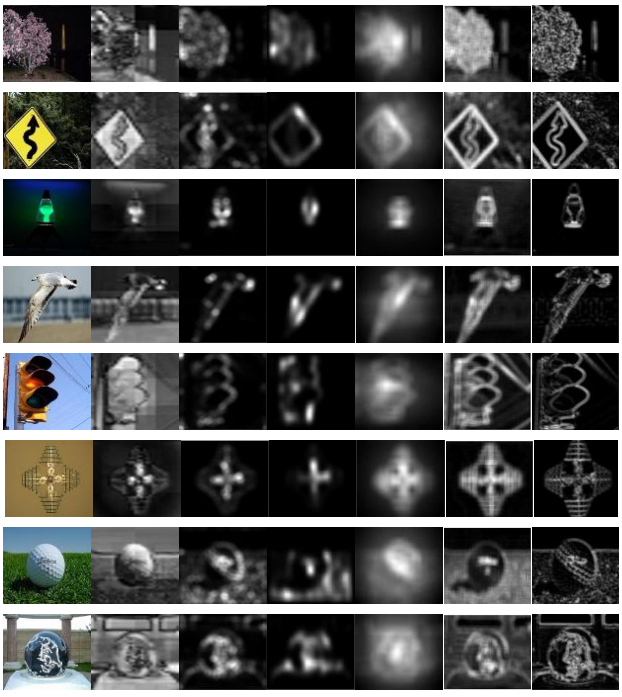


Figure.4. (a) The input images are shown in the first row.(b) The result of the hybrid PC/BC-DIM based saliency detection model without feedback is shown in the second row. (c) The result of the hybrid PC/BC-DIM based saliency detection model with feedback is shown in the third row.

The comparison of the suggested visual attention mode is included in this section. The suggested visual attention model is compared to five other state-of-the-art models in this section. Ittis' model [21], the SR model [18], the SUN model [37], the WU model (Duan et al., 2011) [11], and the Hou model [19] were compared with the output of the suggested model. The outputs of the suggested model and other cutting-edge models are shown in Fig. 5. The textured picture outputs of Itti's model [21] contain superfluous, ill-defined boundaries. Because portions of salient objects overlap, Ittis' model's outputs [21] are not well characterized. The output of the SR model [18] has low resolution for the salient object and lacks clearly defined borders. In addition, the Hou [19] model has a poor resolution for the prominent item and no boundaries. The WU [13] model can precisely predict the borders of salient objects and produces high-resolution output. Although this model contains the most pixels in the area of the prominent object, it is unable to predict the object with any degree of accuracy.

In comparison to the Itti s' model [21], SR [18], WU [13], and Hou [19] models, the SUN model [37] has produced superior results. The conspicuous objects are correctly predicted by this model. The model predicts an additional non-salient zone in textured and cluttered images where the color contrast between the foreground and background is not significantly different. Salient objects are accurately predicted by the suggested hybrid saliency detection model. When dealing with textured or chaotic images, the model forecasts the limits of the prominent object. If there is little color contrast, the model also predicts the salient object. The suggested PC/BC-DIM based hybrid model can precisely identify salient objects in a cluttered background, but Cheng's contrast-based model is unable to predict prominent objects in an image with a cluttered background.



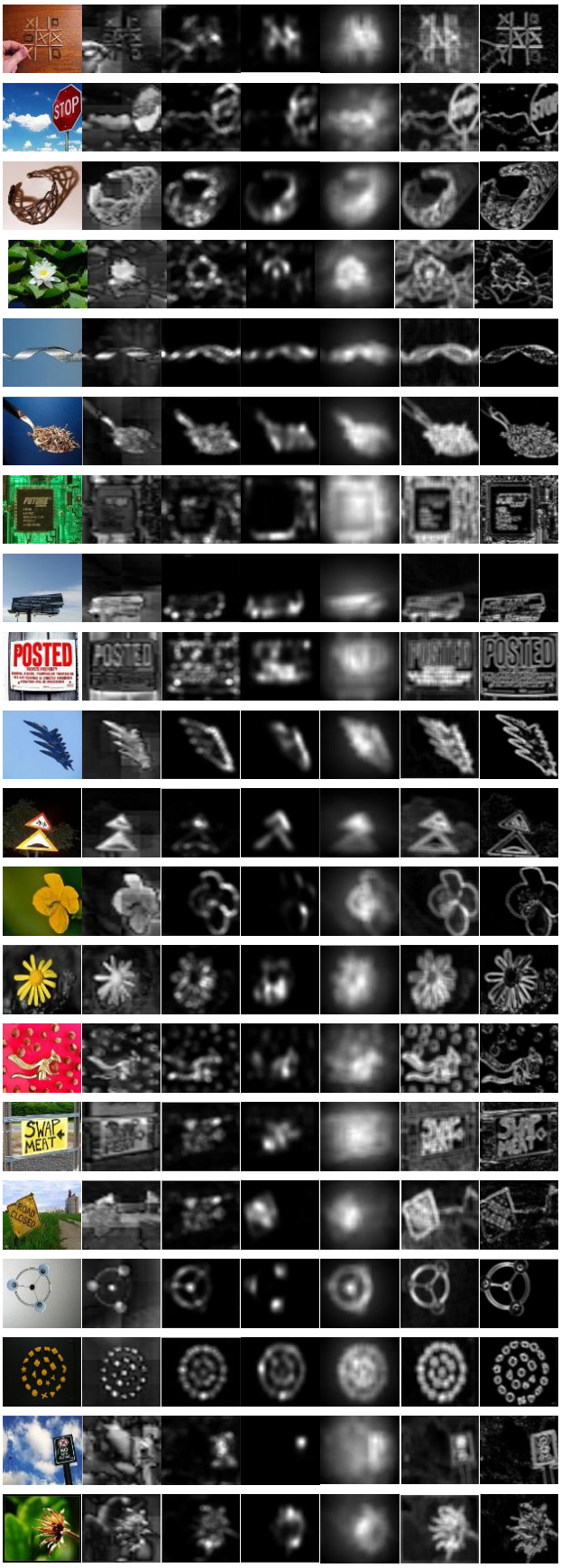


Figure.5. (a) The first column displays the original input photos. (b) The second column (c) displays the result of the Itti model.The third column (d) displays the results of the SR model. The fourth column (e) displays the outcome of the Hou's model. The fifth column (f) displays the outcome of

the WU model. The sixth column (g) displays the outcome of the SUN's model. The last column displays the saliency detection model based on the hybrid PC/BC-DIM.

3.2 Demonstration of Stereo Saliency Model Results

The disparity map technique was used to obtain results from the proposed stereo PC/BC-DIM-based saliency detection model. The difference of an image each pixel value from the pixel value of another image was used to generate results from the proposed PC/BC-DIM-based stereo saliency detection model. The difference of an image each pixel value of the left image was taken from the same pixel value of the right image. If the pixel value in the left image was greater than the right image, then the disparity value was added to the right image and a stereo saliency map was produced as shown in Fig .6. The saliency maps had been shown in the last column of the Fig. 6 depict that the most salient region is present wherever there is more contrast available with respect to background. The outputs of the proposed model are not better for images having low contrast as compared to the background. The outputs of the Fig. 6 depict that the PC/BC-DIM stereo is more efficient and robust for such images where the value for color contrast of the salient regions is more as compared to the background.

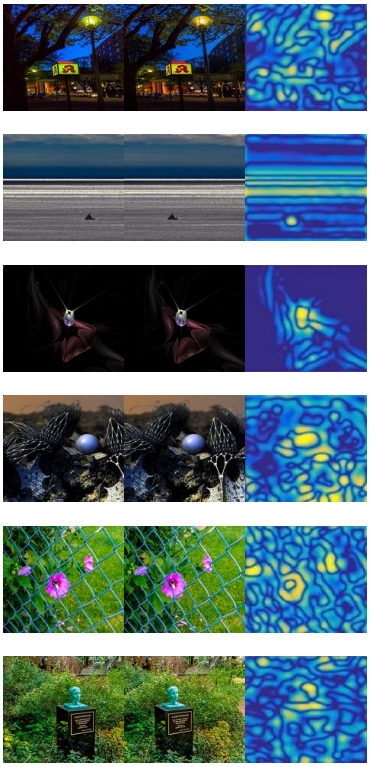
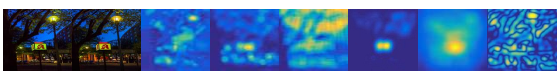


Figure 6. (a) Left camera image. (b) Right camera image (c) PC/BC stereo saliency map.

The proposed stereo model output was compared with Itti model [21], SR model [18], Zhang model [37], SUN model [37], HOU [19], WU model [13]. Fig .7 represents the outputs for Itti model [21] which have noise and blurred as well. The SR model cannot predict most of the salient objects accurately. The SUN [37] model output contains more noise as compared to previously discussed models. The outputs of the SUN model [37] were better if foreground objects have much difference from the background. The results of the HOU (X. Hou & Zhang, 2009) model depict that it is a more plausible model than the previous models. The outputs of HOU (X. Hou & Zhang, 2009) model were better in the case of simple images as compared to complex images. The results for the WU [12] model indicate that this model is better than the previous model but gives results that lack well-defined boundaries. The findings of the suggested PC/BC-DIM stereo saliency detection model demonstrate that it outperforms all of the earlier 2D models that are mentioned here. The highlighted objects in the suggested model have clearly defined boundaries. The suggested model displays boundaries around the salient objects but is unable to produce output with a filled region that represents a salient object.



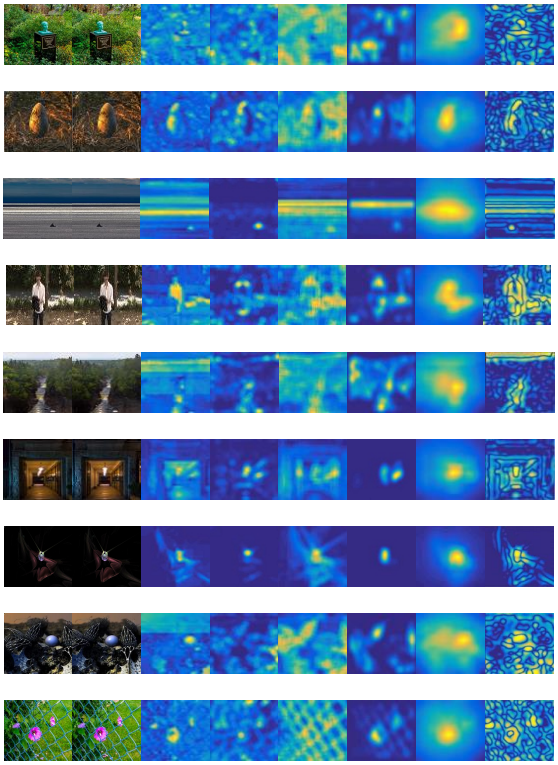


Figure 7. (a) Left camera image. (b) Right camera image (c) Itti (d) SR (e) SUN (f) Hou (g) WU (h) PC/BC stereo Saliency map.

3.3. Performance Evaluation

The MAE score of the suggested hybrid PC/BC-DIM based saliency detection model is compared with various cutting-edge techniques, including the Itti model [30], SUN model (L. Zhang et al., 2008) [37], and WU model [12], in Fig. 8. Between the ground truth and its saliency map, the MAE score was computed. For every model, the MAE graph was created using around 200 photos. The model will be more accurate and precise if the MAE score is lower.

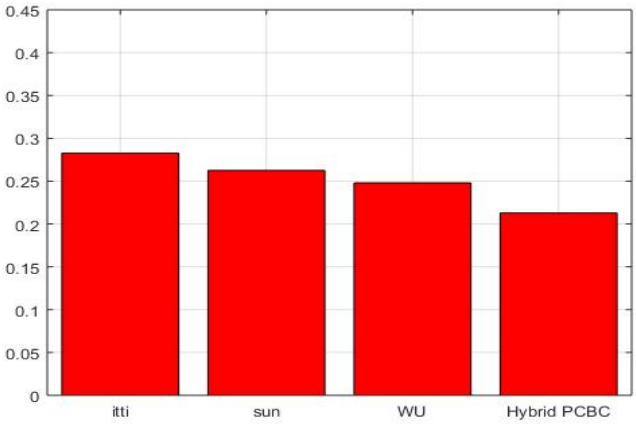


Figure 8. Comparison of proposed hybrid model MEA score with three state-of-art methods.

The Itti [21], SUN, and WU [13] models are inferior to the hybrid PC/BC-DIM. MAE bar graphs were used to statistically analyze PC/BC-DIM stereo saliency detection. Figure 9 shows how the MAE score of the suggested stereo PC/BC-DIM based saliency detection model compares to other cutting-edge techniques including the WU model [13], SUN model, and Itti model [21]. The greater the MAE score of a model, the lower its efficiency will be. One hundred original photos from the MSRA10K dataset and their corresponding

ground-truth images were used to generate MAE bar graphs [11]. The PC/BC-DIM stereo model outperforms the SUN model [37] and the SR model [18]. These models perform less well in crowded settings.

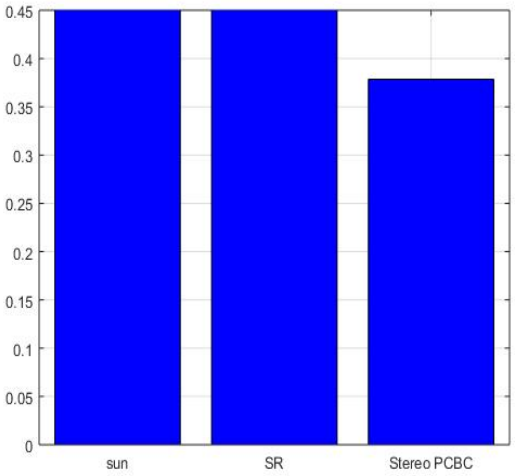


Figure.9. Comparison of proposed stereo model MEA score with three state-of-art methods.

3.4. Computational Cost of PC/BC-DIM Model and Other State-Of-Art Models

The PC/BC-DIM-based hybrid and stereo models are very cost-effective. On core-i3, CPU 1.8 GHZ, 4 GB RAM, and 32-bit bottom-up models like SUN [37], AIM [2], CA [38] and Achanta model [3] the proposed model is cost-effective as compared to these top-down models. While Itti [21], SR [18], HOU (X. Hou & Zhang, 2009) [19], and WU [13] models have less computational cost as compared to the proposed models. A comparison has been shown in the Table 1.

Table 1. Computational cost of PC/BC model and the state-of-art models.

Method	**Achantha2008**	**AIMS2009**	**PC/BC**	**CA**	**SUN**
Code	Matlab	Matlab	Matlab	Matlab	Matlab
Times(s)	24.159	5.915	2.845	141.011	3.564

4. Conclusions

The study that was presented was the first to provide a way for calculating salient items in natural photos using a PC/BC-DIM network model. The input RGB image, which had dimensions of 250 by 250, was used to extract the color channels. The intensity image was derived from the average of the color channels. To provide an orientation response of the input image, the intensity image was linearly filtered at four orientations [0,45,90,135] using Gabor [22] [24] [16] filters. [36] retrieved the attributes of the double opponent color. To obtain the X0 response of each feature, the intensity and orientation pictures at [0,45,90,135] were preprocessed using a LOG [31][24][35] filter.

The intensity and orientation response at the center on off-surround and center off on-surround fields are contained in X0. In the third stage, the neural PC/BC model is given the pre-processed intensity image, orientation response of the pre-processed image, and double opponent colors [36]. It creates an image with the salient object by combining color response, orientation response, and intensity response. By calculating prediction error neurons (E), the PC/BC model forecasts the prediction response (Y). Rather of predicting only one prominent object, the suggested model predicts the entire scene. The edges of every noteworthy object in the picture are returned by the model. Within foreground items, the model provides orientation contrast. Every object is depicted as a prominent object with a complicated orientation and foreground edge contrast. In the

case of textured and congested images, the suggested hybrid PC/BC-DIM based saliency model can precisely identify the borders of salient objects.

This model was tested on MSRA10K [11] and MSRA300 datasets and have the same results with ESD and other databases. The Receiver Operating Characteristics (ROC) and Precision Recall (PR) curves had not gained high scores due to the low resolution of the saliency map, a smaller number of pixels of predictive coding neurons, and hollow regions inside objects where there is no orientation, color, and edge contrast. The proposed model is more suitable for scene recognition of more than one salient object, due to which it can be used for humanoid robot development and other applications. The PC/BC-DIM model is a consistent model according to the theory of cortical anatomy and physiology. The PC/BC-DIM model is a biologically plausible model. The proposed models predict the complete scene instead of only one salient object. The proposed model returns edges of all the salient objects present in the image. The hybrid PC/BC-DIM based model gives orientation contrast inside of foreground objects. This model tends to perform object recognition in this situation due to which this model can be implemented on humanoid robotics and indoor mobile robots. All the objects are detected as salient objects in complex orientation and edges contrast in the foreground. The proposed hybrid saliency detection model and stereo saliency model give better results for images with no orientation and edge contrast. There are some limitations in the model but still, their performance can be improved. Both hybrid and stereo visual attention models are computationally efficient and cost-effective models for implementation on humanoid robots. It is necessary to use more appropriate synaptic weights for the higher-level abstraction of neurons and include global contrast features of the salient object to the background to obtain filled salient regions in the output image. A contrast filter or center-surround contrast followed by the PC/BC network can enhance better precision of salient objects.

References

1. Abouelaziz, I., & El Hassouni, M. (2015). New models of visual saliency: Contourlet transform based model and hybrid ``model. *2015 Intelligent Systems and Computer Vision, ISCV 2015*, *May*, 21–23. https://doi.org/10.1109/ISACV.2015.7105547
2. Achanta, R., Hemami, S., Estrada, F., S., & S. (n.d.). Frequency-tuned salient region detection. In 2009 IEEE Conference on Computer Vision and Pattern Recognition .IEEE. Retrieved from https:// Doi.Org/10.1109/Cvpr.2009.5206596
3. Achanta, R., Estrada, F., Wils, P., & Süsstrunk, S. (2008). Salient region detection and segmentation. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 5008 LNCS, 66–75. https://doi.org/10.1007/978-3-540-79547-6_7
4. Benli, E., Motai, Y., & Rogers, J. (2020). Visual Perception for Multiple Human-Robot Interaction from Motion Behavior. *IEEE Systems Journal*, *14*(2), 2937–2948. https://doi.org/10.1109/JSYST.2019.2958747
5. Bonci, A., Cheng, P. D. C., Indri, M., Nabissi, G., & Sibona, F. (2021). Human-robot perception in industrial environments: A survey. *Sensors*, *21*(5), 1–29. https://doi.org/10.3390/s21051571
6. Borji, A., & Itti, L. (2013). State-of-the-art in visual attention modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*(1), 185–207. https://doi.org/10.1109/TPAMI.2012.89
7. Butko, N. J., Zhang, L., Cottrell, G. W., & Movellan, J. R. (2008). Visual saliency model for robot cameras. *Proceedings - IEEE International Conference on Robotics and Automation*, 2398–2403. https://doi.org/10.1109/ROBOT.2008.4543572
8. Bylinskii, Z., DeGennaro, E. M., Rajalingham, R., Ruda, H., Zhang, J., & Tsotsos, J. K. (2015). Towards the quantitative evaluation of visual attention models. *Vision Research*, *116*, 258–268. https://doi.org/10.1016/j.visres.2015.04.007
9. Cerf, M., Harel, J., Einhäuser, W., Neural, C. K.-A. in, & 2008, U. (2007). Predicting human gaze using low-level saliency combined with face detection. *Advances in Neural Information Processing Systems*, 1–8. http://papers.nips.cc/paper/3169-predicting-human-gaze-using-low-level-saliency-combined-with-face-detection
10. Chen, H., & Tian, G. (2018). A computing model of selective attention for service robot based on spatial data fusion. *Journal of Robotics*, *2018*. https://doi.org/10.1155/2018/5368624
11. Cheng, M., Mitra, N. J., Huang, X., Torr, P. H. S., & Hu, S. (n.d.). *SaliencyTPAMI-1*. *XX*(Xx), 1–14.
12. Duan, L., Wu, C., Miao, J., Qing, L., & Fu, Y. (n.d.). Visual Saliency Detection by Spatially Weighted Dissimilarity School of Information Science and Engineering , Graduate University of the Chinese Academy of. *Image (Rochester, N.Y.)*.

13. Duan, L., Wu, C., Miao, J., Qing, L., & Fu, Y. (2011). Visual saliency detection by spatially weighted dissimilarity. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, *July*, 473–480. https://doi.org/10.1109/CVPR.2011.5995676
14. Elaziz, I. A. (2015). New Models Of Visual Saliency : Contourlet Transform Based Model and Hybrid Model.
15. Frintrop, S. (2006). VOCUS: A visual attention system for object detection and goal-directed search. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 3899 LNAI, 1–228.
16. Greenspan, H., Belongie, S., Goodman, R., Perona, P., Rakshit, S., & Anderson, C. H. (1994). Overcomplete steerable pyramid filters and rotation invariance. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 222–228. https://doi.org/10.1109/cvpr.1994.323833
17. Hou, Q., Cheng, M. M., Hu, X., Borji, A., Tu, Z., & Torr, P. H. S. (2019). Deeply Supervised Salient Object Detection with Short Connections. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *41*(4), 815–828. https://doi.org/10.1109/TPAMI.2018.2815688
18. Hou, X., & Zhang, L. (2007). 2007CVPR_Houxiaodi_04270292.pdf. Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference On, 800, 1–8.
19. Hou, X., & Zhang, L. (2009). Dynamic visual attention: Searching for coding length increments. *Advances in Neural Information Processing Systems 21 - Proceedings of the 2008 Conference*, *800*, 681–688.
20. Huang, Y., & Rao, R. P. N. (2011). Predictive coding. *Wiley Interdisciplinary Reviews: Cognitive Science*, *2*(5), 580–593. https://doi.org/10.1002/wcs.142
21. Itti, L., Koch, C., & Niebur, E. (2005). Short papers meeting, Royal Society of Medicine, London, Section of Coloproctology, 24 November 2004. *Colorectal Disease*, *7*(3), 295–297. https://doi.org/10.1111/j.1463-1318.2005.00780.x
22. Jones, J. P., & Palmer, L. A. (1987). The two-dimensional spatial structure of simple receptive fields in cat striate cortex. *Journal of Neurophysiology*, *58*(6), 1187–1211. https://doi.org/10.1152/jn.1987.58.6.1187
23. Kachurka, V., Madani, K., Sabourin, C., & Golovko, V. (2015). Visual saliency based approach to object detection in computer vision systems: Real life applications. *Proceedings of the 2015 IEEE 8th International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications, IDAACS 2015*, *1*(September), 239–244. https://doi.org/10.1109/IDAACS.2015.7340736
24. Kong, H., Akakin, H. C., & Sarma, S. E. (2013). A generalized laplacian of gaussian filter for blob detection and its applications. *IEEE Transactions on Cybernetics*, *43*(6), 1719–1733. https://doi.org/10.1109/TSMCB.2012.2228639
25. Li, N., Sun, B., & Yu, J. (2015). A weighted sparse coding framework for saliency detection. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, *07-12-June*, 5216–5223. https://doi.org/10.1109/CVPR.2015.7299158
26. Li, N., Xu, H., Wang, Z., Sun, L., & Chen, G. (2017). A salient region detection model combining background distribution measure for indoor robots. *PLoS ONE*, *12*(7), 13–19. https://doi.org/10.1371/journal.pone.0180519
27. Mohammadpour, M., & Mozaffari, S. (2015). A new method for saliency detection using top-down approach. *2015 7th Conference on Information and Knowledge Technology, IKT 2015*. https://doi.org/10.1109/IKT.2015.7288763
28. Qiao, H., Zhong, S., Chen, Z., & Wang, H. (2022). Improving performance of robots using human-inspired approaches: a survey. In *Science China Information Sciences* (Vol. 65, Issue 12). https://doi.org/10.1007/s11432-022-3606-1
29. Rea, F., Metta, G., & Bartolozzi, C. (2013). Event-driven visual attention for the humanoid robot iCub. *Frontiers in Neuroscience*, *7*(7 DEC), 1–11. https://doi.org/10.3389/fnins.2013.00234
30. Schizas, A. M. P., Reid, R., George, M., & Rai, S. (2005). Short papers meeting, Royal Society of Medicine, London, Section of Coloproctology, 24 November 2004. *Colorectal Disease*, *7*(3), 295–297. https://doi.org/10.1111/j.1463-1318.2005.00780.x
31. Sotak, G. E., & Boyer, K. L. (1989). The Laplacian-of-Gaussian kernel: a formal analysis and design procedure for fast, accurate convolution and full-frame output. *Computer Vision, Graphics, & Image Processing*, *48*(2), 147–189. https://doi.org/10.1016/S0734-189X(89)80036-2
32. Spratling, M. W. (2012). Predictive coding as a model of the V1 saliency map hypothesis. *Neural Networks*, *26*, 7–28. https://doi.org/10.1016/j.neunet.2011.10.002

33. Spratling, M. W. (2017). A Hierarchical Predictive Coding Model of Object Recognition in Natural Images. *Cognitive Computation*, *9*(2), 151–167. https://doi.org/10.1007/s12559-016-9445-1
34. Spratling, M. W., De Meyer, K., & Kompass, R. (2009). Unsupervised learning of overlapping image components using divisive input modulation. *Computational Intelligence and Neuroscience*, *2009*. https://doi.org/10.1155/2009/381457
35. Wu, W. (2016). Paralleled Laplacian of Gaussian (LoG) edge detection algorithm by using GPU. *Eighth International Conference on Digital Image Processing (ICDIP 2016)*, *10033*(Icdip), 1003309. https://doi.org/10.1117/12.2244599
36. Zhang, J., Barhomi, Y., & Serre, T. (2012). *Color Image Descriptor*. 312–324.
37. Zhang, L., Tong, M. H., Marks, T. K., Shan, H., & Cottrell, G. W. (2008). SUN: A Bayesian framework for saliency using natural statistics. *Journal of Vision*, *8*(7), 1–20. https://doi.org/10.1167/8.7.32
38. Zhao, R., Ouyang, W., Li, H., & Wang, X. (2015). Saliency detection by multi-context deep learning. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, *07-12-June*, 1265–1274. https://doi.org/10.1109/CVPR.2015.7298731