Research on Salient Object Detection Using Deep Learning and Segmentation Methods

M. Indirani, S. Shankar

Abstract---Detecting and segmenting salient objects in natural scenes, often referred to as salient object detection has attracted a lot of interest in computer vision and recently various heuristic computational models have been designed. While many models have been proposed and several applications have emerged, yet a deep understanding of achievements and issues is lacking. The aim of this review work is to study about the details of methods in salient object detection. It not only focuses on the methods to detect saliency objects, but also reviews the works related to spatio temporal video attention detection technique in video sequences. It also discusses the open issues in terms of evaluation metrics and dataset bias in model performance and suggests future research directions. The evaluation metrics are classified into mean absolute error (MAE), Accuracy and Run-Time complexity.

Keywords--- Spatiotemporal Constrained Optimization Model (SCOM), Context-Aware (CA), Graph-Based Manifold Ranking (GMR), Bootstrap Learning (BL), Deep Learning.

I. INTRODUCTION

Modern day life has overwhelming amount of visual data and information available and being created every minute. This growth in image data has led to new challenges of processing them fast and extracting correct information, so as to facilitate different tasks from image search to image compression and transmission over network.

One specific problem of computer vision algorithms used for extracting information from images is to find objects of interest in an image.

Human visual system has an immense capability to extract important information from a scene. This ability enables humans to focus their limited perceptual and cognitive resources on the most pertinent subset of the available visual data, facilitating learning and survival in everyday life.

Visual saliency is an intriguing phenomenon observed in biological neural systems which received extensive attention by both psychologists and computer vision researchers [1].

It is based on the visual attention, human's ability to quickly locate the most important parts of the scene.

Visual saliency is the perceptual quality that makes an object, person, or pixel stand out relative to its neighbors and thus captures our attention [2].

Therefore, the saliency object detection algorithms attempt to locate dominant, prominent or interesting objects in an image, objects on which humans may also pay more attention.

Generally speaking, there are two different processes that influence visual saliency, one is top-down visual attention model, which uses high-level semantic features and knowledge-driven to compute visual saliency [6], [7].

The other is a bottom-up visual attention model, which is data driven and it relies on image features.

Automatic extraction of high level information is hard and sometimes impossible because it doesn't exist in the particular image at all.

Opposite to that, low level features like contrast, color, orientation are always available. This research work focuses on various salient object detection methods. Deep learning is a well-known process used to detect the object.

Deep learning is a machine learning technique that teaches computers to do what comes naturally to humans: learn by example.

Deep learning is a key technology behind driverless cars, enabling them to recognize a stop sign, or to distinguish a pedestrian from a lamppost.

It is the key to voice control in consumer devices like phones, tablets, TVs, and hands-free speakers. Deep learning is getting lots of attention lately and for good reason. It's achieving results that were not possible before.

Deep learning [3] is revolutionising the way that many industries operate, providing a powerful method to interpret large quantities of data automatically and relatively quickly.

Deterioration is often multi-factorial and difficult to model deterministically due to limits in measurability, or unknown variables.

Deploying deep learning tools to the field of materials degradation should be a natural fit. In this paper, we review the current research into deep learning for detection of object.

In deep learning, a computer model learns to perform classification tasks directly from images, text, or sound.

Deep learning models can achieve state-of-the-art accuracy, sometimes exceeding human-level performance.

Models are trained by using a large set of labelled data and neural network architectures that contain many layers.

The organization of this work is given as follows:

1. The focus of this research is discussed.
2. Discusses about various research methodologies such as background and foreground methods for salient object detection.

Manuscript received September 16, 2019.

M. Indirani, Assistant Professor, Department of IT, Hindusthan College of Engg. and Technology, Coimbatore, T.N, India. (e-mail: mindirani2008@gmail.com)

Dr.S. Shankar, Professor& Head, Department of CSE, Hindusthan College of Engg. and Technology, Coimbatore, T.N, India.(e-mail: shanx80@gmail.com)

Retrieval Number: B10460982S1119/2019©BEIESP
DOI: 10.35940/ijrte.B1046.0982S1119

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

3. Performance evaluations that were conducted to know the better results are analysed.
4. Findings of this overall research work are concluded shortly.

II. REVIEW OF SALIENT OBJECT DETECTION & RESULTS

The researchers have shown a great interest towards pre-attentive or bottom-up saliency detection. Early methods have mostly concentrated on humans in eye fixation prediction and they have introduced the basic principles of saliency detection.

Then, the important problem that has been addressed in the literature is salient object detection and segmentation.

Since, bottom up saliency is stimulus driven and does not look for any particular object in the scene, it can be used for unsupervised segmentation of all the prominent objects in an image. This leads to a solution of the problem of generic object segmentation.

Additionally, literatures in saliency and object segmentation show that graphical model based techniques give efficient modelling and promising results in this field.

2.1. Bottom-up Saliency Models

Bottom up saliency models are mostly inspired by neurophysiology, which adapt the concepts of feature integration theory (FIT) [4] and visual attention [5].

The very first model [6], henceforth referred as IT in the following chapters, uses three features, namely color, intensity and orientation, similar to the simple cells in primary visual cortex.

Center-surround differences over these feature channels generate feature maps that are then normalized across scales and linearly combined to give the saliency. Most computational models are based on either spatial or spectral processing.

Spatial models use different local or global features, like color, intensity, spatial distance, or a combination. Spectral models use a spectral domain analysis of the image and inherently use global features.

Again, all different saliency methods have mainly two approaches- finding a fixation map or generating a saliency map.

Fixation maps [7] try to capture the human eye gaze behaviours and eye fixation points. While they are suitable for many different tasks, e.g., finding fixation scan paths, human gaze pattern analysis, advertise placement in a video, they are not applicable to the problem of salient object segmentation in the field of Computer Vision and Pattern Recognition.

This was discussed by **Zhang et al.,[8]** A new bottom-up salient object detection approach by constructing two graphs using colour and texture features within the manifold ranking framework.

First, it calculates the saliency of boundary patches and excludes the ones with high saliency which might be a part of saliency object.

Second, it adopts a two-stage scheme for salient detection via affinity propagation clustering and graph-based manifold ranking. The background-based saliency detection aims to obtain the salient object regions as much as possible. In the foreground-based saliency detection, a similar computation is processed as that in the former step and yet slightly different.

Instead of simultaneously using all the extracted boundary patches or foreground patches as queries, we compute saliency by using the patches in each cluster in turn and integrating them. At last, the final saliency map is generated by linearly combining two saliency maps respectively exploring color and texture cues.

2.1.1. Background and Foreground Subtraction

Background subtraction is an accurate method for detection of objects in a video, but it requires a priori knowledge of the background image.

The basic idea is to subtract background image pixel by pixel from the current frame and find high differences to detect objects.It is also possible to use background construction methods in case of long scenes where objects enter and exit the scene many times.

Background construction averages all frames in the video, possibly for an initialization time that no objects appear in the scene, to obtain an estimate of the background image.

Zhai et al., [9] developed the model for detection of salient objects in video sequences. Their model is composed of temporal and spatial mode. It was interested only in spatial attention model that is used on still images.

Color statistics of the images are used to reveal the color contrast information in the scene. Based on color contrast they managed to construct a hierarchical representation of the saliency at the pixel level. Given the pixel-level saliency map, attended points are detected by finding the pixels with the local maxima saliency values.

The region-level attention is constructed based on the attended points. This spatial model is one of the fastest salient detection/ attention models in field.

Rahtu et al., [10] proposed the system without any user interaction extracts foreground objects of interest. Foreground and background regions within and across video frames are separated by using proposed method which utilizes visual saliency information extracted from the input video.

The Saliency Map is used to automatically detect salient object from image, and Grab Cut and Adaptive Thresholding for segmentation. It presents the use of saliency maps in segmenting salient objects. The segmentation is used to simplify and split the image in parts with consistent information to analyze.

The Gaussian filter works in characterizing fundamental image edges, i.e. salient edges and can simultaneously reduce insignificant details, hence it produces more accurate boundary information and reduce the effect of noise and achieve a better performance.

Achanta et al., [11] managed to achieve the following requirements: emphasizing largest salient objects; establishing well-defined boundaries of salient objects; disregarding high frequencies arising from texture, noise and blocking artifacts; efficiently output full resolution saliency maps.

These requirements are achieved in frequency domain by filtering low, and high frequency values. They used DoG as a simple band-pass filter.

Sokalski et al., [12] proposed multi-stage salient detection system which combines low-level contrast features, mean-shift segmentation with additional histogram information and multichannel edge features gathered over several feature maps. Due to small number of test images they didn't provide any quantitative analysis of model.

A qualitative result suggests very good accuracy of proposed model.

After time performance analysis of proposed model we concluded that some of the phases are quite time-consuming and impractical on large size images.

Cheng et al., [13] suggested that generic objects with a well-defined boundary share strong correlation in normed gradients space and proposed a saliency-based objectness method.

This method achieves good results in certain conditions, but they still need an exhaustive search before object locating which is time consuming when dealing with large remote sensing images. An efficient coarse object locating method based on a saliency mechanism is proposed, which can generate a small number of bounding boxes as object candidates.

Experiments based on remote sensing images acquired by QuickBird shows that the method could achieve higher accuracy, and the detection time is also reduced significantly.

Goferman et al., [14] presented a context-aware saliency detection algorithm based on four psychological principles: local low-level considerations, global considerations, visual organization rules and high-level factors.

The smoothness prior is used to modify the initial saliency map. The bag-of-features (BOF) framework has been demonstrated to be one of the most successful approaches to scene categorization and object recognition. The BOF strategy is simple and effective.

However, since the classical BOF representation is applied to the entire image, which gives the chance for background clutter to disturb, or even overwhelm the object information, many content irrelevant local features may result in the noisy and non- descriptive visual words in images. Moreover, it also discards the spatial information of local features in the images.

Hou et al., [15] proposed a method whose principle is based on the spectral domain and information theory. Efficient coding decomposes information from the image into innovation part and the redundant part which is already known.

This known part is necessary to suppress by the coding system. Innovation part of the image is called spectral residual and calculated as a difference between logarithmised Fourier transformation of image and generalized shape Fourier. With Inverse Fourier Transform, the spectral residual is converted to spatial domain where it is used to construct saliency map.

Li et al., [16] proposed a method to detect objects using background construction. They claim that their work is usable on even complex and moving backgrounds. The background image is maintained using Bayesian classification of all pixels in the image.

The pixels are classified based on color co-occurrences of inter-frame changes of the pixel. To detect objects, they subtract the maintained background image from the current image. Their method is accurate even in case of dynamic background.

Zhong et al., [17] proposed a novel video saliency detection model for detecting the attended regions that correspond to both interesting objects and dominant motions in video sequences. In spatial saliency map, it inherits the classical bottom-up spatial saliency map.

In temporal saliency map, a novel optical flow model is proposed based on the dynamic consistency of motion. The spatial and the temporal saliency maps are constructed and further fused together to create a novel attention model.

The attention model is evaluated on three video datasets. Empirical validations demonstrate the salient regions detected by our dynamic consistent saliency map highlighting the interesting objects effectively and efficiency.

More importantly, the automatically video attended regions detected by proposed attention model are consistent with the ground truth saliency maps of eye movement data.

Yang et al.,[18] proposed a few methods focus on segmenting out background regions and thereby salient objects. Instead of considering the contrast between the salient objects and their surrounding regions, they considered both foreground and background cues in a different way.

It ranks the similarity of the image elements (pixels or regions) with foreground cues or background cues via graph-based manifold ranking. The saliency of the image elements is defined based on their relevance to the given seeds or queries. It represents the image as a close-loop graph with super pixels as nodes.

These nodes are ranked based on the similarity to background and foreground queries and affinity matrices. Saliency detection is carried out in a two-stage scheme to extract background regions and foreground salient objects efficiently.

Imamoglu et al., [19] analysed to use the multi-scale wavelet transformation to create features and feature maps which represent the contrast or center-surround difference, taking both local and global factors into account.

The wavelet decomposition has the advantage in extracting oriented details (horizontal, vertical and diagonal) in the multi-scale perspective, and enables high spatial resolution with higher frequency components and low spatial resolution with lower frequency components without information loss in details during the decomposition process.

This model creates local saliency map using pixel-level combination of feature maps generated with inverse multilevel wavelet transformation. Additionally, global saliency computation is generated using Probability Density Function (PDF) with a normal distribution.

Zou et al., [20] presented a novel unsupervised algorithm to detect salient regions and to segment out foreground objects from background.

In contrast to previous unidirectional saliency-based object segmentation methods, in which only the detected saliency map is used to guide the object segmentation, our algorithm mutually exploits detection/segmentation cues from each other.

To achieve this goal, an initial saliency map is generated by the proposed segmentation driven low-rank matrix recovery model.

Such a saliency map is exploited to initialize object segmentation model, which is formulated as energy minimization of Markov random field. Mutually, the quality of saliency map is further improved by the segmentation result, and serves as a new guidance for the object segmentation.

The optimal saliency map and the final segmentation are achieved by jointly optimizing the defined objective functions.

Tong et al., [21] proposed a bootstrap learning algorithm for salient object detection in which both weak and strong models are exploited.

First, a weak saliency map is constructed based on image prior to generation of training samples for a strong model. Second, a strong classifier based on samples directly from an input image is learned to detect salient pixels. Multiscale saliency maps are integrated to further improve the detection performance.

Wang et al., [22] proposed video co saliency approach accounts for both inter-video foreground correspondences and intra-video saliency stimuli to emphasize the salient foreground regions of video frames and, at the same time, disregard irrelevant visual information of the background.

Compared to image co-saliency, it is more reliable due to the utilization of temporal information of video sequence. Benefiting from the discriminability of video co-saliency, a unified framework for segmenting out common salient regions of relevant videos, guided by video co-saliency prior is presented.

Unlike naive video co-segmentation approaches employing simple color differences and local motion features, the presented video co-saliency provides a more powerful indicator for the common salient regions, thus conducting video co-segmentation efficiently.

Chen et al., [23] presented a novel model for video salient object detection called spatiotemporal constrained optimization model (SCOM), which exploits spatial and temporal cues as well as a local constraint to achieve a global saliency optimization.

For a robust motion estimation of salient objects, a novel approach is proposed to modelling the motion cues from optical flow field and saliency map of the prior video frame with the motion history of change detection, this approach is enable to distinguish the moving salient objects from diverse changing background regions.

Furthermore, an effective objectness measure is proposed with intuitive geometrical interpretation to extract some reliable object and background regions, which provided as the basis to define the foreground potential, background potential and the constraint to support saliency propagation.

These potentials and the constraints are formulated into the proposed SCOM framework to generate an optimal saliency map for each frame in a video.

Eitel et al., [24] proposed a novel Red Green Blue Depth (RGB-D) architecture for object recognition. The architecture is composed of two separate Convolutional Neural Networks (CNNs) processing streams - one for each modality - which are consecutively combined with a late fusion network.

Focus on learning with imperfect sensor data, a typical problem in real-world robotics tasks. For accurate learning, introduce a multi-stage training methodology and two crucial ingredients for handling depth data with CNNs.

The second, a data augmentation scheme for robust learning with depth images by corrupting them with realistic noise patterns. Presented the state-of-the-art results on the RGB-D object dataset [15] and show recognition in challenging RGB-D real-world noisy settings.

Zhu et al., [25] presented a latent hierarchical structural learning method for object detection. The nodes can move spatially to allow both local and global shape deformations.

The models can be trained discriminatively using latent structural Support Vector Machine (SVM) learning, where the latent variables are the node positions and the mixture component. In this work describe an incremental concave-convex procedure (iCCCP) which allows us to learn both two and three layer models efficiently.

Results show that iCCCP leads to a simple training algorithm which avoids complex multi-stage layer-wise training, careful part selection, and achieves good performance without requiring elaborate initialization.

Perform object detection using our learnt models and obtain performance comparable with state-of-the-art methods when evaluated on challenging public Pattern Analysis, Statistical Modelling and Computational Learning (PASCAL) datasets.

Han et al., [26] proposed high-quality object detection techniques, especially for those based on advanced deep-learning techniques, is still lacking.

To this end, this article delves into the recent progress in this research field, including 1) definitions, motivations, and tasks of each sub direction; 2) modern techniques and essential research trends; 3) benchmark data sets and evaluation metrics; and 4) comparisons and analysis of the experimental results.

More importantly, we will reveal the underlying relationship among Object Detection (OD), Salient Object Detection (SOD), and Category-specific Object Detection (COD) and discuss in detail some open questions as well as point out several unsolved challenges and promising future works.

Girshick et al., [27] proposed a simple and scalable detection algorithm that improves mean Average Precision (mAP) by more than 50% relative to the previous best result on Visual Object Classes (VOCs) 2012-achieving a mAP of 62.4 percent.

The proposed approach combines two ideas: (1) one can apply high- Convolutional Neural Networks (CNNs) to bottom-up region proposals in order to localize and segment objects and (2) when labeled training data are scarce, supervised pre-training for an auxiliary task, followed by domain-specific fine-tuning, boosts performance significantly.

Since combine region proposals with CNNs, call the resulting model a Region-based Convolutional Network (R-CNN).

Chen et al., [28] studied a high-quality 3D objects proposals in the context of autonomous driving. Method exploits stereo imagery to place proposals in the form of 3D bounding boxes.

Formulate the problem as minimizing an energy function encoding object size priors, ground plane as well as several depth informed features that reason about free space, point cloud densities and distance to the ground.

Experiments show significant performance gains over existing Red Green Blue (RGB) and Red Green Blue Depth (RGB-D) object proposal methods on the challenging Karlsruhe Institute of Technology and Toyota Technological Institute (KITTI) benchmark.

Combined with Convolutional Neural Network (CNN) scoring, approach outperforms all existing results on all three KITTI object classes.

Ding et al., [29] proposed a prior knowledge-based deep learning method aimed to enable the robot to recognize indoor objects on sight.

First, integrate the public Indoor dataset and the private Frames of Videos (FoVs) dataset to train a Convolutional Neural Network (CNN). Second, mean images, which are used as a type of colour knowledge, are generated for all the classes in the Indoor dataset.

Finally, when a detection request is launched, the two vectors together with a vector of classification probability instigated by the deep model are multiplied to produce a decision vector for classification.

Experiments show that detection precision can be improved by employing the prior colour and scene knowledge. The results showed potential application of the method for robot vision.

Delforouzi et al., [30] presented a comparative study of outstanding learning-based object detectors such as Aggregate Channel Features (ACF), Region-Based Convolutional Neural Network (RCNN), Fast RCNN, Faster RCNN and You Only Look Once (YOLO) for object tracking.

Use an online and offline training method for tracking. The online tracker trains the detectors with a generated synthetic set of images from the object of interest in the first frame.

The offline tracker uses the detector for object detection in still images and then a tracker based on Kalman filter associates the objects among video frames.

Research is performed on a Tracking-Learning-Detection (TLD) dataset which contains challenging situations for tracking. The results demonstrate that ACF and YOLO trackers show more stability than the other trackers.

Tianet., [31] proposed a video based objection detection method for traceability system with deep learning method.

The surveillance video is collected first, from which an annotated image database of target object such as people or vehicle was constructed to train CNN model off-line. With the trained model, a real-time target detection and recognition system is designed and implemented.

The proposed method mainly includes three aspects: video processing, target detection and object recognition.It provides a variety of video interfaces to support the downloaded video and real-time video stream.

The experimental results indicate that the proposed deep learning based detection method is efficient for the traceability application.

Zhang et al., [32] proposed a deep learning based block-wise scene analysis method equipped with a binary spatio-temporal scene model.

Based on the stacked denoising auto encoder, the deep learning module of the proposed method aims to learn an effective deep image representation encoding the intrinsic scene information, which leads to the robustness of feature description.

Furthermore, the proposed binary scene model captures the spatio-temporal scene distribution information in the Hamming space, which ensures the high efficiency of moving object detection. Experimental results on several datasets demonstrate the effectiveness and efficiency of the proposed method.

Zhao et al., [33] provided a review on deep learning based object detection frameworks. Review begins with a brief introduction on the history of deep learning and its representative tool, namely Convolutional Neural Network (CNN).

Then we focus on typical generic object detection architectures along with some modifications and useful tricks to improve detection performance further.

As distinct specific detection tasks exhibit different characteristics, also briefly survey several specific tasks, including salient object detection, face detection and pedestrian detection.

Experimental analyses are also provided to compare various methods and draw some meaningful conclusions.

Finally, several promising directions and tasks are provided to serve as guidelines for future work in both object detection and relevant neural network based learning systems.

2.2. Analysis

The merits and demerits of the approaches discussed above are given in table 1.

Table 1: Analysis of the discussed methodologies

S.No	Title	Author	Methods	Merits	Demerits
1	Shifts in selective visual attention	Koch, [1987]	Shift of selective visual attention and related visual operations.	1. Processing images at varying resolutions. 2. Image processing with higher level recognition.	More complicated
2	A model of saliency-based visual attention for rapid scene analysis	Itti, [1998]	Feature Integration Theory.	Computationally in efficient manner.	The target detection is critically depends on the feature types implemented.
3	Learning to predict where humans look	Judd, [2007]	Combines both bottom-up image based saliency cues and top-down image semantic dependent cues.	Machine learning is to be used for training a bottom-up, top-down model of saliency based on low, mid and high-level image features.	Eye trackers are expensive and interactive techniques for a more memory consumption.
4	Salient object detection via color and texture cues	Zhang, [2017]	Bottom-up salient object detection approach.	The robustness and efficiency	The background smoothing is very complex
5	Visual attention detection in video sequences using spatiotemporal cues	Zhai, [2006]	Spatiotemporal video attention detection technique.	High control and repeatability of the video sequence.	To quickly highlight the abnormal regions alone, not in the active stage region.
6	A simple and efficient saliency detector for background subtraction	Rahtu, [2009]	Sliding window approach.	Higher accuracy and outperforming the reference methods.	Accurate camera movement estimation is not an easy problem and rapid background model updating is often technically difficult, if not impossible
7	Frequency-tuned salient region detection	Achanta, [2009]	Frequency-tuned approach.	Computational efficiency.	1. Much more complicated, initial alignment can be difficult without proper instruments. 2. Low-frequency content in the image.
8	Automatic Salient Object Detection in UAV Imagery	Sokalski, [2010]	Multi-scale mean-shift segmentation with novel histogram enhancement.	Avoid neighbors annoyed by the noise.	Occasional false positive detection due to the noise in the environment
9	Global contrast based salient region detection	Cheng,[2011]	Regional contrast based saliency extraction algorithm.	High quality saliency maps at the cost of reduced computational efficiency.	Saliency maps should be low and difficult to generate to allow processing of large video collections
10	Context-aware saliency detection	Goferman, [2012]	Context-aware saliency.	To produce compact, appealing, and informative.	Reduce video size in the thumbnail.
11	Image signature: Highlighting sparse salient regions	Hou, [2012]	Image Signature on Synthetic Images.	1. Flexibility 2. Scalability without changing the image quality.	The file size is growing very fast, so it is difficult to find the large number of small elements.
12	Visual saliency based on scale-space analysis in the frequency domain	Li, [2013]	A new bottom-up paradigm for detecting visual saliency.	1. Improved the performance of current saliency models in predicting human attention. 2. The similar salient object to be detected.	It is more computational cost.
13	Video Saliency Detection via Dynamic Consistent Spatio-Temporal Attention Modelling	Zhong, [2013]	A novel optical flow model	1. Effective prominent object detection and coverage. 2.Better efficiency	Limited storage capacity.
14	Saliency detection via graph-based manifold ranking	Yang, [2013]	Graph-Based Manifold Ranking	It unify local and nonlocal methods.	The relatively high dimensionality which make them less suitable for nearest neighbour.
15	A Saliency Detection Model Using Low-Level Features Based on Wavelet Transform	İmamoğlu, [2013]	A novel bottom-up computational model of visual attention.	The statistical relation among the feature maps for the saliency in a global perspective.	1. Global irregularities of the scene can be more dominant than the local irregularities. 2. High down-sampling requirement for images, which would yield spatial information loss.
16	Unsupervised joint salient region detection and object segmentation	Zou, [2015]	Unsupervised algorithm.	1. Iterative and joint optimization. 2. Energy minimization in the salient object.	1. High computational cost. 2. Run-time sensitive applications.
17	Salient object detection via bootstrap learning	Tong, [2015]	Bootstrap learning algorithm.	Alleviating the time-consuming and off-line training process	Restricted within multiple scales of the input image and is unsupervised
18	ViCoS2: Video co-saliency guided co-segmentation	Wang, [2017]	Video cosaliency approach	Foregrounds share similar appearances across videos and salient areas highly contrast with surrounding backgrounds.	Salient objects with complex motion patterns, even in the presence of a cluttered background.
19	SCOM: Spatiotemporal Constrained Optimization for Salient Object Detection	Chen, [2018]	Spatiotemporal constrained optimization model.	High-level features can be transformed to shallower side-output layers.	Limited video tracking.
20	Multimodal Deep Learning for Robust RGB-D Object Recognition	Eitel, [2015]	Multi-stage training methodology.	1. This method is accurate and it is able to learn rich features from both domains. 2. Noise aware training is effective.	It is a time consuming process.
21.	Latent hierarchical structural learning for object detection	Zhu, [2010]	Latent hierarchical structural learning method.	1. Deep structure outperforms shallow structures and that simpler part structures are sufficient to obtain strong results. 2. It reduces the computational cost	Simple structure leads to a meaningful model with good performance.
22.	Region-based convolutional networks for accurate object detection and segmentation.	Girshick, [2016]	Region-based Convolutional Network (R-CNN)	Highly effective for a variety of data-scarce vision problems.	It still takes a huge amount of time to train the network
23.	Prior knowledge-based deep learning method for indoor object recognition and application	Ding, [2018]	A prior knowledge-based deep learning method.	It provides the better performance for detection (or) improves the precision vale of detection.	It is also cause of computational complexity, and taking a long time.
24.	Video object detection for tractability with deep learning method	Tian, [2017]	A video based objection detection method	This method is efficient for the traceability application.	1. This model is computationally expensive. 2. Needed a lot of training data
25.	Deep learning driven block wise moving object detection with binary scene modeling	Zhang, [2015]	A deep learning based block-wise scene analysis method equipped with a binary spatiotemporal scene model.	This method is highly effective and efficient for moving object detection based on binary scene modeling.	High computational cost.

From this analysis table, we can conclude that every methodologies proposed previously consists of various merits and demerits. All the merits and demerits involved in these works are considered for the review from which new methodology can be proposed by combining the merits of all the methodologies. The performance analysis were conducted to check the consistent level of the various proposed methodologies which is described detailed in the following sections. However, the performance of these approaches may decrease when applied to dynamic video scenes. The reasons behind this phenomenon may be four-fold. First, moving targets are generally blurred in video frames, whereas still images are typically captured with sharp focus which guarantees salient objects appearing clearly. Second, salient objects are easily to be partly or even totally occluded by background regions in some video frames, which often leads saliency detection to failure. Third, for objects in videos usually move rapidly and arbitrarily, previous spatial priors that largely determine saliency performance in still images may not work well in dynamic scenes. Last but not least, some background regions may change dramatically during the moving of salient objects, which leads to the difficulty of extracting salient objects. Therefore, it is challenging to detect salient objects from videos. So it left as scope of future work.

III. CONCLUSION AND FUTURE WORK

The research work concerned with the spatiotemporal video attention detection technique for detecting the attended regions that correspond to both interesting objects and actions in video sequences. Deep learning method is giving a very good performance for the detection of object. Future work will focus on introducing the objectness measure used to extract the motion of salient objects from both static and changing background regions in a frame.

REFERENCES

1. Bruce, N. and Tsotsos, J., 2006. Saliency based on information maximization. *In Advances in neural information processing systems* (pp. 155-162).
2. Yantis S., "How visual salience wins the battle for awareness," *Nature neuroscience,* vol. 8, no. 8, pp. 975–977, 2005.
3. Nash, W., Drummond, T. and Birbilis, N., 2018. A review of deep learning in the study of materials degradation. *npj Materials Degradation*, *2*(1), p.37.
4. Jiang, H., Wang, J., Yuan, Z., Wu, Y., Zheng, N. and Li, S., 2013. Salient object detection: A discriminative regional feature integration approach. *In Proceedings of the IEEE conference on computer vision and pattern recognition,* pp. 2083-2090.
5. Koch, C. and S. Ullman (1987). Shifts in selective visual attention: *Towards the underlying neural circuitry.* 188, 115-141.
6. Rasi.D and Suganthi.J, 2016. "A survey on Image Segmentation algorithms", *International Journal of Computer trends and technology,*Vol.35, No.4, pp:170-174.
7. Sasi Kala Rani .K, Rasi.D and Deepa.S.N, 2018. "Developed global biotic cross pollination algorithm for CIS" *International Journal of Business Intelligence and Data Mining,* Vol. 13, Nos. 1/2/3, pp:108-128.
8. Zhao, Q. and Koch, C., 2013. Learning saliency-based visual attention: *A review. Signal Processing,* 93(6), pp.1401-1407.
9. Judd, T., K. Ehinger, F. Durand, and A. Torralba, Learning to predict where humans look. *In CVPR.* 2007.
10. Zhang, Q., Lin, J., Tao, Y., Li, W., & Shi, Y. (2017). Salient object detection via color and texture cues. *Neuro computing,* 243, 35-48.
11. Y. Zhai and M. Shah, "Visual attention detection in video sequences using spatiotemporal cues," *in Proc. ACM Int. Conf. Multimedia,* 2006, pp. 815–824.
12. Rahtu, E., &Heikkilä, J. (2009). A simple and efficient saliency detector for background subtraction. *IEEE 12th International Conference on Computer Vision Workshops (ICCV Workshops),* pp. 1137-1144.
13. R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, "Frequency-tuned salient region detection," *in CVPR,* 2009, pp.-18.
14. Sokalski, J., Breckon, T.P. and Cowling, I., 2010. Automatic salient object detection in UAV imagery. *Proc. of the 25th Int. Unmanned Air Vehicle Systems,* pp.1-12.
15. M.-M. Cheng, G.-X. Zhang, N. J.Mitra, X. Huang, and S.-M. Hu, "Global contrast based salient region detection," *in Proc. IEEE CVPR,* 2011, pp. 409–416.
16. S. Goferman, L. Zelnik-Manor, and A. Tal, Context-aware saliency detection, *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 34, pp. 1915-1926, 2012.
17. Hou, X., Harel, J. and Koch, C., 2012. Image signature: Highlighting sparse salient regions. *IEEE transactions on pattern analysis and machine intelligence,* 34(1), pp.194-201.
18. Li, J., M. D. Levine, X. An, X. Xu, and H. He (2013). Visual saliency based on scale-space analysis in the frequency domain. *TPAMI,* 35(4), 996-1010.
19. Zhong, S. H., Liu, Y., Ren, F., Zhang, J., &Ren, T. (2013). Video Saliency Detection via Dynamic Consistent Spatio-Temporal Attention Modelling. *In AAAI,* pp. 1063-1069.
20. Yang, C., Zhang, L., Lu, H., Ruan, X., & Yang, M. H. (2013). Saliency detection via graph-based manifold ranking. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* pp. 3166-3173.
21. N. İmamoğlu, W. Lin, and Y. Fang, "A Saliency Detection Model Using Low-Level Features Based on Wavelet Transform," Multimedia, *IEEE Transactions,* vol. 15, no. 1, pp. 96 - 105, 2013.
22. Zou, W., Liu, Z., Kpalma, K., Ronsin, J., Zhao, Y., &Komodakis, N. (2015). Unsupervised joint salient region detection and object segmentation. *IEEE Transactions on Image Processing,* 24(11), 3858-3873.
23. Tong, N., Lu, H., Ruan, X., & Yang, M. H. (2015). Salient object detection via bootstrap learning. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* pp. 1884-1892.
24. Wang, W., Shen, J., Sun, H. and Shao, L., 2018. Video co-saliency guided co-segmentation. *IEEE Transactions on Circuits and Systems for Video Technology,* 28(8), pp.1727-1736.
25. Chen, Y., Zou, W., Tang, Y., Li, X., Xu, C., &Komodakis, N. (2018). SCOM: Spatiotemporal Constrained Optimization for Salient Object Detection. *IEEE Transactions on Image Processing,* 27(7), 3345-3357.

26. Eitel, A., Springenberg, J. T., Spinello, L., Riedmiller, M., &Burgard, W. (2015). Multimodal deep learning for robust RGB-D object recognition. *In IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS),*pp. 681-687.
27. Zhu, L., Chen, Y., Yuille, A., & Freeman, W. (2010). Latent hierarchical structural learning for object detection. *In IEEE Computer Society Conference on Computer Vision and Pattern Recognition,* pp. 1062-1069.
28. Han, J., Zhang, D., Cheng, G., Liu, N., &Xu, D. (2018). Advanced deep-learning techniques for salient and category-specific object detection: a survey. *IEEE Signal Processing Magazine,* 35(1), 84-100.
29. Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2016). Region-based convolutional networks for accurate object detection and segmentation. *IEEE transactions on pattern analysis and machine intelligence,* 38(1), 142-158.
30. Chen, X., Kundu, K., Zhu, Y., Berneshawi, A. G., Ma, H., Fidler, S., &Urtasun, R. (2015). 3d object proposals for accurate object class detection. *In Advances in Neural Information Processing Systems, pp*. 424-432.
31. Ding, X., Luo, Y., Li, Q., Cheng, Y., Cai, G., Munnoch, R., & Wang, B. (2018). Prior knowledge-based deep learning method for indoor object recognition and application. *Systems Science & Control Engineering,* 6(1), 249-257.
32. Delforouzi, A., Pamarthi, B., &Grzegorzek, M. (2018). Training-Based Methods for Comparison of Object Detection Methods for Visual Object Tracking. Sensors, 18(11), 3994.
33. Tian, B., Li, L., Qu, Y., & Yan, L. (2017). Video object detection for tractability with deep learning method. In 2017 *Fifth International Conference on Advanced Cloud and Big Data (CBD)*,pp. 397-401.
34. Zhang, Y., Li, X., Zhang, Z., Wu, F., & Zhao, L. (2015). Deep learning driven blockwise moving object detection with binary scene modeling. *Neurocomputing,* 168, 454-463.
35. Zhao, Z. Q., Zheng, P., Xu, S. T., & Wu, X. (2019). Object detection with deep learning: A review. *IEEE transactions on neural networks and learning systems,* vol. 14, no. 8,pp.1-21.

AUTHORS PROFILE

M. Indirani, M.E., pursued B.E. Computer Science and Engineering in the year 2000 from Bharathiar University, Coimbatore and M.E. Computer and Communication Engineering in the year 2007 from Anna University, Chennai. She is pursuing Ph.D., in Anna University, Chennai since 2016 and currently working as an Assistant Professor in the Department of Information Technology in Hindusthan College of Engineering and Technology, Coimbatore since 2007. She has published 4 international Journals and Conferences. Her main research focuses on Image Processing. She has 16 years of teaching experience.

Dr.S. Shankar is currently working as Professor and Head of the Department in Computer Science and Engineering at Hindusthan College of Engineering and Technology, Coimbatore. Prior to this, he was working as a Professor and Department Head in Information Technology at Sri Krishna College of Engineering and Technology, Coimbatore. He has completed his PhD on Data Mining in the year 2012 at Anna University. He has more than 17 years of experience in teaching and research and added to his credit 20 International Journal Publications with good impact factor and citations. He has written and published a book on Python in Mcgraw Hill. He is a recognized supervisor in Anna University Chennai and currently guiding 10 PhD scholars. He is a Wipro certified faculty.