# Chapter 2

# Theoretical Foundation and Literature Survey

This chapter presents the survey of state-of-the-art methods in salient object detection and image annotation. Section 2.1 introduces the topic of salient object detection and automatic image annotation. Section 2.2 presents the literature survey in the field of salient object detection, image annotation, and colon tumor localization. Issues and challenges of both the fields are discussed in Section 2.3. Section 2.4 lists the databases used to implement the proposed models. The proposed models are evaluated based on the metrics given in Section 2.5. Section 2.6 concludes the chapter.

## 2.1 Background

This section defines and briefly describe salient object detection and automatic image annotation.

### 2.1.1 Salient Object Detection

Salient object detection locates salient objects in an image. The algorithms developed for salient object detection try to imitate human visual mechanism. Some models first try to locate salient regions in an image and then extract out the salient objects. Other models, try to pixel-wise classify the image region as salient/non-salient. A variety of methods are developed by the researchers which are discussed in Section 2.2.1.

### 2.1.2 Automatic Image Annotation

Automatic image annotation is the process of assigning tags to the images specifying the objects present in the image and/or associated context. It is primarily a multi-label classification task in which features are extracted from the image. The feature vector is then used to assign keywords. It has several applications like content-based image retrieval, image database management, automatic image captioning, and medical image annotation. A general model for image annotation can be explained in FIGURE 2.1. In the proposed model, first, salient objects are extracted from the

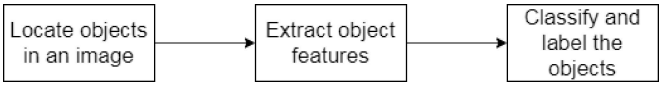


FIGURE 2.1: Block diagram for a general automatic image annotation model.

image, and then texture and color features are extracted, and then multi-label classification is performed. Researchers for automatic image annotation have proposed several methods. A survey of those methods is given Section 2.2.2.

## 2.2 Literature Review

This section reviews some of the state-of-the-art techniques for salient object detection and image annotation. Section 2.2.1 discusses the work in the field of salient object detection, and Section 2.2.2 explores methods for image annotation. Section 2.2.3 shows some of the research done in the field of colon tumor localization.

### 2.2.1 Salient Object Detection

This subsection discusses the various techniques used for salient object detection. The techniques are classified into statistical, machine learning, and deep learning methods. Section 2.2.1.1 provides the review for statistical and machine learning-based methods and Section 2.2.1.2 reviews deep learning methods.

#### 2.2.1.1 Statistical and Machine Learning Methods

Various methods and models are developed by researchers who use the statistical properties of an image to locate the salient object. Most of the researchers use ensemble feature techniques because any particular handcrafted feature is not enough to locate the salient object. The problem of salient object detection has been approached in different ways. Some of the models view the problem as a segmentation framework. A few models try to locate objects in the image and then classify them for saliency. Prior based models are mostly found because they simplify the location of the salient object to some extent. They also correspond more to the human visual attention system. Various graph-based techniques have also been proposed for finding the salient object. Sparse representations are also used in detecting salient objects. Researchers have also used the matrix decomposition model where the feature matrix is decomposed to reflect salient and non-salient regions. Some of the models generate a hierarchical representation of image saliency and produce a result by fusion of the output at different levels.

For *segmentation based approaches*, the procedure is to localize the salient object and then apply segmentation. For localization eye-fixation models [2], edges, texture [12] and color [19] features are used.

*Superpixel algorithms* allow shallow segmentation of images. Borji [3] uses superpixels to get an idea about the complexity of the image. Simple images have less number of superpixels than a complex image. Seo *et al.* [20] uses superpixels to develop a bipartite dictionary assigning one of the two labels - salient or background.

The aim is to maximize inter-class reconstruction error and minimize inter-class reconstruction error.

*Color* is the most widely used feature for salient object detection. Though, it is not capable of finding salient objects on its own. It is used as a supplement in most of the algorithms. Zhang *et al.* [21] utilize color contrast with boosting Harris to highlight salient points. Zhu *et al.* [9] focus on finding the most striking color of the scene to locate the salient object. Zhang *et al.* [22] use the color distribution-based model to refine the coarse saliency map generated by the symmetric surround model. Abinash [17] presents a model of salient object detection based on color channels. Hu *et al.* [23] use color distribution to measure the similarity of superpixels. Kim *et al.* use prior maps generated from various combinations of Red, Green and Blue (RGB) color components [24]. Vu and Chandler [25] use color distance as a feature for salient object detection. Kapoor *et al.* [26] also use features generated from color components of images [26].

Some of the researchers emphasize using *focus* as an important prior for finding the salient object. The assumption is that salient objects in an image are always in focus. Thus, non-focused objects can be marked as background. Sun *et al.* [4] develop a sparse dictionary based on focus prior map. Zeng and Tsai [27] also use this assumption together with intensity differences to locate the salient object.

Another method to locate the salient object is to separate *background and foreground.* When the image is segmented as per these criteria, foreground regions are then processed to extract the salient object. Huang *et al.* [5] select foreground seeds

based on surroundedness cue. Further, geodesic refinement is applied to the foreground region for salient object detection. Wang and Liu [28] separate background using geometrical interpretation. Seo and Yoo [29] use edges to generate a convex hull for creating a background and foreground dictionary. Reconstruction error optimization is done to extract the salient object in the foreground. Kong *et al.* [30] use images annotated with foreground and background. They transfer this information to the over segmented input images to locate salient objects.

*Objectness proposals* are also an interesting method to find salient objects. In these methods, the models generally first try to find as many complete objects they can find in an image and, after that, finalize a salient object among them. Zhang *et al.* [15] use objectness map and fuse it with the saliency map generated by the Markov chain using background seeds. Wang *et al.* [31] use objectness cue with superpixels. They localize the salient object using contrast, center, and background priors. Wang and Yang [32] propose the use of saliency-guided object proposals. Zhou *et al.* [33] uses fuzzy theory to integrate maps obtained from regional saliency measure and objectness proposals. Li and Lun [34] generate object proposals and differentiate them as foreground and background using the low-rank decomposition method. Srivatsa and Babu [35] estimate foreground from object proposals using foreground connectivity cue.

*Boundary* is an important prior used by various models. The assumption is that generally, salient object lies away from the boundary. The boundary marks the

background region of an image. Using this assumption, the background can be separated from the image, and this salient object can be extracted. Abkenar *et al.* [6] use distribution-based boundary contrast map. The graph representation of the image is used to compute the connectivity of the image regions to the image boundary as well as to their local neighbors and the image foreground. Connectivity maps obtained are then fused with the boundary contrast map. Huo *et al.* [36] use boundary homogeneity as a prior to estimate the background and eventually use it to mark foreground regions of the image. Wang *et al.* [37] estimate foreground regions using boundary connectivity cues.

Another approach of locating salient objects is to use *graphs*. In using graphs, initially, some seed values are assigned that become the nodes, and edges connect these nodes. The nodes and edges are updated to converge to find the salient object. Nouri *et al.* [7] use contrast as a feature to generate graph. A threshold for edge weight is used to eliminate non-salient edges. Li *et al.* [38] generates a hypergraph using information from a pixel's similarity to its neighborhood and its dissimilarity from the background. Wu *et al.* [39] proposes a graph-based method in which seeds are generated from the background cues. Wang and Lv [10] use eye fixation prediction as a prior to apply graph-cut on an input image.

*Contour-based models* have been primarily used for segmentation. Infused with saliency information, these have been used to detect salient objects. The advantage of using them is getting neat object boundaries. Du and Chen [11] use patch rarities to form the object contours. The patch rarities are computed by using the random

forest. Liu *et al.* [40] use completeness and closure measure as saliency cues to form the contour of the object.

Some researchers use *sparse representation* to highlight the difference between salient and non-salient regions. Wang *et al.* [8] use the visual attraction level of foreground and background regions to generate the sparse representation of the image. Reconstruction errors are used as a measure of saliency indicators. Yan *et al.* [41] generate sparse representation of images from features extracted at local, global and semantic level. Zhang *et al.* [42] build a nonconvex structure matrix decomposition model. They explore the relationship between each superpixel to make the salient object highlighted consistently. Laplacian regularization is used to increase the distance between salient regions and non-salient regions in feature space. Peng *et al.* [43] use structured matrix decomposition model with two structural regularizations.

Using *spatial priors* and location information from an image also helps to locate the salient object. Zhang *et al.* [18] use spatial priors in addition to color and central bias to construct the graph. The graph is built by connecting nodes that are spatially close in an image. The edge weights are provided based on color similarity and spatial proximity. Wang *et al.* [44] use spatial relationship to develop a geodesic weighted Bayesian model. Xiang *et al.* [45] exploit a location-aware strategy to identify the optimal saliency map across multiple scales of the image.

Some researchers rely upon various *low level features* to find the salient object. Xu *et al.* [46] propose a model that uses distinctive features like lightness, sharpness, and magnitude. Hendrawati *et al.* [47] use color features, central moments, texture and

shape features - energy, entropy, contrast, and homogeneity and invariant moments for salient object detection.

*Contrast* is an important feature related to salient object detection. Salient objects generally have high contrast than the background object. Nikhila and Rawat [48] use local contrast and compactness visual cues. Xiao and Zhou [49] combine fine-grained contrast prior to rough-grained object consistency. They use focusness prior for guiding the contrast map. Yang *et al.* [50] use contrast, center and smoothness priors for salient object detection. Fu *et al.* [51] use contrast, center priors and surroundedness cue.

Researchers also generate various *hierarchical representations* of image to extract information at various levels. Islam *et al.* [52] uses a hierarchical representation of relative saliency and perform stage-wise refinement to locate the salient object. Xiao and Wang [53] apply hierarchical Boolean map approach to generate attention maps.

A comprehensive analysis of few methods is given in Table 2.1.

TABLE 2.1: Comprehensive details of few existing methods for salient object detection using statistics and machine learning.

| **S.No.** | **Paper** | **Techniques** | **Result** |
|---|---|---|---|
| 1 | Multi-scale analysis of color and texture (Tang *et al.* 2011) [19] | Segmentation at 3 scales, texture and color features | **Precision/Recall/F1**: MSRA: 0.80/0.79/0.80 |
| 2 | A Dataset and a Baseline Model (Borji 2015) [3] | Relation between eye-fixation and salient object, Superpixel | **F-Measure/AUC**: MSRA-5K: 0.727/0.781 Bruce-A: 0.308/0.780 Judd-A: 0.551/0.662 |
| 3 | Cognitive Neuroscience (Zhu *et al.* 2017) [9] | Depth, luminance, center-bias | **MAE**: RGBD1: 0.1065 RGBD2: 0.1007 |
| 4 | Background Model (Zhang *et al.* 2018) [22] | Symmetric surround model, color-distribution based mode, Gauss filter, background model | **AUC** MSRA: 0.9671 |
| 5 | Graph theory (Abinash 2018) [17] | Color channels, neighborhood information | **JC/DC/MSE**: Berkeley: 0.94/0.97/0.03 Corel: 0.92/0.96/0.04 |

| S.No. | Paper | Techniques | Result |
|---|---|---|---|
| 6 | Compactness measurement (Kim *et al.* 2013) [24] | RGB color prior maps, compactness measure | **Precision/Recall/F1** MSRA: 0.87/0.80/0.85 |
| 7 | Fusing Foreground and Background Priors (Huang *et al.* 2018) [5] | Foreground and background seeds, surroundedness cue, geodesic refinement | **MAE/AUC**: ASD: 0.0596/0.9446 PASCAL-S: 0.1868/ 0.6917 |
| 8 | Joint Latent Space Embedding (Kong *et al.* 2018) [30] | Relationship between objects of same class, annotated reference image, joint latent embedding of superpixels | **F-Measure/AUC**: ECSSD: 0.7243/0.9350 HKU-IS: 0.6903/0.9349 SOD: 0.6072/0.8524 DUT-OMRON: 0.5178/ 0.8951 |
| 9 | Two-stage absorbing Markov chain (Zhang *et al.* 2017) [15] | Random walk on absorbing Markov chain, background seeds, objecntess map | **AUC/MAE**: ECSSD: 0.917/0.167 MSRA5K: 0.959/0.108 MSRA10K: 0.964/0.110 SED2: 0.897/0.159 |
| 10 | Saliency-guided object proposal (Wang *et al.* 2016) [32] | Object proposals | **F-Measure**: MSRA: 0.8627 PASCAL-S: 0.6846 |
| 11 | Fuzzy Theory and Object-Level Enhancement (Zhou *et al.* 2019) [33] | Multiple prior maps, object proposals | **MAE**: ASD: 0.0439 ECSSD: 0.1316 PASCAL-S: 0.1710 |
| 12 | Objectness measure (Srivatsa *et al.* 2015) [35] | Foreground connectivity | **MAE**: MSRA-1000: 0.064 CSD: 0.132 |
| 13 | Graph-Based Background and Foreground Connectivity Cues (Abkenar *et al.* 2019) [6] | Distribution-based boundary contrast map, graph | **F-Measure/MAE**: DUT-OMRON: 0.6064/ 0.1202 HKU-IS: 0.7306/0.1151 |
| 14 | Contextual Hypergraph Modeling (Li *et al.* 2013) [38] | Hyper graph, center-versus-surround contextual contrast analysis, Support Vector Machine (SVM) | **VOC Overlap Score**: MSRA: 0.77 SOD: 0.40 SED: 0.52 Imgsal-50: 0.69 |
| 15 | Reliable boundary seeds and saliency refinement (Wu *et al.* 2019) [39] | Graph, superpixels, background seeds, background and foreground-based map | **Precision/F-Measure /AUC**: ASD: 0.934/0.912/0.863 ECSSD: 0.818/0.734/ 0.792 PASCAL-S: 0.682/0.585 /0.769 THUR-15K: 0.600/0.570 /0.822 DUT-OMRON: 0.593 /0.566/0.827 |
| 16 | Fixation priori (Wang *et al.* 2016) [10] | Eye-fixation map | **AUC**: PASCAL-S: 0.867 MSRA-1000: 0.985 |
| 17 | Hierarchical Contour Closure (Liu *et al.* 2017) [40] | Closure completeness, closure reliability, hierarchical segmentation | **F-Measure/MAE/ AUC**: MSRA10K: 0.846/0.097 /0.964 PASCAL-S: 0.631/0.185 /0.846 DUT-OMRON: 0.571/ 0.115/0.889 |
| 18 | Nonconvex Structured Matrix Decomposition (Zhang *et al.* 2017) [42] | L1 norm of logistic function on the singular values of a matrix to approximate rank function, relationship between each superpixel, Laplacian regularization | **MAE** ECSSD: 0.178 |

| S.No. | Paper | Techniques | Result |
|---|---|---|---|
| 19 | Structured Matrix Decomposition (Peng *et al.* 2017) [43] | Image structure, Laplacian regularization | **F-Measure/OR/ AUC/MAE**: MSRA10K: 0.704/0.741 /0.847/0.104 DUT-OMRON: 0.424 /0.441/0.809/0.166 iCoSeg: 0.611/0.598/ 0.822/0.138 SOD: 0.456/0.419/0.733 /0.233 ECSSD: 0.517/0.523/ 0.775/0.227 |
| 20 | Focusness guided (Xiao *et al.* 2017) [49] | Contrast-prior, object consistency, focusness prior | **MAE/F-Measure**: MSRA10K: 0.1107/0.761 SED2: 0.1118/0.7595 PASCAL-S: 0.1935/ 0.5597 ECSSD: 0.2206/0.5928 |

The eye-fixation based models [3, 10] rely heavily on the correct prediction of eye-fixation. Color-based models [17, 19, 22, 24, 33, 39] fail when the foreground and background have similar colors. The model proposed in [30] requires a longer run-time for each image than most of the state-of-the-art algorithms. Since the methods in [6,9,35] uses boundary prior, salient objects located near the boundary are either not detected or are incompletely detected. Structured matrix decomposition [43] highlights distinct elements present in the background even though they are non-salient. It is also noticeable that contour-based model [40], models using both foreground and background maps [5], objectness based models [15,32] and contrast based models [38, 49] provide excellent results.

#### 2.2.1.2 Deep Learning Methods

The deep learning methods came to play when Krizhevsky's [54] 8-layer network won the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) in 2012 with a

very high margin. Since then, the development of various networks from 8 to 350 layers has only propelled the use of deep networks for salient object detection. Within a short span, a variety of architectures have been developed. Simple CNNs have been evolved into Fast R-CNN [55], and Mask R-CNN [56]. YOLO [57] networks are developed for faster implementation of deep learning models. A variety of network models - AlexNet, ResNet [58], GoogleNet [59] and VGGnets [60] are available for implementing the algorithms. A newer variety of these networks - General Adversarial Networks [61] - are brought into play. Here one network produces an image after training. The work of the discriminator network is to measure how much the generated image corresponds to the real image. The auto encoder-decoder network is also widely used to complement information loss at various levels. This section discusses some of the models of salient object detection based on deep learning.

Hou *et al.* [16] introduces short connections to the skip-layer structures within the Holistically-Nested Edge Detector (HED) [62] architecture. They use multi-level and multi-scale features extracted from Fully Convolutional Neural Networks (FCNs). Wang *et al.* [63] employ multi-scale mask-based Fast R-CNN. The regions are segmented using edge-preserved methods. Low-level contrast and backgroundness prior are used for improving the result. Liu *et al.* [64] work by using pixel-level and region-level predictions. Their model uses modified dilated convolution layers and short connections. The authors implement a superpixel based manifold learning algorithm for obtaining a better boundary of the salient object. Post-refinement is done by DenseCRF [65]. Wang *et al.* [66] jointly learn to segment salient object

masks and detect salient object boundaries. Their model makes use of the focal loss to facilitate the learning of the hard boundary pixels. Wang *et al.* [67] utilize edge-preserved neural network based on Fast R-CNN framework. They make use of multi-scale spatial context.

Han *et al.* [68] first model the background and then separate salient objects from the background. They employ stacked denoising autoencoders with deep learning architectures to model the background where underlying patterns are explored, and more powerful representations of data are learned in an unsupervised and bottom-up manner. The perspective of their model is to measure reconstruction residuals of deep autoencoders. The model of Zhang *et al.* [69] operates on a symmetrical, fully convolutional network to effectively learn complementary saliency features under the guidance of lossless feature reflection. The location information, together with contextual and semantic information, of salient objects, is jointly utilized to supervise the proposed network for more accurate saliency predictions. The aim is to optimize the weighted structural loss function. Zhang *et al.* [70] apply unsupervised saliency and take normalized color images as inputs. Results from deep FCNN and Robust Background Detection (RBD) are concatenated to feed into a shallow network to map the concatenated feature maps to saliency maps. They also implement a superpixel level saliency map at a multi-scale. Tang and Wu's [71] cascaded convolutional neural networks (CNNs) based method is proposed to learn this structural information via adversarial learning implicitly. Generator G consists of an encoder-decoder network for global saliency estimation and a deep residual network for local saliency

refinement. Discriminator D is then designed to distinguish the real salient maps (i.e., ground truths) from the fake ones produced by G. Based on this; an adversarial loss is introduced to optimize G. Xiao *et al.* [72] propose dense short- and long-range connections that effectively integrate multi-scale features. They diagnose which regions of an input image are distracting and harmful for saliency prediction. Their method is based on distraction mining approach.

Han *et al.* [73] convert edge convolution constraint to a modified U-Net. They fuse the features of different layers to reduce the loss of information. They also provide a new loss function based on image convolution, which adds an L1 constraint to the edge information of saliency map and ground-truth. Li and Yu [74] utilize hybrid contrast-oriented deep neural networks. They make use of the fully convolutional stream for dense prediction and a segment-level spatial pooling stream for sparse saliency inference. They propose an attentional module that learns weight maps for fusing the two saliency predictions from these two streams. Their model has a customized, fully connected conditional random field model to improve spatial coherence and contour positioning. Guan *et al.* [75] employ global contextual information along with the low-level edge features. Their model works on hierarchical boundary information and edge contours. Zhang and Zu's [76] model works on multi-scaled feature aggression. Tang *et al.* [77] proposes recurrently aggregating and refining features in a cross-level and spatial attention-aware manner. Their spatial attention-aware module suppresses the non-salient regions and highlights salient objects.

Wang *et al.* [78] utilize recurrent fully convolutional networks to work on semantic segmentation data. Zhuge *et al.* [79] integrate multi-level convolutional features recurrently with the guidance of object boundary information. Fu *et al.* [80] employ fully convolutional network augmented with segmentation hypotheses. Najibi *et al.* [81] convert the ground-truth rectangular boxes to Gaussian distributions. Zheng *et al.* [82] operate on annular feature pyramid network to augment information flow and enhance feature hierarchy.

In Fatemi *et al.*'s [83] model color space is used to extract the feature and from self-encoder to remove the noise in the property matrix. Hsu *et al.* [84] apply a weakly supervised approach to object saliency detection where only image-level labels, indicating the presence or absence of a target object in an image are used. Background prior and superpixel- and object proposal-based evidence are also utilized. Wang *et al.* [85] proposes the Attentive Saliency Network (ASNet) that learns to detect salient objects from fixation maps. The network mimics human visual attention mechanisms. The model of Zhang *et al.* [86] integrates multi-level feature maps into multiple resolutions. The model simultaneously incorporates coarse semantics and fine details. Edge-aware feature maps in low-level layers and the predicted results of low-resolution features are recursively embedded into the learning framework. Qi *et al.* [87] use multi-scale contextual information with dynamic routing. They aggregate multi-level features by using multi-crossed skip-layer connections.

These are the few models that use deep learning for their implementation. They have achieved outstanding results when compared to classical methods. Still, newer

models are developed which can surpass human's capability in the presence of highly cluttered background data.

A comprehensive analysis of few methods is given in Table 2.2.

TABLE 2.2: Comprehensive details of few existing methods for salient object detection using deep networks.

| S.No. | Paper | Techniques | Result |
|---|---|---|---|
| 1 | Deeply Supervised with Short Connections (Hou *et al.* 2019) [16] | Short connections, skip-layer structures, HED architecture, multi-level and multi-scale features, FCNs | **F-Measure/MAE**: MSRA-B: 0.927/0.028 ECSSD: 0.915/0.052 HKU-IS: 0.913/0.039 PASCAL-S: 0.83/0.080 SOD: 0.842/0.118 |
| 2 | Fast R-CNN and low-level cues (Wang *et al.* 2016) [63] | Multi-scale mask-based Fast R-CNN, edge-preserved methods, low-level contrast and backgroundness prior, high-level semantic, edge-based propagation method | **F-Measure**: ECCSD: 0.841 DUT-OMRON: 0.768 JuddDB: 0.516 |
| 3 | Pixel Meets Region (Liu *et al.* 2018) [64] | Pixel-level and region-level predictions, modified dilated convolution layers, short connections, FCN, pixel-wise salient classifier, superpixel based manifold learning algorithm, DenseCRF | **F-Measure/MAE**: ECSSD: 0.902/0.066 PACSAL-S: 0.828/0.103 |
| 4 | Focal Boundary Guided (Wang *et al.* 2019) [66] | Object boundaries, focal loss | **F-Measure/MAE/ SSM/ODS/OIS**: MSRA-B: 0.933/0.034/ 0.911/0.770/0.775 ECSSD: 0.931/0.045/ 0.900/0.761/0.765 HKU-IS: 0.925/0.034/ 0.776/0.780/0.784 DUT-OMRON: 0.776/ 0.066/0.784/0.584/0.593 SOD: 0.853/0.092/0.793 /0.639/0.644 |
| 5 | Edge Preserving and Multi-Scale Contextual Neural Network (Wang *et al.* 2018) [67] | Edge-preserved neural network, Fast R-CNN framework, multi-scale spatial context | **F-Measure/AUC**: JuddDB: 0.556/0.545 DUT-OMRON: 0.789/ 0.803 THUR15K: 0.761/0.779 SED2: 0.893/0.918 ECSSD: 0.893/0.937 PASCAL-S: 0.822/0.856 |

| S.No. | Paper | Techniques | Result |
|---|---|---|---|
| 6 | Lossless Feature Reflection and Weighted Structural Loss (Zhang *et al.* 2019) [69] | Location information, contextual and semantic information, weighted structural loss function | **F-Measure/MAE/ SSM** MSRA10K: 0.718/ 0.0643/0.811 DUT-OMRON: 0.742/ 0.0622/0.853 ECSSD: 0.911/0.0421/ 0.916 HKU-IS: 0.906/0.0357/ 0.914 PASCAL-S: 0.813/ 0.0732/0.851 SED: 0.925/0.0442/ 0.913 SOD: 0.822/0.1012/ 0.804 |
| 7 | Integrated deep and shallow networks (Zhang *et al.* 2017) [70] | Results of unsupervised saliency, normalized color images, high-level semantic cues, superpixel level saliency map at multi-scale. | **MAE**: ECSSD: 0.0610 DUT-OMRON: 0.0648 SED: 0.0533 PASCAL-S: 0.0810 SOD: 0.1002 HKU-IS: 0.0486 THUR: 0.0704 |
| 8 | Cascaded Convolutional Neural Networks and Adversarial Learning (Tang *et al.* 2019) [71] | Conditional generative adversarial networks | **F-Measure/SSM/ MAE**: SED: 0.9140/0.9070/ 0.0372 ECSSD: 0.8868/0.8952/ 0.0415 PASCAL-S: 0.7403/ 0.7568/0.0709 HKU-IS: 0.8794/0.8974/ 0.0329 SOD: 0.7412/0.8026/ 0.0827 DUT-OMRON: 0.7243/ 0.8299/0.0418 |
| 9 | Dense Connections and Distraction Diagnosis (Xiao *et al.* 2018) [72] | Dense short-and long-range connections, multiscale features, contexts at multiple levels | **F-Measure/MAE**: DUT-OMRON: 0.816/ 0.147 ECSSD: 0.937/0.058 HKU-IS: 0.933/0.042 MSRA-B: 0.944/0.032 PASCAL-S: 0.869/0.071 SOD: 0.875/ 0.099 |
| 10 | Contrast-Oriented Deep Neural Networks (Li *et al.* 2018) [74] | Fully convolutional stream, segment-level spatial pooling stream, attentional module, fine -tuning pretrained baseline models, customized fully connected conditional random field model | **F-Measure/MAE**: MSRA-B: 0.931/0.042 HKU-IS: 0.913/ 0.041 PASCAL-S: 0.857/0.092 DUT-OMRON: 0.811/ 0.064 ECSSD: 0.925/0.058 SOD: 0.857/ 0.120 |
| 11 | Edge-Aware Convolution Neural Network (Guan *et al.* 2019) [75] | Global contextual information, low-level edge features, holistically-nested edge detection (HED) model, hierarchical boundary information, edge contours, multi-scale pyramid-based supervision | **F-Measure/MAE**: DUT-OMRON: 0.785/ 0.047 ECSSD: 0.901/0.044 THUR15K: 0.717/0.074 HKU-IS: 0.879/0.036 PASCAL-S: 0.812/0.075 |

| S.No. | Paper | Techniques | Result |
|---|---|---|---|
| 12 | Multi-scale feature aggression (Zhang *et al.* 2019) [76] | Four kinds of features in various resolution levels | **F-Measure**: SOD: 0.8068 ECSSD: 0.8713 DUT-OMRON: 0.6938 THUR15K: 0.7010 |
| 13 | Recurrently Aggregating Spatial Attention Weighted Cross-Level Deep Features (Tang *et al.* 2019) [77] | Features integrated from multiple layers, spatial attention-aware module | **F-Measure/MAE**: ECSSD: 0.923/0.043 DUT-OMRON: 0.791/ 0.051 SOD: 0.825/0.118 PASCAL-S: 0.845/0.105 HKU-IS: 0.923/0.031 |
| 14 | Recurrent Fully Convolutional Networks (Wang *et al.* 2019) [78] | Recurrent architecture, semantic segmentation data | **F-Measure/MAE**: HKU-IS: 0.8564/0.0547 ECSSD: 0.8713/0.0668 PASCAL-S: 0.7784/ 0.1049 SED1: 0.8811/0.0750 |
| 15 | Boundary-Guided Feature Aggregation Network (Zhuge *et al.* 2018) [79] | Multilevel convolutional features, object boundary information, attention-based feature fusion module | **F-Measure/MAE**: ECSSD: 0.882/0.051 DUT-OMRON: 0.721/ 0.060 HKU-IS: 0.887/0.043 |
| 16 | Deep Segmentation Assisted Refinement Network (Fu *et al.* 2019) [80] | Segmentation hypotheses, edge-aware saliency maps | **F-Measure**: ASD:0.9339 ECSSD: 0.9197 DUT-OMRON: 0.7915 PASCAL-S: 0.8441 SED2: 0.8730 HKU-IS: 0.8925 |
| 17 | Real-Time Unconstrained (Najibi *et al.* 2018) [81] | Gaussian distributions, ROI, covariance | **Precision/Recall**: MSO: 81.8/85.5 MSRA: 90.1/87.1 DUT-O: 80/44.6 PASCAL-S: 82.1/77.8 |
| 18 | Annular feature pyramid network (Zheng *et al.* 2019) [82] | Annular feature pyramid network | **F-Measure/MAE**: HKU-IS: 0.928/0.032 DUT-OMRON: 0.799/ 0.057 PASCAL-S: 0.871/0.073 ECSSD: 0.934/0.038 SOD: 0.869/0.101 |
| 19 | Weakly Supervised Learning A Classifier-Driven Map Generator (Hsu *et al.* 2019) [84] | Image-level labels, image-level classifier, pixel-level map generator, background prior, superpixel, object proposal | **Precision**: Graz-02: 84.1 PASCAL-VOC 07: 47.2 PASCAL-VOC 12: 56.8 |
| 20 | Inferring Human Fixations (Wang *et al.* 2019) [85] | Fixation map | **F-Measure/MAE**: ECSSD: 0.928/0.043 HKU-IS: 0.920/0.035 PASCAL-S: 0.857/0.072 SOD: 0.835/0.115 |

Heterogeneous objects create problems for models proposed in [16, 69, 71]. The models proposed in [69,70,84] cannot detect salient objects when they are similar to the background. The model proposed in [16] fails in case of complex backgrounds, images with low contrast, and detection of transparent objects. [67] also fails when

the image has low contrast and when there is no clear demarcation between object and background. In the presence of multiple distinctive objects in the image, the model proposed in [69] fails to detect the primary salient object. It also leads to wrong results when the objects are tiny. Shadows and complex background makes the processing of [71] unsuccessful. Occlusion and texture difference in disconnected objects are limitations of the model proposed in [72]. The model proposed in [78] relies highly on initial estimations and has a computational overhead because of the recurrent network architecture. Similarly, the output of the model proposed in [80] depends on the location accuracy of coarse saliency maps. The model proposed in [84] does not support contextual information.

### 2.2.2 Image Annotation

There are various machine learning techniques applied to image annotation. The primary method is to segment the image, extract features from the image, and then perform multi-label classification. Some of the methods have a decision tree for classifiers. The fusion of features - high-level and low-level - obtained from different ways enriches features derived from the image and can help to provide better tags. Some of the methods use a feedback mechanism to improve the results. Semi-supervised methods are also quite useful for image annotation. Some researchers also use neural networks with a single hidden layer for generating annotations. For deep learning methods, some researchers use the features extracted from CNN. Semi-supervised deep learning methods are also used. Recently General Adversarial Networks have

become the latest trend to be used in image annotation. The subsections below discuss some of the machine learning and deep learning methods for image annotation.

#### 2.2.2.1 Machine Learning Methods

Hu and Chen [88] utilize neighborhood rough sets for image annotation. They construct a global rough set model. They try to establish a semantic association between words and images. Sun and Loparo [89] build a context-aware image annotation framework. Ujjwal *et al.* [90] build an interactive Assistive Annotation System that helps annotators by marking salient regions in an image. The salient regions can be further refined manually. Wang *et al.* [91] learn a distance metric based on structural SVM. A collaborative label propagation algorithm is used for modeling the correlation between class labels. Wojnar and Pinheiro [92] extract feature using a Fast-Hessian detector. They also use Speeded Up Robust Features (SURF) descriptor and an SVM classifier for generating image annotations.

Renuse and Bogiri [93] view the image annotation problem as that of multi-label learning and multi-keyword extraction. They build a C4.5 classifier with the help of extracted features. Lee *et al.* [94] also propose a decision tree-based model for image annotation. They extract wavelet-based descriptors. These descriptors are applied to decision trees to build a random forest.

Zhang and Lou [95] use the fusion of color and texture features. They use regression and genetic algorithms for producing sentiment-based image annotation. For regression, they use the least squares support vector machine. For the optimization

of regression, particle swarm optimization is used. Similarly, Chan *et al.* [96] also use an ensemble feature method. They extract color moment, edge oriented histogram, and local binary pattern. They use an SVM classifier, and cosine similarity is used as the distance metric for a test image.

Budikova *et al.* [97] use annotation relevance feedback to improve the quality of the result. Likewise, Zhong and Ma [98] relies on a relevance feedback algorithm for improving the accuracy of annotation results. They also emphasize the order of annotated tags, which they accomplish using a semantic co-occurrence strategy. Sun *et al.* [99] use an improved Markov model for annotation. The annotations are clustered based on semantic relationships. Relevance feedback from people's cognitive learning habits is used for improving the result.

Wu *et al.* [100] annotated medical images using multi-label active learning. The method was based on low-rank modeling. Jing *et al.* [101] use a supervised non-negative matrix factorization-based framework for image annotation. Schaefer and Stuttard [102] propose the idea that database visualization and browsing can prove effective for image annotation. Bhargava *et al.* [103] point out that selection of objects to be annotated is important for image annotation.

Bi and Yin [104] use a graph-based semi-supervised learning method. Zhu *et al.* [105] use a multi-view semi-supervised learning scheme. Each view-specific classifier is trained and re-trained with labeled and pseudo-labeled data. For assigning annotations, the maximum vote entropy principle is followed.

Kulkarni and Kulkarni [106] propose the use of neural networks and fuzzy logic for

image annotation. Neural networks are used for classification, and fuzzy logic is used for assigning tags in natural language. Li *et al.* [107] performs data grouping and uses a single-layer feed-forward neural network for providing tags to images.

Akhilesh and Sedamkar [108] segment the image using a mean shift algorithm. They employ an iterative, probabilistic, and meta-heuristic method for finding an optimal feature subset. Wen and Li [109] use the Jseg algorithm to segment the images. Further, with multi-SVM, they establish relationships between low-level features and attribute concepts. Guo *et al.* [110] use a combined framework for image segmentation and image annotation. Their model is based on probabilistic latent semantic analysis. Mihai and Stanescu [111] segment the image regions based on hexagonal structures. For the regions thus obtained, features are extracted using K-means clustering. The image regions are described using a vocabulary of blobs.

Lan *et al.* use image similarity and vocabulary prior probability for generating image tags [112].

A comprehensive analysis of few methods is given in Table 2.3.

TABLE 2.3: Comprehensive details of few existing methods for image annotation using machine learning.

| **S.No.** | **Paper** | **Techniques** | **Result** |
|---|---|---|---|
| 1 | SURF descriptor (Wojnar *et al.* 2012) [92] | SURF descriptor, SVM classifier, Fast-Hessian detector | **Error**: IRMA radiographic images: 3.4 |
| 2 | Multi label learning and multi feature extraction (Renuse *et al.* 2017) [93] | Extracted features, mapping of tags, features and dictionary learning, C4.5 classifier | **Accuracy**: IAPR TC12: 78 |
| 3 | Random Forest Classifier and Confidence Assigning (Lee *et al.* 2011) [94] | Wavelet-based CSLBP (WCS-LBP) descriptors, body relation graph (BRG). | **Error**: ImageCLEF2007: 20.33 |

| S.No. | Paper | Techniques | Result |
|---|---|---|---|
| 4 | Fusing Content-Based and Tag-Based Technique Using Support Vector Machine and Vector Space Model (Chan *et al.* 2014) [96] | Color moment (CM), edge orientation histogram(EOH), local binary pattern (LBP), cosine similarity | **Accuracy**: 100 test-images from Google image search: 0.65 |
| 5 | RF (Budikova *et al.* 2018) [97] | Annotation relevance feedback, candidate keywords, three multi-modal search techniques | **Mean Precision**: Profiset: 73.8 |
| 6 | Multi-view non-negative matrix factorization and semantic co-occurrence (Zhong *et al.* 2016) [98] | Relevance feedback algorithm | **Average Precision /Average Recall/F1/ N+**: Corel5K: 27/32/29.3/ 139 |
| 7 | Supervised NMF-Based (Jing *et al.* 2012) [101] | Image descriptors, annotation terms, latent image bases, three-block proximal alternating nonnegative least squares algorithm | **Accuracy**: LabelMe: 0.8866 UIUC Sports: 0.8482 Cal Tech 20: 0.7485 |
| 8 | Object based image retrieval (Bhargava *et al.* 2014) [103] | Selection of objects with in an image | **Precision/Recall**: IAPR TC12: 0.38/0.35 |
| 9 | Graph Semi-Supervised Learning (Bi *et al.* 2018) [104] | Normalization and modification of decision boundary, scoring model, image similarity calculation | **Precision**: Clinical Endoscopy Images: 88 |
| 10 | Divide and Conquer Method (Li *et al.* 2016) [107] | Data grouping, constrained clustering, softmax gate network, single-hidden-layer feedforward neural network. | **Precision/Recall/F1**: Corel15K: 32/36/34 ESP Game: 38/22/28 IAPRTC-12: 43/22/29 |
| 11 | Ant colony optimization algorithm (Akhilesh *et al.* 2016) [108] | Mean shift algorithm, feature extraction, optimization of feature descriptor weights | **Precision/Recall**: BSD300: 0.54375/ 0.5136 |
| 12 | PFCO-based (Wen *et al.* 2017) [109] | Jseg algorithm, multi-SVM, inference rules | **Recall/Precision**: ECCV: 0.278/0.236 |
| 13 | Supervised PLSA (Guo *et al.* 2013) [110] | Topic models, probabilistic latent semantic analysis (PLSA), graphical model | **Accuracy**: MSRC-21: 75.8 Corel: 82.8 |
| 14 | Vocabulary prior probability (Lan *et al.* 2010) [112] | Generative model, image similarity | **Precision/Recall**: Corel5K: 0.1679/0.1663 |

The misclassification examples produced by the model [92] point that the image must have a certain level of quality for reliable annotation. The model proposed in [96] can handle only a single concept. Complicated image content affects the classification performance of the model proposed in [101]. In the model proposed by [104], loss of information occurs because of quantization, and the semantic gap exists between visual features. [109] can only be used for small-scale annotation. It can be seen that algorithms based on multi-features and multi-view [93, 97] provide an excellent

result. Model based on genetic algorithm [108] also performs well. Latent semantic analysis [110] gives the best result, which implies such models should be explored more.

#### 2.2.2.2 Deep Learning Methods

Deng *et al.* [113] combine U-shape network with active learning for image annotation. Yan *et al.* [114] use multi-label convolutional neural network which has a multi-scale structure and a noise-robust loss. Jin and Nakayama [115] see the problem of image annotation as a sequence generation problem. They develop a recurrent image annotator model for the purpose. Niu *et al.* [116] use two branches of deep network. The first branch runs deep. The second branch is for the fusion of multi-scale features obtained from the first branch. They use user-provided tags as a supplement to image input to the network. Yao *et al.* [117] use a stacked discriminative sparse autoencoder for the image annotation system.

Ke *et al.* [118] develop a model for image annotation based on a deep convolutional neural network (CNN) and multi-label data augmentation. It is an end-to-end automatic image annotation model that transforms the image annotation problem into a multi-label learning problem. It employs adaptive feature learning. The multi-label data augmentation method is based on Wasserstein's generative adversarial networks (GANs). Wu *et al.* [119] also use GANs. They perform sequential sampling from deterministic point process model.

For image annotation, Hu *et al.* [120] adopt a model based on multi-view semi-supervised learning. Liu *et al.* [121] employ semi-supervised neural network for image annotation.

The method of Zhang *et al.* [122] relies on semantic features and features obtained from convolutional neural networks. These features then form the input for co-graph regularized collective non-negative matrix factorization. Similarly, Ning *et al.* [123] and Wu *et al.* [124] use CNN to extract deep features of the image.

Sapkota *et al.* [125] employ a pixel-wise classification to solve this problem. The CNN is trained to label each pixel with the value of the raw RG values of the patch centered at that pixel. Wang *et al.* [126] obtain multiple intra- and inter-dependencies between image and labels to generate tags. Luo *et al.* [127] perform attribute-specific segmentation. The tags are then generated from these regions from multi-label networks. Jiu *et al.* [128] use spatial geometric context as the weights of a deep network. Markatopoulou *et al.* [129] use deep convolutional neural network architecture for providing image annotation tags by exploiting concept relations at two different levels.

Tumas and Serackis [130] integrate YOLOv3 into image labeling software. Jin *et al.* [131] improve CNN results by combining the object detection and outlier-merging algorithm to deal with outliers. Long *et al.* [132] use pre-trained ResNet model for representing features. They also have a new loss function that adjusts weights as per the difficulty levels of recognizing different labels for the same input. Li and Ye [133] fuse multiple features. They measure the similarity of scenes based on hierarchical

similarity diffusion. Based on the similarity, the scenes are clustered and annotated. Wang *et al.* [134] developed a model based on graph convolutional network.

Ficiu *et al.* [135] develop a complex annotation framework that automatically generates high-quality markings. Wu *et al.* [136] utilize weakly labeled images and calculate triplet similarity loss for unlabeled images. Zhang *et al.* [137] presents a microarchitecture unit for annotation, which is two times deeper CNN with only 10% parameters.

A comprehensive analysis of few methods is given in Table 2.4.

TABLE 2.4: Comprehensive details of few existing methods for image annotation using deep learning.

| S.No. | Paper | Techniques | Result |
|---|---|---|---|
| 1 | New Framework (Deng *et al.* 2019) [113] | Active learning, U-shape network, suggestive annotation strategy | **Accuracy**: IBSR18: 90.91 MRBranS18: 87.4 |
| 2 | Knowledge mined from radiology reports (Yan *et al.* 2019) [114] | Mine training labels, multi-label convolutional neural network, multi-scale structure | **AUC**: DeepLesion: 0.9083 |
| 3 | Recurrent Image Annotator (Jin *et al.* 2016) [115] | Sequence generation problem | **Precision/Recall/F1**: Corel 5K: 32/35/32 ESP Game: 32/32/31 IAPR TC12: 35/34/33 |
| 4 | Multi-Modal Multi-Scale (Niu *et al.* 2019) [116] | Multi-scale deep model, two-branch deep neural network architecture: a very deep main network branch and a companion feature fusion network branch, multi-modal, noisy user-provided tags, label quantity prediction auxiliary task | **Precision/Recall/F1**: NUS-WIDE: 81.46/ 75.83/78.55 MSCOCO: 75.43/71.23/ 73.27 |
| 5 | Weakly Supervised Learning (Yao *et al.* 2016) [117] | Unified annotation framework, discriminative high-level feature learning, weakly supervised feature transferring, stacked discriminative sparse autoencoder, semantic annotation, tile-level annotated training data, alternate iterative optimization method. | **Accuracy**: 87.93% |
| 6 | Multi-Label Data Augmentation (Ke *et al.* 2019) [118] | Deep convolutional neural network (CNN), multi-label data augmentation, multi-label learning problem, adaptive feature learning, multiple cross-entropy loss functions, Wasserstein generative adversarial networks, deformable convolution, spatial pyramid pooling. | **Precision/Recall/F1**: Corel 5K: 0.41/0.55/ 0.47 ESP Game: 0.48/0.39/ 0.43 IAPR TC12: 0.48/0.43/ 0.45 |
| 7 | Diverse and Distinct (Wu *et al.* 2018) [119] | Generative model, sequential sampling, determinantal point process (DPP) model | **Precision/Recall/F1**: ESP: 42.96/32.34/34.93 IAPR TC12: 43.57/ 26.22/31.04 |
| 8 | Exploring Multi-Facet and Structural Knowledge (Hu *et al.* 2017) [120] | Multi-view semi-supervised learning, exploit both labeled images and unlabeled images, intrinsic data structural information, pairwise constraint on outcomes of different views, classifier learning component | **Precision**: NUS-WIDE: 30.29 MIRFLICKR-25000: 52.60 IAPR TC12: 33.76 |

| S.No. | Paper | Techniques | Result |
|---|---|---|---|
| 9 | Cograph Regularized Collective Nonnegative Matrix Factorization (Zhang *et al.* 2019) [122] | Cograph regularized collective nonnegative matrix factorization method, recommending issue, recover the image-to-label relation, image-to-image relation, label-to-label relation, semantic cooccurrence information of labels, semantic features, convolutional neural networks (CNNs)-based visual features, visual-based label cooccurrence information | **Recall/Precision/F1**: Corel 5K:52/46/48.8 IAPR TC12: 41/51/45.4 ESP: 38/49/42.8 |
| 10 | Integration of Image Feature and Word Relevance (Ning *et al.* 2018) [123] | Deep features, label relevance, synthetic minority oversampling technique, symbiotic and semantic relationships of labels, relevance of label sets, joint convex loss function is proposed, co-regularization | **Precision/Recall/F1**: Corel 5K: 38/49/43 IAPR TC12: 49/31/38 |

The accuracy of boundary delineation is not satisfactory in [113]. Similar structures lead to loss of performance in [114]. The performance gain of multi-scale CNN used in [116] is very less as compared to CNN. The performance of [117] is not comparable to fully supervised methods. In [119], the performance decreases as tag subset size increases. The effect of adding more features and more views for predicting tags is not very clear in the model proposed in [120]. The time complexity of the model proposed in [122] is high. The model [123] can be improved by generating tags other than those present in the training set.

### 2.2.3 Colon Tumour Localization

Colorectal cancer is the second leading cause of cancer deaths. Colon endoscopy is the primary method to detect and prevent cancer. Cancer usually starts with the formation of polyps inside the colon. Early detection of cancerous polyps can bring down the deaths caused by a late/missed detection. Computer-aided detection can provide much help in this field by accurately detecting polyps.

Many features are studied for the extraction of the polyp from the colonoscopy

videos. Mostly edge features [138] are used to separate polyps from colon walls. Texture features [139–141] and geometrical features [140] are also used to distinguish the structure of polyps from the surrounding walls. Researchers have also explored the use of shape-based features [142,143], radiomic features [142,144], morphological features [145] and ColorSIFT features [146]. Recently, deep learning is exploited for detecting polyps.

Ornela *et al.* [147] use automatic encoder-decoder with CNN to detect polyps. Younghak *et al.* [148] use conditional networks to generate synthetic colon-polyp images to overcome the issue of lack of labeled training data, and after that, perform detection using them. They use dilated convolutions and edge-filtering based conditioned input image during training. Nima *et al.* [143] use an ensemble of CNN exploding various features – color, shape, and texture – and temporal information on multiple scales to accurately detect the polyp. Ruikai *et al.* [149] attempt to classify the polyps using transfer learning from a deep CNN trained on millions of non-medical images. Eduardo *et al.* [150] use the texture feature of patches of image regions as input to CNN for detecting colon polyps. Other significant works are listed in Table 2.5.

TABLE 2.5: Some popular techniques used for Polyp Detection.

| S.No. | Paper | Techniques | Result |
|---|---|---|---|
| 1 | Part-based multiderivative edge cross-sectional profile (Yi Wang *et al.* 2014) [138] | Edge cross-section profiles | Area under the free-response receiver operating characteristic curve, Average number of false regions per image: Randomly selected 46 de-identified video files captured during routine screening colonoscopy: 0.32 |

| S.No. | Paper | Techniques | Result |
|---|---|---|---|
| 2 | Max-AUC Feature Selection (Jian-Wu Xu *et al.* 2014) [151] | Sequential forward floating selection | Area under the receiver operating characteristic curve, 96% by-polyp sensitivity at false-positive (FP) rates of 4.1 and 6.5 for CTC cases acquired at the University of Chicago Medical Center |
| 3 | Automated Polyp Detection (Alexander V. Mamonov *et al.* 2014) [140] | Geometrical analysis and the texture content | 47% sensitivity per frame, 81% sensitivity per polyp at a specificity level of 90% for Hospital of the University of Coimbra |
| 4 | Imbalanced Learning and Discriminative Feature Learning (Seung-Hwan Bae et al. 2015) [152] | Data sampling-based boosting framework | Maximum score points falling inside the polyp mask, precision, recall, PR -AUC, and speed of the detector for CVC-ColonDB |
| 5 | Revamped fly-over (Marwa Ismail *et al.* 2015) [153] | Virtual fly-over | Visibility coverage and polyp detection rate for Virtual Colonoscopy Center and Walter Reed Army Medical Center |
| 6 | Shape and Context Information (Nima Tajbakhsh *et al.* 2016) [143] | Hybrid context-shape approach | Sensitivity of 88.0% for CVC-ColonDB, sensitivity of 48% for the ASU-Mayo, polyp detection latency of 0.3 seconds |
| 7 | Texture Feature Extraction and Analysis (Yifan Hu *et al.* 2016) [139] | Adaptive approach to extract and analyze the texture features | Area under the curve of receiver operating characteristic, Differentiation capability of 0.8016 for CTC database |
| 8 | Bag of Feature (Yixuan Yuan *et al.* 2016) [141] | Bag of feature, textural features from the neighborhoods of the key points | Accuracy for Qilu Hospital, Shandong University 93.2% |
| 9 | Convolutional Neural Networks and Random View Aggregation (Holger R. Roth *et al.* 2016) [154] | Deep convolutional neural network | Sensitivity for Data from three institutions: 75% at 3 FP |
| 10 | Convolutional Neural Networks (Nima Tajbakhsh *et al.* 2016) [155] | Fine-tune a CNN | Free-response receiver operating characteristics curve for Database of 40 short colonoscopy videos |

| S.No. | Paper | Techniques | Result |
|---|---|---|---|
| 11 | Transferring Low-Level CNN Features From Nonmedical Domain (Ruikai Zhang *et al.* 2017) [149] | Transfer learning from big nonmedical datasets | Precision 87.3%, Recall rate 87.6%, Accuracy 85.9% for PWH Database |
| 12 | Integrating Online and Offline Three-Dimensional Deep Learning (Lequan Yu *et al.* 2017) [156] | 3-D fully convolutional network | Precision and Recall for ASU–Mayo Clinic Polyp Database |
| 13 | Shape Index, Multiscale Enhancement Filters, and Radiomic Features (Yacheng Ren *et al.* 2017) [142] | Shape index, multiscale enhancement filters, and radiomic features | Free-response receiver operating characteristics curve, by-polyp sensitivity, per-scan sensitivity and FP rate for Walter Reed Army Medical Center |
| 14 | Saliency and Adaptive Locality-Constrained Linear Coding (Yixuan Yuan *et al.* 2017) [146] | Color scale invariant feature transform | Accuracy for 40 patients' WCE image video clips 88.61% |
| 15 | Conditional Adversarial Networks (Younghak Shin *et al.* 2018) [148] | Conditional adversarial networks | Precision and Recall for CVC-CLINIC, CVC- ClinicVideoDB |
| 16 | Unified Bottom-Up and Top-Down Saliency Approach (Yixuan Yuan *et al.* 2018) [157] | Superpixels, sparse autoencoder (SAE) to learn discriminative features | 0.818 recall for CVC-clinicDB |
| 17 | 3D Radiomic Features (Yacheng Ren *et al.* 2018) [144] | 3-D radiomic features | 98.5% by-polyp sensitivity at 2.0 FP for Walter Reed Army Medical Center , The Cancer Imaging Archive |
| 18 | Region Based Deep CNN and Post Learning Approaches (Younghak Shin *et al.* 2018) [158] | Region-based convolutional neural network | Precision, Recall and Specificity for CVC-CLINIC, ETIS-LARIB, ASU-Mayo Clinic Colonoscopy Video dataset and CVC- ClinicVideoDB |
| 19 | Ensemble Learning Approach (Xiaolei Xie *et al.* 2018) [159] | Data-driven modeling | Accuracy, sensitivity and specificity for Beijing Friendship Hospital |
| 20 | Morphological Features (Yacheng Ren *et al.* 2019) [145] | New morphological features | FP rate 2.0 at 96.2% by-polyp sensitivity, 2.1 FP at 93.9% per-scan sensitivity for The Cancer Imaging Archive (TCIA) database, Walter Reed Army Medical Center (WRAMC) database |

| S.No. | Paper | Techniques | Result |
|---|---|---|---|
| 21 | Ensemble of Instance Segmentation Models (Jaeyong Kang *et al* 2019) [160] | Object detection neural network "Mask R-CNN" | Mean pixel precision, mean pixel recall and interception over union for CVC-ClinicDB, ETIS -Larib, and CVC- ColonDB |

## 2.3 Issues and Challenges

There are various issues related to the field of salient object detection. They can be summarized as below:

- Saliency is not very clearly defined. It is highly subjective.
- The maps generated by various algorithms contain regions of grayscale. For evaluation of the results obtained from these algorithms, binarization has to be performed. The selection of threshold creates a major problem.
- A persistent question remains that whether the segmentation and saliency location should be treated as a unified or separate task. As a unified task, satisfactory results are not obtained [2]. Over-segmented results lead to the wrong detection of object boundaries [2]. An incorrect segmentation results in the wrong localization of the salient object.
- Some of the datasets have only a single object with a simple background. Also, the salient object in the images is sometimes too obvious for detection.

For cluttered images and complex background, most of the algorithms fail to provide a decent result [3].

- Heterogeneous objects cause incomplete object detection. Objects similar to the background also lead to wrong diagnosis of salient object [22].
- All the regions that are different from the border are not necessarily salient [5].
- The prior knowledge of hand-crafted features is database specific and is not useful in all cases [16].
- Bottom-up methods fail in extracting high-level semantic features [67].

For pixel-based methods, the computation complexity is very high [19]. They also highlight edges more than they highlight the salient object [21]. Most of the methods that use reconstruction errors use only a background dictionary [20], while some methods do not give any importance to spatial locations [9]. Geodesic methods can find interesting points in an image, yet they are far from locating specific salient objects in the images. Graph-based methods cannot characterize the original image too well; hence are not sufficient to locate the salient object [23]. The surroundedness cue does not highlight the salient object uniformly.

The deep learning methods, though, provide good results also suffer from various limitations. They can be summed up as follows:

- Most of the networks developed for salient object detection cannot deal with the scale-space problem [16].

- Due to numerous convolution and pooling operations, the boundary of the salient object is blurred [63, 73].
- Current architecture requires large scale pre-trained CNNs.
- Much of the location information and fine details are dropped [69, 73].
- For the deep network architectures, it is hard to learn the global structural information because of pixel-wise loss functions [75].
- Introduction of multiple stages also reduces the efficiency of the procedure [71].
- Deep learning methods which are region-based suffer the problem of the narrow receptive field.
- The performance is challenged in non-salient regions and the existence of multiple salient objects [72].
- Edge information is not utilized properly by many of the networks [75].
- The training of all the deep learning-based methods is prolonged.
- Overlap region storage causes a lot of data redundancy [76].

Region-based CNNs lose context due to which they cannot accurately locate the salient object [67]. Patch-based CNNs are very inefficient in terms of computation and storage. Methods which directly work with an image instead of patches have problem with multi-scales and weak semantic information.

For the models proposed in this thesis, the issues mentioned above have been kept

in perspective. For example, all the models designed for salient object detection produce output as binary maps instead of grayscale images. This makes the evaluation of the proposed model easy and accurate. The algorithms proposed have been evaluated for simpler and complex datasets to avoid biased feature extraction. For avoiding problems of heterogeneous objects, minimum directional contrast has been used such that the complete object is highlighted uniformly. Location-prior has not been used to avoid missing objects located at corners and boundaries. For gathering semantic features, object proposals have been used. To avoid burred object boundary because of convolution and pooling, the hybrid model has been generated, which produces output with sharp boundaries. The single-stage YOLOv2 network is used for object detection to maintain the efficiency of the procedure.

Various challenges have to be faced while designing methods for image annotation. A few of them are listed below:

- The primary challenge is that the evaluation methods are inconsistent.
- Subjective evaluations of tags conflict.
- Images with distortion, blur, occlusion, brightness problem, color anomaly, and signal loss lead to wrong results.
- The category definition of objects is also not very clear [88].
- The order of tag generation is also a less explored area [89].
- The applications of annotation in the medical field had been very limited [90].

- Most of the methods use image similarity measure for generating tags. But, there is no distance metric good enough. Nearest neighbors selected by visual similarity do not share the same tags.
- Efficient label transfer mechanism is also complicated to define [98].
- The correlation between labels is ignored [91]. So, the connection between them cannot be utilized.
- The training information does not have a complete annotation of the image [93].
- Various models that have been provided rely on low-level features that have no semantic information. Other models reduce the high-dimension feature to low-dimension. Thus, significant improvement is not achieved [95].
- It is difficult to rank results obtained from different classifiers [96].
- The use of CNN is limited to medium-sized target vocabularies, for which reliable training data is required.
- Search-based image annotation methods do not produce satisfactory results [97].
- The correlation between features and labels is complicated to map.

The image annotation models proposed in this thesis deal with the issue of large feature vectors required for image tagging. An extensive study on feature selection

algorithms and reducing the feature vector length results in lesser tagging time and leads to an accurate result.

There are many issues related to the images provided by colonoscopy. There are areas of white light reflection. Sometimes texture of colon walls and polyp is almost similar, which creates difficulty in segmenting the polyps. The angle of endoscopy also can lead to a missing polyp. There is also much human error involved in detecting polyps due to fatigue. Lack of illumination also makes it difficult to locate the polyps.

In the proposed model, the white light reflection is handled using YIQ color images. Low-level feature problems are removed by using deep learning methods for colon detection.

## 2.4 Benchmark Databases

For evaluating the salient object detection algorithms, a variety of publicly available datasets are used. The following datasets are used to evaluate the models presented in this thesis:

- *Bruce-Annotation (Bruce-A) Dataset* [161]: This dataset contains 120 indoor and outdoor scenes for eye-tracking prediction.
- *CVC-Clinical Database* [162]: From 29 different sequences, 612 still images are obtained. Together with each image, a ground truth data is provided, which

is a binary mask covering the polyp area. There are a total of 672 instances of polyps, as there are few images with multiple polyp instances.

- *Extended Complex Scene Saliency Dataset (ECSSD)* [163]: The dataset has 1000 images with complex background.

- *DUT-OMRON Image Dataset* [164]: The dataset contains 5168 images with a highly complex background.

- *Judd Annotation (Judd-A) Dataset* [165]: The dataset has 300 images of outdoor and indoor scenes and is broadly used for eye-fixation evaluation.

- *Microsoft Research Asia (MSRA) Dataset* [166]: It is the most widely used dataset. It has 10000 images with a single salient object, mostly located in the center of the image. The background of the images is simple and clutter-free. Researchers have also used subsets of this dataset with 5000 and 1000 images. They are called as MSRA-5k and ASD, respectively.

- *PASCAL-S Dataset* [167]: The dataset has 850 images on eight subjects.

- *THUR 15K* [168]: The database contains five categories of images. For each group, there are 3000 images. The salient regions are marked at the pixel level.

For image annotation, the following databases are used by different researchers:

- *IAPR TC-12* [169]: This database is a collection of 20000 images of various categories. Every image has an attached caption.

- *LabelMe Dataset* [170]: They provide an online automatic annotation tool that helps developers test various computer vision algorithms.

- *Microsoft Research Cambridge (MSRC) Dataset* [171]: The dataset has 591 images for 23 object classes.

- *National University of Singapore(NUS)-Wide Dataset* [172]: NUS-WIDE Lite holds a total of 73 attributes with much variety like the airport, animal, beach, etc.

- *UIUCSports Dataset* [173]: The image dataset contains 1579 annotated images in 8 different sports categories.

## 2.5 Performance Metrics

The following metrics are used for the evaluation of proposed salient object detection models:

- Area Under Curve (AUC): It is the area under the curve of Receiver Operating Characteristics (ROC) Curve.

- F-Score: It is the harmonic mean of precision and recall.

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{2.1}$$

- Intersection-over-union (IoU) score: It is defined as the ratio between area of overlap and area of union between actual and predicted salient region.

$$\text{IoU} = \frac{\text{Area of overlap}}{\text{Area of union}} \tag{2.2}$$

- Mean Absolute Error (MAE): It is the average of incidents every time a salient pixel is marked as non-salient and vice-versa.

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n} |x_i - x| \tag{2.3}$$

  where $x_i$ represents predicted value and $x$ ground truth value.

- Precision Recall (PR) Curve: It is a curve drawn between precision and recall. It is helpful in the proposed models as the data for the proposed models is imbalanced.

- Receiver Operating Characteristics (ROC) Curve: It is a curve drawn between true positive rate and false positive rate.

For image annotation, the following evaluation metrics are used:

- Accuracy: It is the ratio of the number of relevant keywords to the total number of relevant keywords assigned to the image in the ground truth.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{2.4}$$

where, True Positive (TP) = Total number of tags which are present in the predicted result as well as in ground truth;

True Negative (TN) = Total number of tags which are absent in the predicted result as well as in ground truth;

False Positive (FP) = Total number of tags which are present in the predicted result but not present in ground truth;

False Negative (FN) = Total number of tags that are not present in the predicted result but are present in the ground truth.

- Precision: It is the fraction of relevant keywords among all the predicted keywords.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{2.5}$$

- Recall: It is the fraction of relevant tags predicted by the model.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{2.6}$$

- Coverage: Coverage [174] is the number of steps required to move down to cover all labels in the list. The lesser number of steps required the better the algorithm.

- Macro-averaged F1: Macro-averaged F1 calculates the F1-score of each class independently and then calculates the average.

- Micro-averaged F1: Micro-averaged F1 is calculated by aggregating the contributions of each class to calculate the average metric.
- One-error: One-error calculates the fraction of examples whose top-ranked predicted label is not in the actual label list.
- Ranking Loss: The ranking loss is the fraction of example where an irrelevant tag has a higher rank than a relevant tag.
- True Negative Rate: The fraction of irrelavant tags correctly identified.

$$\text{True Negative Rate} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{2.7}$$

## 2.6 Conclusion

This chapter reviewed state-of-the-art methods in the fields of salient object detection and image annotation. It also explored the issues and challenges of these fields and the existing ways. The benchmark databases on which the proposed models are evaluated were also described. Finally, the metrics on which the performance of the proposed model is evaluated were explained.