Effective Analysis of Multimedia Data: a Human Attention Perspective

Amit Bora, Shanu Sharma

Abstract*: *The advancement in the technology related to capturing, storage and transmission of data, have made multimedia data like images and videos pervasive today. But proper tools to describe and organize them are still not efficient. These images and videos are mostly targeted to human beings, whereas most of the existing multimedia assessment techniques are lack of subjective efficiency. Thus, more and more researchers are engaged in this filed to provide effective techniques of multimedia data analysis by accommodating user's perception and cognition. Keeping user's perception in mind, in this paper human attention-based perspective is discussed for multimedia data analysis. Recent work done in the field of attention, eye tracking, based image and video analysis along with the description of various benchmark datasets are summarized. The paper aims at providing the recent advances in field of user oriented effective multimedia data analysis along with the future directions.

***Index Terms*: Saliency, Image, Video, Attention, Multimedia**

I. INTRODUCTION

Due to the advancement of multimedia technologies, a huge amount of multimedia data is generated, transmitted and stored on daily basis. This rapid increase in multimedia data also increased the problem of handling and storing it. A wide range of this multimedia data based real time applications also exists in field of online gaming, web streaming, robotics etc. Efficient management of this huge amount of data and development of effective applications is nowadays the current interest of multimedia research community.

Due the enormous growth in the field of vision science over past half century, today we have deep understanding about the processing of visual information by brain like which portion of brain is used in process of recognition, functional modelling of human neurons etc. A sufficient amount of information about low level visual processing which is quantitive is accessible to multimedia data processing researchers. Based on this information till now many effective computer vision-based technologies have been developed in field of multimedia data analysis like compression, enhancement and retrieval etc.

Revised Manuscript Received on April 05, 2019.

Amit Bora, Department of Computer Science & Engineering, Amity University, Uttar Pradesh, India

Shanu Sharma, Department of Computer Science & Engineering, Amity University, Uttar Pradesh, India

For example over last years, researchers have successfully implemented the concept of human visual models in designing or refining the video processing systems. Some examples are use of quantization weights in JPEG image compression, early color perception based CIEXYZ color space proposed by Wright and Guild, generation of temporal contrast sensitivity function which is based on the concept of temporal frequency response of the human visual system. The problem related to these techniques is that they are based on objective aspects (like Mean Square Error, bit arrangements, data received etc) which neglects the perceptive features of the data. To bridge this gap, subjective aspects (perceptual features) are required. As multimedia data is mostly targeted to human beings, the incorporation of user oriented subjective features can provide more efficiency to existing techniques.

Generally proper assessment of image or video is based on its characteristics, using these characteristics appropriate features are then selected. For human's perception-oriented assessment, user-oriented features need to be extracted from spatial, chromatic and temporal domain. Multimedia data processing using these features can provide high correlation with subjective scores of human observers. Usually manual process of video analysis involves the user's perception, thinking and action steps in sequence. As human's perception is directly dependent on human's attention, this paper aims at exploring the modelling of human's attention i.e saliency for development of effective multimedia applications. Since the human visual system is very complex and replicating such a visual system through computer vision makes the accurate saliency detection as a challenging task. In this paper work done in past five years in field of saliency-based multimedia data analysis is discussed and analyzed. Further various state of the art models and benchmark datasets available in current field is also discussed.

The paper is organized as section II explains about basic concepts of human attention modelling, saliency, its types and applications; Section III explain about different methods by which we can perform saliency detection and eye fixation in Images and videos; Section IV summarizes the various benchmark datasets available for saliency algorithm testing for images and videos. Section V summarized the work done in past five years in tabular format, followed by Conclusion and future

directions in the related field.

II. HUMAN ATTENTION MODELLING: SALIENCY

Eye rotations are a complicated and integral part of human vision, they occur both voluntarily and involuntarily, and allow a person to fixate on features in the world, even in the case where head or target features are moving. One of the main reasons for eye movement is to position the feature of interest on the fovea. Another reason for eye movement is that the photoreceptors are slow to respond to stimuli due to their chemical nature. They take up to 10ms to fully respond to stimuli and produce a response for up to 100ms. Eye movements help keep the image fixed on the same set of photoreceptors so that they can fully charge [1].

A. Visual Saliency

Saliency is one of the subjective measure by which we can impart the human attention perspective to machine. Saliency detection is the process of estimating the attention mechanism of individuals for facilitating the various cognitive and perceptual factors related to them. It is based on the eye fixation mechanism of humans on particular part of an image which occurs at the position of stimulus in a scene.

Visual Saliency estimation is the process of estimating the attractive visual signals in images and videos. Saliency estimation can provide the analysis of images and videos in the same way as the human generally do and thus can meet the human perception in a better way. It is a quality by which an object stands out from its neighbors and surrounding. As visual saliency imitates human vision perception, it becomes a hot and challenging research area in the field of multimedia and computer vision applications. Visual saliency detection has a wide range of applications including eye gaze prediction, object importance detection, video interestingness estimation, image quality assessment, image memeorability assessment.

B. Computational Assessment of Saliency

Saliency is a distinct subjective perceptual quality which makes something emerges as more interesting others and quickly catches our attention. In computer vision, Saliency is performed using saliency maps. Saliency map provides an image showing each pixel's uniqueness in binary threshold form. It simplifies the image representation into something more meaningful and easier to analyze.

For example, a pixel having high grey or other unique color quality will obviously have higher value in saliency map than others as shown in Fig 1. It is a kind of image segmentation but provides better object extraction than segmentation. Each pixel in a region is similar to some other with respect to some characteristic like color, texture, intensity etc.

Fig. 1. Fine grained saliency maps for different objects in Images

Computational assessment of saliency can be achieved using low level and high-level features of an image or video. Saliency detection methods using low level features are referred as Bottom up methods whereas high level features-based saliency detection approach is known as Top down methods. Bottom up features are the core features which constitutes a visual element. Localize objects from surrounding which show distinct characteristics in free viewing condition [11]. This includes color, intensity, orientation, local entropy etc. Many benchmark models have been proposed for detection and estimation of salient regions and objects in Images but very less work is done for saliency in videos. Top down saliency or features refers to querying based on predetermined objects like faces, cars, person etc. [8][2]. These are target and goal oriented which looks for target object in scene by employing appearance characteristics of target object [11]. Some authors [2] believes that at early stages bottom-up features guides attention and later, high level factor (actions, events) directs eye movements and comes with their own model which was a combination of both approaches. Basic models based on these two approaches along with related work done in past five years is discussed in detail in next section.

III. SALIENCY DETECTION: APPROACHES AND RELATED WORK

A. Saliency Estimation Approaches

The current research field has atleast 42 state-of-the-art models. These state-of-the-art models have worked well and have set a benchmark. The models are based on both top-down approach and bottom-up approach. This domain has attracted lot of attention in the field of computer vision. Some popular bench mark models using are discussed below:

Classical Methods: These are the methods which are based on biological phenomena's of human visual system and uses simple models and hardware to implement them whereas advance methods deals with methods which includes neural networks, superpixel segmentations etc. Some of the classical methods used for bottom-up and top-down saliency estimation are explained below:

[3]Koch and Ullman 1985 provided the biological plausible architecture which was theoretical and explained the role of color, orientation, and direction etc for the attention modelling of humans. They proposed selected visual attention works in three stages: First, computing a set of elementary features (color, orientation or motion) as topographical maps in parallel across the visual field in which locations differed from the surrounding as per elementary features are singled out. Second, a Winner Take All (WTA) network on map selects the most salient region or location and fuse information from different maps in a single whole map. Third, WTA networks automatically shifts the attention to next locations. These shifts can be biased with proximity (searching for interesting object in nearby locations) and similarity preferences (searching for salient targets similar to the current one). After their theory Itti and Koch 1999 [4] proposed a practical implementation of model which was based on the bottom up attention psychology of human vision and feature integration theory ie center

surround mechanism. In their work four ways were provided to integrate the feature maps together. Simple normalized summation (naive method), linear combination with learned weights (increase in weights for those regions which show higher peak values inside target region), global nonlinear normalization followed by summation (globally multiplying the saliency map with difference of global maxima and average local maxima) and local nonlinear competition between salient locations followed by summation (convolving each feature map with 2D Difference of Gradient (DoG) filter and truncated Gaussian filter for boundary condition within multiscale framework of Gaussian pyramids and further explained this model in 2000). They performed the comparison among these methods on natural images database with aluminum can, vehicles emergency triangle and traffic signs as target. [6] Bruce and Tsotsos 2006 proposed bottom-up model based on information maximization sampled from scenes and Shannon's self-information measure which measures the local contrast. For each local image patch, independent component assumption (ICA) was employed with probability of image location as likelihood using gaussian kernel density estimate and self-information for final saliency estimation. [7]Harrel and Koch 2007 provided Graph Based Visual Saliency (GBVS) approach which generates activation maps on some feature channels and normalizing the maps using markovian approach. [10]Hou and Zhang 2007 proposed method based on log spectrum of input image by extracting image residual in spectral domain. Log spectrum helps in detecting low level frequencies without the need of any prior feature information of object. The difference between log spectrum and shape information is obtained as spectral residual whose inverse fourier transform provides the resultant saliency map. [11]Achanta et al 2009 introduced salient region detection method which provides full resolution saliency maps with well-defined object boundaries. They analyzed 5 state-of-art methods in frequency domain and found the deficiencies due to improper frequencies. To remove or reduce those deficiencies, they performed frequency tuning using color and luminance features which provided better results than others. Limitation to this method was that it worked best for large objects and fails in case, if object is not distinct from background.

Advanced Approaches: Some advanced techniques which uses concept of Convolution neural network (CNN), Recurrent Neural Network(RNN), superpixel segmentation and histogram were also developed for saliency detection.

Simple Linear Iterative Clustering (SLIC) superpixel segmentation clusters the pixels in 5 dimensional CIELAB color space and x-y pixel coordinates to generate compact, nearly uniform superpixels which provides low computation with better results than normal segmentation methods. CNN eliminated the dependency of hand crafted features and the dependency on centre bias knowledge. It consists of multiple layers with 10-1000 neurons per layer where each layer can perform tasks like augmentation, rotation, convolution etc. It uses a kind of multilayer perceptron network that is designed to require minimal preprocessing. They have applications in image/video recognition, recommender systems, NLP, image classification, image analysis etc. [16], [20], [23], [26] provided different CNN networks for saliency or salient region detection in images and videos. Similarly, more models like SALiency in CONtext(SALICON), Multi Level NET (ML-NET), Multiscale Deep Feature (MDF), MultiTask, Unsupervised Hierarchical Model (UHM) are present, based on CNN's for different applications. [23] provided a good literature about CNN model in their work. Histogram based approaches includes Kl-divergence and EMD methods to generate the dissimilarity histograms and find different regions in data. These dissimilarities are then used to generate the salient regions in the data.

B. Related Work

Saliency plays a vital role in many applications related to images, videos and audio files. Almost every multimedia file contains some information which needs to be determined and distinguished from other information for some particular task. Here, in this section, we will discuss the work done based on saliency concepts on Images, Videos and for both category.

Saliency detection in Images: Saliency detection in images has wide range of applications including image segmentation, region-based image retrieval, adaptive image compression, mobile robot navigation, surveillance and many more. Saliency in Image data is useful in tasks like Object detection and recognition, Image compression, Image Thumbnailing, Automatic collage creation, Non-photorealistic rendering, Advertisement design, Image Retargeting, Scene classification, Object based image retrieval [18], Adaptive region of interest based image compression [18], Smart image resizing [18] etc.

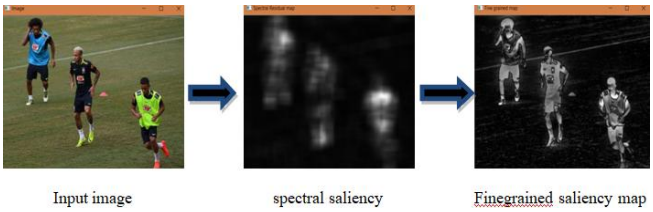

Fig. 2. Example of Saliency detection in Images

In [2], Authors developed saliency maps by fusing color, texture saliency maps with prior information (like color and location priors) by dividing the input image into N segments with SLIC superpixel algorithm in CIELAB color space. Model was able to provide better results as compared to other models but makes the superpixel and color conversion process mandatory. [5] provided an approach to locate objects using graph based manifold ranking technique on boundary priors (salient object rarely occupies three or all sides of image). Exploited boundary priors by using nodes on each side of image as labeled background query helped in computing saliency of nodes on relevance and used 4 labeled maps to generate saliency map by integrating them, applied binary segmentation on saliency map to take labeled foreground nodes as salient queries. Final saliency map was computed by relevant nodes to foreground queries. In [8], authors provided deep information of top-down and Bottom-up methods for saliency detections and estimation. It focuses on recent attention modeling efforts and computational model of attention as defined by Tsotsos and Rottenstein (2011). [12] provided three models based on Top down saliency estimation to generate

category specific saliency map addition to dictionary based salient approaches (patch based, provide better representation of objects) and super pixel based bottom up salient object detection methods. Resulted in better performance than other compared models but lagged in some images like sofa, bottles, Tv/monitor, chair. [22] provided work on traffic data analysis by proposing the idea of vanishing points. They considered front as vanishing point for the scene and used it as a guiding medium for their experimentation. Vanishing point detection algorithm by Hui kong was used in their work and combined both top-down and bottom-up approaches with vanishing point concept to get the saliency maps. They also rewrote the ROC algorithm and considered only True Positive Rate as their model significantly increased the True positive values. In [23], authors proposed SalClassNet framework consisting of two networks which are jointly trained. First network generates the top-down saliency maps of input images (with 13 convolution and 5 max pooling layers) while the second network performs the visual classification (inspired by Inception network) using saliency maps. Also provided a dataset of top-down saliency maps, subset of standford dogs dataset. [21] addressed the problem of missing boundary and structural information in salient objects using a deep CNN with low level enhanced features (LFE). Low level features were obtained using shallow sub-networks fused with heuristics and guided features. [25] All Object detection algorithms results in objects surrounded by a bounded box. This box limits the ability of machine to compute the features of objects like shape and size. In this paper authors tried to cope this problem, using the box location, bag-of-features (HMAX model with linear classifier) as top-down guiding medium to create a contour around the object by Ratio contour algorithm along with a proposed low threshold technique and refined binary classifier for lower True Negative Rate and False Positive Rate respectively. [26] presented a pixel saliency based encoded CNN (PS-CNN). They obtained a saliency matrix with pixel-wise saliency in feature map which is used to threshold the original feature map segments with multiple binary masks and squeezed masked feature maps to get encoded maps using 1x1 convolution kernel. Finally, concatenate original maps with encoded maps to get the fine-grained representation. The presented method provided better classification accuracy on 3 different datasets than other CNN based models but lack on different view angles, strong illumination change and big occlusions.

Saliency detection in Videos: Saliency in Visual data is useful in tasks like Activity recognition, Video compression, Video summarization, Object Tracking, Adaptive content delivery, Motion/ movement detection etc. General framework for saliency detection in videos includes applying subtraction between two consecutive frames; dividing frames into different features like intensity, color, orientation, motion etc., estimating center surround difference and combining the results from previous steps to generate saliency map is shown in Fig. 3. [14] proposed a method to handle dynamic background in videos. For this, they introduced spatial-temporal regional filters that assign high saliency values to stable coherent regions (using KNN Histograms (KNNH) and Markov Random Field Graph (MRFG) for motion distribution) and eliminate repeating dynamic backgrounds. [15] authors presented an approach which identifies and locates salient regions in videos by combining the saliency map both in spatial and temporal domains with the significant regions obtained using external camera and proposed NMF(Non-negative Matrix Factorization) algorithm. AUC score metrics was used to evaluate the models. Limitation was dependency on external camera data. [16] proposed a novel approach based on 3D CNN known as Deep 3D Video saliency (Deep3DSaliency). It consists of 2 modules STSM (Spatio-Temporal Saliency Model to extract ST features) and SSAM (Stereoscopic Saliency Aware Model to infer depth and semantic features). Performed better in terms of ROC, CC, KL Divergence, NSS and Time to process each frame than other networked models. Problem with these networked models is that they need alot of time to be trained and this time is also dependent on the hardware used.

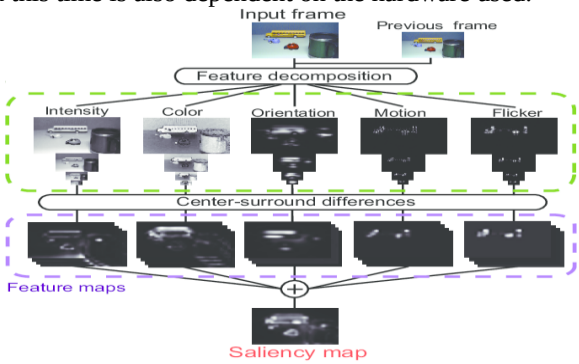

Fig. 3. General framework for saliency detection in videos

[17] provided a model based on Dynamic GMM (DGMM) on a eye tracking database and analysis of face and mouth features while changes were learned and updated using particle filters (PF) to smooth DGMM saliency distribution. Hence the model was named as PF-DGMM. Performed better in terms of AUC, NSS and CC but unable to perform in occluded videos. [18] proposed to integrate actions like speech and head movements along with static features into process of salient region detection and using Multiple Hidden Markov Model (MHMM) which determines the saliency transition between frames for face detection and recognition and implementing it for video conference compression (HEVC) High Efficiency Video Coding. [19] detected the salient objects using spatial and temporal saliency analysis with background and center contrast priors and termed them as object proposals (object proposed to be salient). Then, using a ranking function ranked the proposals according to the spatial and temporal saliency cues. [20] proposed a deep video saliency network having two parts: static which estimates spatial saliency and dynamic which uses static saliency values to generate spatiotemporal saliency results. Using spatial saliency values for 2nd layer removed the need of computing optical flow values in dynamic scenes and helps increasing computational speed. They also proposed a data augmentation technique to prevent over-fitting of network with the help of synthetic dataset and annotating input data. [24] authors used top-down features to analyze the behavior of patients suffering from dementia to predict their movements and object manipulation. Besides using segmentation for foreground objects, they used annotated images with bounding boxes in interested region and proposed a learning framework using Expectation Maximization (EM) which automatically

calculates parameters to optimize data likelihood. They used a combination of global (geometric configuration of arms and saliency on these) and local features (adaptive relocation of saliency priors), estimating the log likelihood from the global and local feature distributions to form the saliency maps based on likelihood. Proposed model worked better than compared models but problem was that it used annotated data which may not be available every time.

Combined Approach: In [13] authors proposed unified framework for salient spatiotemporal region detection and segmentation of objects for videos (STS). Also, provided approach for temporal salient region detection (LGT) along with faster variants of both proposed approaches. They computed the final saliency map by combining the spatial and temporal saliency maps. These two maps were computed as follows:

i. Spatial salient region detection: Patch level image abstraction, Region level image abstraction, Color contrast estimation, Color distribution estimation, spatial saliency assignment and refinement, Center prior integration

ii. Temporal salient region detection is of two types: motion channel (temporal gradient) and optical flow. Authors used optical flow.

Patch level optical flow abstraction, Local temporal saliency estimation with multilevel center surround, Global temporal saliency estimation, temporal saliency assignment, spatiotemporal saliency assignment. In [8], Novel approach Encoding Based Saliency (EBS) to support human activity to recognize and weakly supervise training of activity detection algorithm was proposed with fully unsupervised algorithm (Kmeans for finding minimum number of bins) to find salient regions within videos. Modeled joint distribution of motion or appearance features which yield favorable performance.

IV. SALIENCY EVALUATION

A. Benchmark Datasets

Most of the state-of-the-art methods have provided benchmark datasets for comparing the efficiency of any newly developed algorithm. In Table 1, we are providing information about some benchmark datasets which were used by various authors in evaluation of saliency models.

B. Evaluation Measures

These are used to evaluate the performance of a model and prove its accuracy in comparison to existing models. Some of the important measures/ metrics used for model evaluation are discussed below:

- ROC AUC: treats saliency map as binary classifier for every pixel. Pixels with high saliency value than threshold are salient object while others are not. Human labeled masks are used as ground truth. By varying threshold, ROC curve drawn as False Positive rate and True Positive rate and generates an Area Under Curve.

Predicted Actual

	Negative	Positive
Negative	True negative (TN)	False positive (FP)
Positive	False negative (FN)	True positive (TP)

Table 1. Bench mark Datasets for Saliency Detection

Name	Types of Data	Description	Link
MSRA10K (THUS10K)	Images	Pixel level saliency labeling of 10000 images	https://mmcheng.net/msra10k/
MSRA-B	Images	Pixel accurate salient object labeling of 5000 images	https://mmcheng.net/msra10k/
ECSSD and CSSD	Images	1000 images and 200 images with ground truth masks	http://www.cse.cuhk.edu.hk/leojia/projects/hsaliency/dataset.html
Toronto	Images	120 indoor and outdoor color images of 681×511 resolution	http://www-sop.inria.fr/members/Neil.Bruce/#SOURCECODE
NUSEF	Images	Eye fixation dataset in free view stimuli,1024×768 resolution	http://mmas.comp.nus.edu.sg/NUSEF.html
MIT	Images	1003 natural indoor and outdoor scenes with maximum 1024 resolution	http://people.csail.mit.edu/tjudd/WherePeopleLook/index.html
FBMS	Videos	26 and 59 video sequences with pixel accurate moving object pixel annotations.	https://lmb.informatik.uni-freiburg.de/resources/datasets/
DUT-OMRON	Images	5168 images resized at 400×x or x×400 where x<400 with 1 or more salient object and complex background.	http://saliencydetection.net/dut-omron/
CD2014	Videos	Change detection dataset 2012 and 2014. CD2014 consist of 2012 dataset with more updated data in 11 different categories	http://changedetection.net/ http://jacarini.dinf.usherbrooke.ca/dataset2014/
CVPR PAPERS	Images, Videos	Contains all datasets of papers published in CVPR	http://www.cvpapers.com/datasets.html
HKU-IS	Images	4447 images with pixel wise saliency annotations	https://sites.google.com/site/ligb86/mdfsaliency/
DUTS-TR	Images	10553 training images and 5019 testing images	http://saliencydetection.net/duts/
UCF sports	Videos	150 sequences in 720x480 resolution natural dataset.	http://crcv.ucf.edu/data/UCF_Sports_Action.php

- Precision: ratio between True positive and total predicted positive values. Red portion indicates precision in above table

  P = TP/TP+FP
- Recall: ratio between True positive and total Actual positive. Blue portion indicates recall in above table.

R = TP/TP+FN

- Fscore: it is the combination of precision and recall which provides harmonic mean of data.

  F1 = 2* ((precision*recall)/ (precision+recall))
- NSS: quantifies saliency map values at eye fixation location and normalize it with saliency map variance.

  Nss score = 1/N ∑NSS(p) (N is total no. of eye fixation)
- Linear correlation coefficient (CC/Pearson coefficient): range between -1 and 1. When value close to -1 or 1 almost a linear relationship between 2 variable.

$$Cc = cov(SM, FM)/\sigma_{SM} * \sigma_{FM}$$

V. DISCUSSION AND FUTURE DIRECTION

In this paper we tried to cover the all information on saliency estimation and detection models. These models have been categorised in spatial, temporal and spatiotemporal by different authors and works on using bottom-up and top-down features. For simplicity, the recent work has been summarized in Table 2. Further various benchmark datasets used for modelling saliency along with evaluation metrics is summarized in Table 2. The review presented in the paper provides the summarized information on saliency detection for the efficient implementation of multimedia-based applications. Based on the review it has been noticed that till now a lot of work has been done for saliency estimation in images, whereas very have explored the saliency work for video analysis. As of future work, we will try to incorporate the concept of human attention for estimating the interestingness of video along with various applications for video processing i.e enhancement and retrieval etc.

Table 2. Analysis of Related Work

Ref No.	Year	Methodology	Data	Remarks
[12]	2014	Super-pixel based discriminative dictionaries	Images	lagged in some images like sofa, bottles, Tv/monitor, chair.
[22]	2014	Traffic data analysis with vanishing point (VP) algorithms	Images	Focuses more on objects near to the vanishing points than nearby
[25]	2014	Contour level object detection with HMAX model and linear classifier	Images	Fails in cases of group of objects
[5]	2015	Graph based manifold ranking with boundary priors	Images	Not performed well for center-biased images
[8]	2015	Histogram based baye's classification and kmeans	Video and images	Estimating number of bins and joint distribution of points could be confusing and difficult
[13]	2015	Spatial and temporal saliency detection	Video and images	Optical flow calculation is difficult and time consuming
[24]	2015	Expectation maximization with global and local features for likelihood estimation	Videos	Use annotated data which could not be present every time
[2]	2016	SLIC based segmentation with color and texture maps	Images	SLIC superpixel segmentation and color conversion is necessary
[19]	2017	Ranking based spatial-temporal saliency analysis and object proposals	Videos	Performed better in terms of Precision, Recall, Mean Absolute Error (MAE)
[23]	2018	SalClassNet with 2 networks-one for saliency and other for classification	Images	Classification layer is highly dependent on Saliency layer
[14]	2018	Filters based on KNNH and MRFG model	Videos	Dependent on optimal value of KNN structure and high computational power
[15]	2018	Camera dependent spatial and temporal region detection using NMF	Videos	Used AUC score to measure performance but dependent on external camera data
[14]	2018	Filters based on KNNH and MRFG model	Videos	Dependent on optimal value of KNN structure and high computational power
[15]	2018	Camera dependent spatial and temporal region detection using NMF	Videos	Used AUC score to measure performance but dependent on external camera data
[17]	2018	Particle filter based Dynamic GMM (PF-DGMM)	Videos	Performed better in terms of AUC, NSS, CC but unable to perform in occluded data.
[18]	2018	Static features with Multiple Hidden Markov Model(MHMM)	Videos	Needs high computational power
[20]	2018	FCN (Fully Convolutional Network)	Videos	Final results dependent on results of static layer.
[21]	2018	LFE network with guided information	Images	Hard to detect small objects and objects with same color as background
[26]	2018	PS-CNN	Images	lagged on different view angles,strong illumination change and big occlusions.
[16]	2019	3D CNN	Videos	Performed better in terms of AUC,CC,NSS, KLD, Time but need complex hardware and time to train model

REFERENCES

1. AL C. Bovik, Perceptual Video Processing: Seeing the Future, Proceedings of the IEEE, Vol 98, No. 11 (2010)
2. L. Zhang, L. Yang & T. Luo, "Unified saliency detection model using color and texture features". PloS one, 11(2), e0149328, 2016.
3. C. Koch & S.Ullman, "Shifts in selective visual attention: towards the underlying neural circuitry". In *Matters of intelligence*, page no:115-141, Springer, Dordrecht,1987.
4. L. Itti & C. Koch, "Feature combination strategies for saliency-based visual attention systems", *Journal of Electronic imaging*, *10*(1), page no:161-170, 2001.
5. D. Zhang & C. Liu, "Salient Object Detection Based on Context and Location Prior". *International Journal of Signal Processing, Image Processing and Pattern Recognition*, *8*(4), 125-134, 2015.
6. N. Bruce & J. Tsotsos, "Saliency based on information maximization", In *Advances in neural information processing systems*, page no: 155-162, 2006.
7. J. Harel, C. Koch & P. Perona, "Graph-based visual saliency". In *Advances in neural information processing systems*, page no: 545-552, 2007.
8. T. Mauthner, H. Possegger, G. Waltner & H. Bischof. "Encoding based saliency detection for videos and images". In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 2494-2502), 2015.
9. L. Itti & A. Borji, "Computational models: Bottom-up and top-down aspects". *arXiv preprint arXiv:1510.07748*, 2015.
10. X. Hou & L. Zhang, "Saliency detection: A spectral residual approach". In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, page no: 1-8). IEEE, 2007.
11. R. Achanta, S. Hemami, F. Estrada & S. Susstrunk, "Frequency-tuned salient region detection", 2009.
12. A. Kocak, K. Cizmeciler, A. Erdem & E. Erdem, "Top down saliency estimation via superpixel-based discriminative dictionaries". In *BMVC*, 2014, September.
13. R. Kannan, G. Ghinea & S. Swaminathan, "Discovering salient objects from videos using spatiotemporal salient region detection". *Signal Processing: Image Communication*, *36*, 154-178, 2015.
14. C. Chen, Y. Li, S. Li, H. Qin & A. Hao, " A novel bottom-up saliency detection method for video with dynamic background", *IEEE Signal Processing Letters*, *25*(2), 154-158, 2018.
15. X. Sun, Y. Hu, L. Zhang, Y. Chen, P. Li, Z. Xie & Z. Liu, "Camera-Assisted Video Saliency Prediction and Its Applications", *IEEE transactions on cybernetics*, *48*(9), 2520-2530, 2018.
16. Y. Fang, G. Ding, J. Li & Z. Fang, "Deep3DSaliency: Deep Stereoscopic Video Saliency Detection Model by 3D Convolutional Networks", *IEEE Transactions on Image Processing*, *28*(5), 2305-2318, 2019.
17. M. Xu, Y. Ren, Z. Wang, J. Liu & X. Tao, "Saliency detection in face videos: A data-driven approach", *IEEE Transactions on Multimedia*, *20*(6), 1335-1349, 2018.
18. M. Xu, Y. Liu, R. Hu & F. He, "Find who to look at: Turning from action to saliency", *IEEE Transactions on Image Processing*, *27*(9), 4529-4544, 2018.
19. F. Guo, W. Wang, J. Shen, L. Shao, J. Yang, D. Tao & Y. Y. Tang, "Video saliency detection using object proposals", *IEEE transactions on cybernetics*, (99), 1-12, 2017.
20. W. Wang, J. Shen & L. Shao, "Video salient object detection via fully convolutional networks", *IEEE Transactions on Image Processing*, *27*(1), 38-49, 2018.
21. T. Zhao & X. Wu, "Image Saliency Detection with Low-Level Features Enhancement". In *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)* (pp. 408-419). Springer, Cham, 2018.
22. T. Deng, A. Chen, M. Gao & H. Yan, "Top-down based saliency model in traffic driving environment". In *Intelligent Transportation Systems (ITSC)*, IEEE 17th International Conference on (pp. 75-80). IEEE, October, 2014.
23. F. Murabito, C. Spampinato, S. Palazzo, D. Giordano, K. Pogorelov & M. Riegler, "Top-down saliency detection driven by visual classification", In *Computer Vision and Image Understanding*, 2018.
24. V. Buso, I. Gonzalez-Díaz & J. Benois-Pineau, "Object recognition with top-down visual attention modeling for behavioral studies". I*n Image Processing (ICIP)*, 2015 IEEE International Conference on (pp. 4431-4435). IEEE.
25. H. Yu, Y. Chang, P. Lu, Z. Xu, C. Fu & Y. Wang, "Contour level object detection with top-down information". *Optik International Journal for Light and Electron Optics*, 125(11), 2014, 2708-2712.
26. C. Yin, L. Zhang & J. Liu, "Pixel Saliency Based Encoding for Fine-Grained Image Classification". In *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)* (pp. 274-285).

AUTHORS PROFILE

Amit Bora is currently pursuing his M.Tech at Amity University, Noida. His area of interest includes, Image, Video processing and Computer Vision.

Shanu Sharma is currently working as an Assistant Professor at Amity University, Noida. Her are of interest if Cognitive Science, Computer Vision and Artificial Intelligence. She is an active member of IEEE, SCRS and IAENG. She has published many research papers in International Conferences and Journals