Saliency Detection for Stereoscopic 3D Images using Neural Network

Rakesh Y[1] , K Sri Rama Krishna[2]

Asst. Professor[1], HOD & Professor[2]

Department of Electronics and Communication Engineering, URCE[1], VRSEC[2], Vijayawada

Abstract - It is the era of rapidly growing digital image usage; automatic image categorization has become prominent research area. Saliency detection is a pre-processing step for a wide area of applications which includes object detection and recognition, face recognition, image compression, visual tracking, object retargeting, image categorization and image segmentation. In this paper, we propose a fast and compact saliency detection method to meet the essential application requirement of salient object detection task. We introduce a computational model for detecting visual saliency for stereoscopic image using artificial neural network model which extracts features of images. The neural network architecture is capable of extracting feature hierarchies from the image pixels automatically. To achieve feature extraction task neural network architecture has to be trained by image dataset. The feature extracted by the neural network having wide semantic information which is helpful in detecting visual saliency. We evaluate our approach on several datasets, including challenging scenarios with different parameters, as well as salient object detection in images. Overall, we demonstrate favourable performance compared to state-of-the-art methods in estimating both ground-truth eye-gaze and activity annotations.

Keywords: Saliency detection stereoscopic 3D image ground-truth eye-gaze activity annotations.

I. INTRODUCTION

Recently saliency detection attracted much research interest. Now, saliency detection focused on finding the most important part of the image. While detecting saliency and retrieval of saliency maps or graphs some important information can be obtained. Based on this information irrelevant images or part of it can be filtered. Salient regions contain important information which in general is contrasted with its arbitrary surrounding. Saliency detection is a pre-processing step for a wide area of applications which includes object detection and recognition, face recognition, image compression, visual tracking, object retargeting, image categorization and image segmentation.

Saliency detection can be categorized as either top-down or bottom-up approaches. Top-down methods are task-driven and require supervised learning with manually labelled ground truth. To better distinguish salient objects from background, high-level information and supervised methods are incorporated to improve the accuracy of saliency map. In contrast, bottom-up methods usually exploit low-level cues such as features, colors and spatial distances to construct saliency maps. The bottom-up strategy of saliency detection is pre-attentive and data-driven. It is usually fast to execute and easy to adapt to various cases compared to top-down approaches, and therefore has been widely applied. One of the most used principles, contrast prior, is to take the color contrast or geodesic distance against surroundings as a region's saliency. Saliency is resulted from visual contrast as it intuitively characterizes certain parts of an image that appear to stand out relative to their neighbouring regions or the rest of the image. Thus, to compute the saliency of an image region, the technique should be able to evaluate the contrast between the considered region and its surrounding area as well as the rest of the image.

II. RELATED WORK

Bottom-up vision based saliency detection and training models to estimate the eye fixation behaviour of humans, either based on local patch or pixel information which is still of interest today. In contrast to using fixation maps as ground-truth, proposed a large dataset with bounding-box annotations of salient objects. By labelling 1000 images of this dataset, refined the salient object detection task. Grouping image saliency approaches, we see methods working on local contrast or global statistics. Recently, segmentation based approaches have emerged which often impose an object-center prior, i.e. the object must be segregated from image borders, mainly motivated by datasets.

Human eye-gaze or annotations as ground truth information for training stereoscopic images saliency methods are another alternative. Eye-gaze tracking data, captured by for activity recognition data sets, emphasized differences between spatio-temporal key-point detections and human fixations. Later, utilized such human gaze data for weakly supervised training of an object detector and saliency predictor which learned the transition between saliency maps of consecutive frames by detecting candidate regions created from analyzing magnitude, image saliency by and high level cues like face detectors.

Recently, much effort has been made to design discriminative features and saliency priors. Most methods essentially follow the region contrast framework, aiming to design features that better characterize the distinctiveness of an image region with respect to its surrounding area. In, three novel features are integrated with a conditional random field. A model based on low-rank matrix recovery is presented in to integrate low-level visual features with higher-level priors. Saliency priors, such as the center prior and the boundary prior, are widely used to heuristically combine low-level cues and improve saliency estimation. These saliency priors are either directly combined with other saliency cues as weights or used as features in learning based algorithms. While these empirical priors can improve saliency results for many images, they can fail when a salient object is off-center or significantly overlaps with the image boundary.

Before the introduction of large stereo datasets, relatively few stereo techniques used ground-truth information to learn parameters of their models. For a general overview of stereo algorithms Kong and Tao used sum of squared distances to compute an initial matching cost. They trained a model to predict the probability distribution over three classes. The initial disparity is correct, the initial disparity is incorrect due to fattening of a foreground object, and the initial disparity is incorrect due to other reasons. Centers. Ground-truth data was also used to learn parameters of graphical models. Zhang and Seitz used an alternative optimization algorithm to estimate optimal values of Markov random field hyper parameters. Scharstein and Pal constructed a new dataset of 30 stereo pairs and used it to learn parameters of a conditional random field. Li and Huttenlocher presented a conditional random field model with a non-parametric cost function and used a structured support vector machine to learn the model parameters.

Recent work focused on estimating the confidence of the computed matching cost. Haeusler et al. used a random forest classifier to combine several confidence measures. Similarly, Spyropoulos et al. trained a random forest classifier to predict the confidence of the matching cost and used the predictions as soft constraints in a Markov random field to decrease the error of the stereo method.

Nowadays, more and more images are appearing and shared on the Internet. With such a large amount of images, we can rely on intelligent image understanding techniques to automatically process and analyze the images. Deep neural networks, more specifically the convolutional neural networks (CNNs), have been extensively studied for recognition, and understanding. CNN is a biologically inspired learning model. The features are learned end-to-end from raw data for classification or prediction. More specifically, CNN takes the raw images as input, and ensemble the feature learning and the training as a whole process. With a designed deep structure, CNN can effectively learn the complicated mapping relations between the raw image and the labels. Moreover, the spatial structure of images is adequately considered and used in CNN for regularization through restricted connectivity between layers (local filters), parameter sharing (convolutions),and special local invariance-building neurons (max pooling). Furthermore, parameters in local filters and between layers are connected and trained as a whole to encode some characteristics about human visual system (HVS), such as the edges and contours, which are vital for human to perceive and understand an image.

III. PROPOSED WORK

While studying the human visual and cognitive system, it is observed that it is composed of interconnected layers of neurons. The layered human visual system consists of simplex and complex cells determined by input signals. Convolution neural network resembles to human visual system so it is well suited for building a computational model of detecting visual saliency of images. In this paper we introduce a computational model for detecting visual saliency using artificial neural network model which extracts features of images. The neural network architecture is capable of extracting feature hierarchies from the image pixels automatically. To achieve feature extraction task neural network architecture has to be trained by image dataset. The feature extracted by the neural network having wide semantic information which is helpful in detecting visual saliency.

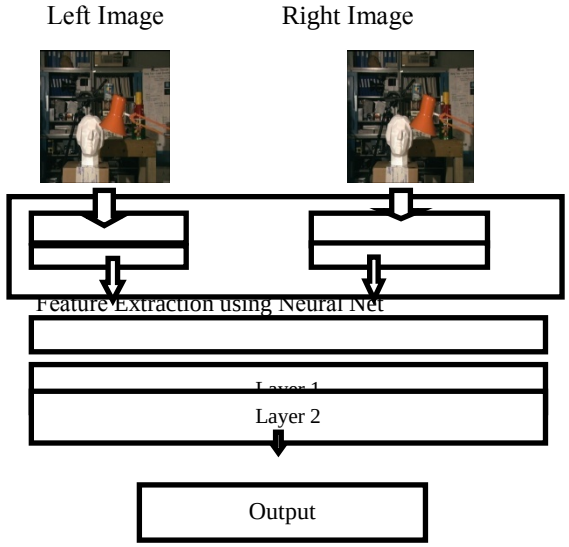


Fig.1. Layered Neural Network Architecture for saliency Detection

As in most other saliency systems, we base our computations mainly on intensity and features. This is in correspondence to human perception, since color is one of the basic features that guide visual attention. It is likely

that it is useful for other applications such as modeling eye fixations. The color computation is performed in an opponent color space, which corresponds to the opponent theory of human perception. This theory states that there are three opponent channels in the human visual system: red versus green, blue versus yellow, and black versus white. We have experimented with the LAB color space, but obtained better results with the simple color space in the intensity channel is obtained by I = R+G+B /3 , and the two color channels by RG = R − G and BY = B − R+G/ 2 . We can treat now all three channels I, RG, and BY equally to determine feature-specific saliencies.

We extract features for each image region with a deep convolutional neural network originally trained over the Image dataset using, an open source framework for CNN training and testing. A CNN pre-trained on large image data-set can be exploited as generic feature extractor through learning process. In learning process parameters of first n layers of source (pre-trained CNN) are transferred to the first n layers of target (new task) network and left without updates during training on new data-set, while the rest of the layers known as adaptation layers of target task are randomly initialized and updated over the training. If a fine-tuning strategy is taken then back propagation process will be carried out through the entire (copied + randomly initialized layers) network for calibrating the parameters of the copied layers in the new network so that the CNN responses well to the new task. In this experiment, we take pre-trained networks and extract features from their respective penultimate layers. These networks have been trained on ImageNet2, where the final logits layer of each network has 1000 output neurons. That final layer is decapitated, and then rest of the CNN is employed as fixed feature extractor on the new data-sets, where number classes per data-set may differ.

The architecture of this CNN has eight layers including five convolutional layers and three fully-connected layers. Features are extracted from the output of the second last fully connected layer, which has neurons. The CNN was originally trained on a dataset for visual recognition; automatically extracted CNN features turn out to be highly versatile and can be more effective than traditional handcrafted features on other visual computing tasks. Since an image region may have an irregular shape while CNN features have to be extracted from a rectangular region, to make the CNN features only relevant to the pixels inside the region, as in, we define the rectangular region for CNN feature extraction to be the bounding box of the image region and fill the pixels outside the region but still inside its bounding box with the mean pixel values at the same locations across all training images. These pixel values become zero after mean subtraction and do not have any impact on subsequent results. The warped RGB image region is then fed to the deep CNN and a 4096-dimensional feature vector is obtained by forward propagating a mean-subtracted input image region through all the convolutional layers and fully connected layers. We name this vector feature A.

Feature A itself does not include any information around the considered image region, thus is not able to tell whether the region is salient or not with respect to its neighborhood as well as the rest of the image. To include features from an area surrounding the considered region for understanding the amount of contrast in its neighborhood, we extract a second feature vector from a rectangular neighborhood, which is the bounding box of the considered region and its immediate neighboring regions. All the pixel values in this bounding box remain intact. Again, this rectangular neighborhood is fed to the deep CNN after being warped. We call the resulting vector from the CNN feature B. As we know, a very important cue in saliency computation is the degree of (color and content) uniqueness of a region with respect to the rest of the image. The position of an image region in the entire image is another crucial cue. To meet these demands, we use the deep CNN to extract feature C from the entire rectangular image, where the considered region is masked with mean pixel values for indicating the position of the region. These three feature vectors obtained at different scales together define the features we adopt for saliency model training and testing. Since our final feature vector is the concatenation of three CNN feature vectors, we call it S-3CNN.

IV. NEURAL NETWORK TRAINING

In extraction of CNN features, we train a neural network with one output layer and two fully connected hidden layers. This network plays the role of a repressor that infers the saliency score of every image region from the CNN features extracted for the image region. It is well known that neural networks with fully connected hidden layers can be trained to reach a very high level of regression accuracy. Concatenated CNN features are fed into this network, which is trained using a collection of training images and their labelled saliency maps that have pixel wise binary saliency scores. Before training, every training image is first decomposed into a set of regions. The saliency label of every image region is further estimated using pixel wise saliency labels. During the training stage, only those regions with 70% or more pixels with the same saliency label are chosen as training samples, and their saliency labels are set to either 1 or 0 respectively. During training, the output layer and the fully connected hidden layers together minimize the least-squares prediction errors accumulated over all regions from all training images. Traditional regression techniques, such as support vector regression and random

forests, can be further trained on this feature vector to generate a saliency score for every image region.

V. LEARNING OF STEREOSCOPIC IMAGE WITH CNN

Nowadays, convolutional neural networks (CNNs) have been successfully employed to learn the image representation for various applications, such as image classification, object detection, human parsing and activity recognition. In this paper, we rely on CNN for learning local structures of the stereoscopic images. The structures are learned via multiple layers of convolution and max-pooling, which are expected to be sensitive to the quality perception of the stereoscopic images.

The stereoscopic images differ from the 2D natural images, as the left and right views together can provide depth perception. Therefore, perceptual evaluation of the stereoscopic images needs to consider the information from both the left and right views. We propose two CNNs to fully exploit the structures of the stereoscopic images, which are expected to be sensitive for quality perception. As demonstrated in, the difference image between the left view and right view is more important than the left and right views for quality assessment. After performing two layers of convolution and pooling processes, the final representation is obtained. MLP with two fully-connected layers are utilized to summarize the representation and generate the final score as follows:

$$S = \omega_S(\sigma(\omega_h(\theta_{im}) + b_h)) + b_s$$

Where σ is the nonlinear activation function. θ_{im} denotes the learned representation with two layers of convolution and max-pooling. ω_h and b_h are used to map the obtained image representation θ_{im} to the representation in the hidden layer. ω_s and b_s are the parameters to compute the final score of the input image patch. S is the learned score to indicate the perceptual quality of the input image patch.

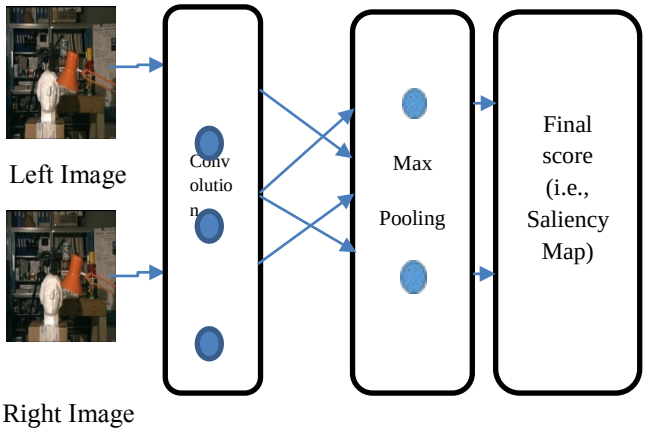


Fig.2. Learning of Stereoscopic image

Given left and right image, the saliency map is computed in following steps:

1. Input process of images and object perception by CNN;
2. Image boundary information propagation within the max pooling with super-pixel graph;
3. Coarse-grained saliency information fusion; and
4. Fine grained saliency map generation by nonlinear regression-based propagation, as illustrated in Figure 2.

The trained FCNN is used to adaptively capture the semantic structural information on object perception, resulting in a pixel-wise object ness probability map (ranging from 0 and 1), which we refer as a Deep Map. This stage focuses on modelling the underlying object properties from the perspective of foreground discovery using CNN. In contrast, the stage 2) aims to explore the influence of the image boundary information in saliency detection from the viewpoint of background propagation to estimate the saliency values on the super-pixel level where the ones on the image boundary are initialized -1 and others as 0. After the propagation process, we have a saliency map denoted as Boundary then we perform saliency fusion of the Deep Map and Boundary Map to generate the refinement over the super-pixel graph, resulting in the final fine-grained saliency map.

VI. EXPERIMENTAL SETUP

Detailed architecture of the proposed method can be found in the supplementary materials1. We pre-train the RFCN on the PASCAL VOC 2010 semantic segmentation data set with 10103 training images belonging to 20 object classes. The pre-training is converged after 200k iterations of SGD. We then fine-tune the pre-trained model for saliency detection on the THUS10K data set for 100k iterations. In the test stage, we apply the trained RFCN in three different scales and fuse all the results into the final saliency maps. Our method is implemented in MATLAB and runs at 4.6 seconds per image on a PC with a 3.4 GHz CPU and a TITANX GPU. The source code will be released.

We evaluate the proposed algorithm (RFCN) on five benchmark data sets: SOD, ECSSD, PASCAL-S, SED1, and SED2. The evaluation result on SED2 and additional analysis on the impact of recurrent time step are included in the supplementary materials. Three metrics are utilized to measure the performance, including precision-recall (PR) curves, F-measure and area under ROC curve (AUC). The precision and recall are computed by thresholding the saliency map, and comparing the binary map with the ground truth. The PR curves demonstrate the mean precision and recall of saliency maps at different thresholds. The F-measure can be calculated by

$$F_\beta = \frac{(1+\beta^2)\,\mathrm{Pr}\,ecision \times \mathrm{Re}\,call}{\beta^2\,\mathrm{Pr}\,ecision - \mathrm{Re}\,call}$$

Where Precision and Recall are obtained using twice the mean saliency value of saliency maps as the threshold, and set β 2= 0:3.

The precision refers to the fraction of salient pixels which are assigned correctly in the detected saliency maps. While the recall refers to the fraction of correct salient pixels in the ground truth:

$$\mathrm{Pr}\,ecision = \frac{|M \cap C|}{|M|}, \mathrm{Re}\,call = \frac{|M \cap C|}{|C|}$$

All the precision and recall scores are combined to plot the PR curve, Receiver operating characteristics (ROC). The ROC curve is generated based on true positive rates (TPR) and false positive rates (FPR) when binarizing saliency maps with a set of fixed thresholds:

$$TPR = \frac{|M \cap C|}{C}, FPR = \frac{|M \cap C|}{|C|}$$

Where G denotes the oppositive of the ground truth G. The ROC curve plots the TPR versus FPR by varying the threshold. Area under ROC curve (AUC). The AUC score is computed as the area under the ROC curve. A perfect AUC performance gets a score of 1, while the AUC performance of random guessing gets a score of 0.5.

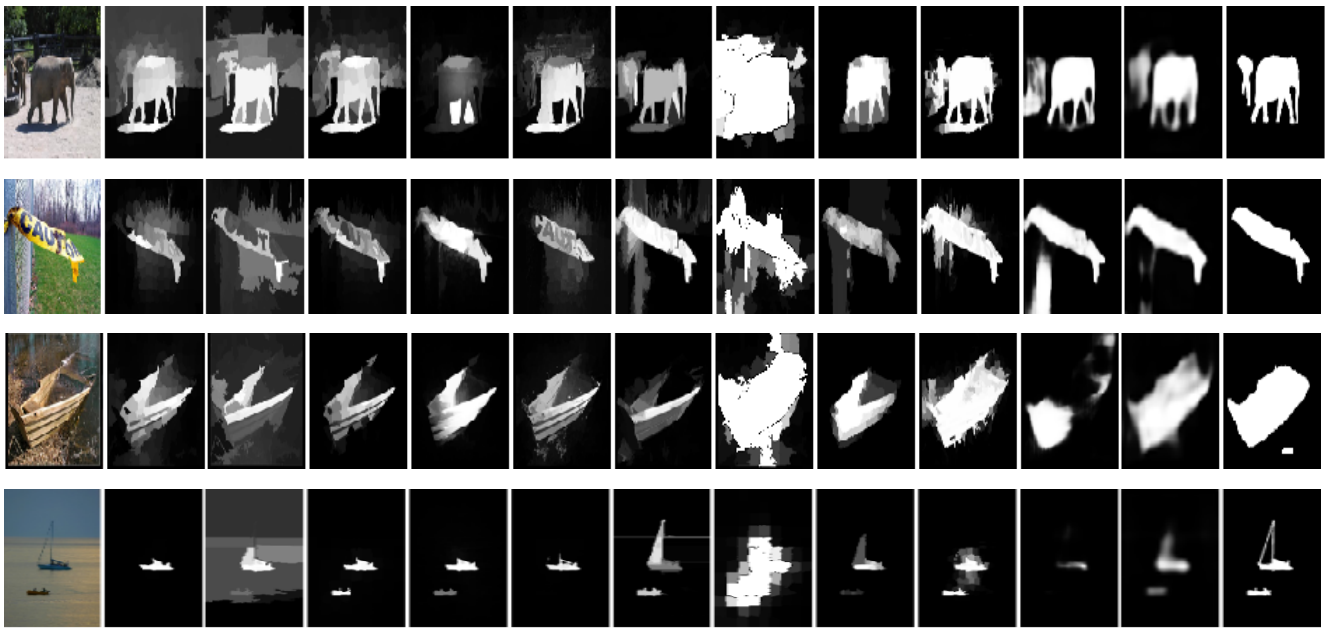
The MAE score is calculated as the mean of pixel-wise absolute errors between the saliency map S and the ground truth G:

$$MAE = \frac{1}{W_I / H_I} \sum_{x=1}^{W_I} \sum_{y=1}^{H_I} |S(x,y) - G(x,y)|$$

where WI and HI are the width and height of the saliency map S. S(x; y) and G(x; y) are the continuous saliency score and the binary ground truth at pixel (x; y), which are normalized in the range [0; 1]. Smaller MAE score means better performance.

VII. PERFORMANCE COMPARISION

We compare the proposed method with 10 recent state-of the-art methods on aforementioned datasets, which include SR, FT, SF, GS, HS, RC, MR, wCtr, DRFI, GMR (Figure 3). We use either the implementations or the saliency maps provided by the authors for fair comparison. We present the comparison results from both qualitative and quantitative aspects for comprehensively revealing the characteristics of our method. Table 1 With respect to AUC, F-measure, MAE and runtime on the SOD, ECSSD, PASCAL-S datasets. Our proposed methods rank first and second on the taken data sets. As can be easily seen, the best detection precisions are all obtained by the deep learning based methods. From Table 1, we can find that our method makes a significant improvement in processing speed compared with other deep learning based methods. Our method can be comparable with best approaches in terms of F-measure and MAE.



(a)Source(b)SR (c)FT (d)SF (e)GS (f)HS (g)RC (h)MR (i)wCtr (j)DRFI (k)GMR (l)Ours (m)GT

Fig.3. Visual comparison of saliency maps generated from 10 different methods, including ours. The ground truth (GT) is shown in the last column. MDF consistently produces saliency maps closest to the ground truth. We compare MDF against spectral residual (SR), frequency-tuned saliency (FT]), saliency filters (SF), geodesic saliency (GS), hierarchical saliency (HS), regional based contrast (RC),

manifold ranking (MR), optimized weighted contrast (wCtr) , discriminative regional feature integration (DRFI) and graph-based manifold ranking(GMR).

TABLE 1. COMPARISON OF QUANTITATIVE RESULTS INCLUDING AUC, F-MEASURE (LARGER IS BETTER) AND MAE, RUNTIME (SMALLER IS BETTER).

Measure	SR	FT	SF	GS	HS	RC	MR	wCtr	DRFI	GMR	Ours
AUC	0.862	0.85	0.897	0.871	0.856	0.923	0.913	0.874	0.943	0.924	0.954
F-measure	0.611	0.616	0.63	0.589	0.606	0.68	0.677	0.668	0.716	0.717	0.737
MAE	0.231	0.264	0.201	0.198	0.219	0.165	0.148	0.157	0.123	0.109	0.128
Runtime(s)	0.31	0.322	0.142	0.01	1.178	1.3	1.625	1.214	0.575	0.135	0.029

VIII. CONCLUSION

In this paper, we propose a fast and compact saliency detection method to meet the essential application requirement of salient object detection task. After training the network can directly predict a dense full-resolution saliency map when being fed into an image. The end-to-end work manner effectively simplifies the processing procedure. By comprehensive experiments, we verify that our method can achieve comparable or better precision performance than the state-of-the-art methods while get a significant improvement in detection speed (processing in real time). The compact architecture, fast processing speed, small parameter storage and decent precision performance make it possible to employ our method practically as a pre-processing step before other visual tasks. In learning process parameters of first n layers of source (pre-trained CNN) are transferred to the first n layers of target (new task) network and left without updates during training on new data-set, while the rest of the layers known as adaptation layers of target task are randomly initialized and updated over the training. In the proposed method our goal is to analyze if the accuracy improves when multiple CNN bottleneck features are fused. We conclude that, without manual feature selection, effective features for edit propagation can be automatically extracted by (i) deep learning from user inputs in a single image and (ii) efficiently learning the CNN in propagation system using deep features has generated better results than previous work in several applications such as grayscale image colorization, image recoloring, and foreground segmentation.

REFERENCES

[1] X. Liu, W. Yang, L. Lin, Q. Wang, Z. Cai, J. Lai (2015) “Data-driven scene understanding with adaptively retrieved exemplars”, IEEE Multimeda.

[2] X. Liang, S. Liu, X. Shen, J. Yang, L. Liu, L. Lin, S. Yan (2015) “Deep human parsing with active template regression”, IEEE Trans. Pattern Anal. Mach. Intell.

[3] G. Li and Y. Yu (2015) “Visual saliency based on multiscale deep features,’ in Proc. IEEE Conf. CVPR, pp. 5455–5463.

[4] Dingwen Zhang et al (2015) “Co-saliency Detection via Looking Deep and Wide”, in IEEE Conf. CVPR.

[5] Rakesh Y (2017) “Super-Pixel Based Saliency In 3d-Image Object Detection Using Content Based Image Retrieval”, Journal Of Theoretical And Applied Information Technology 15th February.

[6] Walid Hachicha et al., (2012) “Combining Depth Information And Local Edge Detection For Stereo Image Enhancement” in 20th European Signal Processing Conference (EUSIPCO).

[7] Yuting Zhang (2015) “Improving Object Detection with Deep Convolutional Networks via Bayesian Optimization and Structured Prediction” in Proc. IEEE Conf. CVPR, pp. 249-258.

[8] Christoph Feichtenhofer (2015) “Dynamically Encoded Actions based on Spacetime Saliency” in IEEE Conf. CVPR.

[9] R. Arandjelovic, A. Zisserman (2012) “Three things everyone should know to improve object retrieval”, Proc. CVPR,

[10] Chen Gong et al (2015) “Saliency Propagation from Simple to Difficult” in IEEE Conf. CVPR.

[11] Wei Zhang et al. (2016) “Learning structure of stereoscopic image for no-reference quality assessment with convolutional neural network”, in Pattern Recognition PP 176–187

[12] Hyun Soo Park and Jianbo Shi (2015) “Social Saliency Prediction”, in IEEE Conf. CVPR.

[13] Lijun Wang (2015) “Deep Networks for Saliency Detection via Local Estimation and Global Search”, in IEEE Conf. CVPR.

[14] Jingbo Zhou (2012) “Multiscale saliency detection using principle component analysis” in WCCI 2012 IEEE World Congress on Computational Intelligence June, 10-15, Brisbane, AustraliaYao Xiao et al. (2015) “Complexity-Adaptive Distance Metric for Object Proposals Generation” in IEEE Conf. CVPR.

[15] Yuki Endo et al. (2016) “DeepProp: Extracting Deep Features from a Single Image for Edit Propagation” in EUROGRAPHICS / J. Jorge and M. Lin Volume 35, Number 2. A.Borji, M.-M. Cheng, H. Jiang, and J. Li (2015) “Salient object detection: A benchmark”, IEEE Trans. Image Process., vol. 24, no. 12, pp. 5706– 5722.

[16] S. Battiato, G. M. Farinella, G. Puglisi, and D. Ravi (2014) “Saliency-based selection of gradient vector flow paths for content aware image resizing,” IEEE Trans. Image Process., vol. 23, no. 5, pp. 2081–2095.

[17] W. Zhu, S. Liang, Y. Wei, and J. Sun (2014) “Saliency optimization from robust background detection”, CVPR.

[18] M.-M. Cheng, N. J. Mitra, X. Huang, P. H. S. Torr, and S.-M. Hu (2014) “Global contrast based salient region detection”. TPAMI,

[19] M. Jiang, S. Huang, J. Duan, and Q. Zhao. Salicon (2015) “Saliency in context”, In IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

[20] X. Huang, C. Shen, X. Boix, and Q. Zhao. Salicon (2015) “Reducing the semantic gap in saliency prediction by adapting deep neural networks”. In IEEE International Conference on Computer Vision (ICCV).

[21] F. Visin, K. Kastner, K. Cho, M. Matteucci, A. C. Courville, and Y. Bengio (2015) “Renet: A recurrent neural network based alternative to convolutional networks”, CoRR, vol. abs/1505.00393.

[22] J. Pont-Tuset, F. Perazzi, S. Caelles, P. Arbel´aez, A. Sorkine-Hornung, and L. Van Gool (2017) “The 2017 davis challenge on video object segmentation”, arXiv:1704.00675.