A Benchmark of Computational Models of Saliency to Predict Human Fixations in Videos

Shoaib Azam[1], Syed Omer Gilani[2], Moongu Jeon[1], Rehan Yousaf[2] and Jeong Bae Kim[3]
[1]*School of Information and Communication, Gwangju Institute of Science and Technology, Gwangju, Republic of Korea*
[2]*National University of Sciences and Technology, Islamabad, Pakistan*
[3]*Department Management of Technology, Pukyong National University, Busan, Republic of Korea*

Keywords: Saliency Benchmark of Videos, Scoring Metrics, Fixation Maps.

Abstract: In many applications of computer graphics and design, robotics and computer vision, there is always a need to predict where human looks in the scene. However this is still a challenging task that how human visual system certainly works. A number of computational models have been designed using different approaches to estimate the human visual system. Most of these models have been tested on images and performance is calculated on this basis. A benchmark is made using images to see the immediate comparison between the models. Apart from that there is no benchmark on videos, to alleviate this problem we have a created a benchmark of six computational models implemented on 12 videos which have been viewed by 15 observers in a free viewing task. Further a weighted theory (both manual and automatic) is designed and implemented on videos using these six models which improved Area under the ROC. We have found that Graph Based Visual Saliency (GBVS) and Random Centre Surround Models have outperformed the other models.

1 INTRODUCTION

Saliency has been discussed and implemented in so many ways on images and many computational models have been defined in this respect. Each computational model is at best as an individual technique and also has proven the psychology of the human visual system. There is no such method to judge an individual technique until and unless it is compared to some standard or state of the art. But comparing different models gives you the basic idea about how good a model is and how bad a model is.

Previously, many people have done comparison between models and got results to improve the technique but this has been done only on static images (Judd et al., 2012); (Privitera and Stark, 2000); (Parkhurst et al., 2002); (Ouerhani, 2003); (Elazary and Itti, 2008); (Henderson et al., 2007); (Bruce and Tsotsos, 2006); (Itti, 2005); (Peters et al., 2005); (Yubing et al., 2010); (Yubing et al., 2011). Our goal is to pick up certain saliency models (Borji and Itti, 2010) and apply them on videos. After applying these models on the videos we will compare these models with each other and generate fixation map, through which we compute the Area Under Receiver Operating Characteristic Curve (AUC score) (Borji et al., 2013a); (Borji et al., 2013b).

(AUC) and set our own benchmark in which we will only be dealing with videos. Further, we have designed a weighted algorithm for saliency computation. It includes both manual and automatic weight assigning. For this AUC is also calculated.

For performing our goal we have taken a dataset (Hadizadeh et al., 2012) which includes 15 observer and 6 models which we will be comparing. These models will be applied on 12 videos having the following details (name and frames) are Foreman (300 frames), Bus (150 frames), City (300 frames), Crew (300 frames), Flower Garden (250 frames), Mother and Daughter (300 frames), Soccer (300 frames), Stefan (90 frames), Mobile Calendar (300 frames), Harbor (300 frames), Hall Monitor (300 frames) and Tempete (260 frames).The dataset which we are using is generated with a free viewing task, it states that the participants are not restricted to see particular object in the scene.

No doubt the models are very near to the human visual system but a small difference still shows there is a big room for improvement.

Azam, S., Gilani, S., Jeon, M., Yousaf, R. and Kim, J.
A Benchmark of Computational Models of Saliency to Predict Human Fixations in Videos.
DOI: 10.5220/0005678701340142
In *Proceedings of the 11th Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2016) - Volume 4: VISAPP*, pages 134-142
ISBN: 978-989-758-175-5

2 CONTRIBUTION

We have made the following contribution to the computational models and the human fixations:

- We have used only 6 saliency models (which are the major attention models in last ten years).
- We have applied these models on 12 videos.
- For each video with having specified number of frames we have computed their saliency maps with all the said 6 models.
- We have used the eye tracking dataset of 15 observers for creating the fixation map for each video.
- We have calculated the AUC (Borji et al., 2013a); (Borji et al., 2013b) score as scoring metric for all the videos, each containing specified frames.
- Most importantly we have designed our own "weighted optimal algorithm" which assigns weight to the models automatically also manually as "user weighted algorithm" and computed the AUC (Borji et al., 2013a); (Borji et al., 2013b) scores, hence giving out weighted combination of models showing the best possible results.

3 SALIENCY MAP AND COMPUTATIONAL MODELS

All the computational models follow the same procedure in computing the saliency map. The main ides was first introduced by Itti and Koch (Koch and Ullman, 1985). The saliency map is basically generated on the basis of features in parallel and by combining the features values gives us the saliency map.

To compute different features at different scales the model computes many pyramids of the input image. Most commonly used image features are intensity, colour and orientation. These features are again divided into sub features. The feature maps are generated through centre surround method or difference of Gaussian (DoG) (Young, 1987). These features map are added up and finally normalized and weighted combined giving up the saliency map. This technique is purely bottom up. Models do use top down technique too (Borji and Itti, 2010).

In videos models are applied to each frame for computing the saliency map. Then these frames are again synchronized to the videos. The procedure for generating the saliency map varies a little bit according to the each computational model. But the main idea remains the same.

The models which we are using for generating saliency map are as follows.

1) *The first model which we are using is by Achanta et al., (2008). This model uses a simple technique which extracts objects from the background. This is done through subtraction of common part which is the background.*
2) *Secondly we have context aware saliency (Goferman et al., 2010) model. This model emphasizes on locating the most important part of the image which has the main contents of information but not just salient object.*
3) *Saliency detection based on wavelet (Nevrez et al., 2013) is the third model. This model focuses on extracting the low level features through wavelet transform then it creates the feature map which clearly points out the features such as edges and texture.*
4) *Fourthly we have graph based visual saliency model (Harel et al., 2007) (GBVS) is a two steps procedure which contains making of activation maps on specific feature channels and then normalizes it according to the others map which it has to combine with.*
5) *The fifth model we have is called random surround (Vikram et al., 2012) model. This model follows a simple procedure to calculate the image saliencies which consists of computing local saliencies over some random regions in the rectangular shape according to the task interest. Then the Gaussian filter is applied to remove the noise from the images. Colour space is generated and divided in to L*,a* and b* channels. Saliency maps are generated in these channels. Besides n random sub windows are produced over each of L*,a* and b* channel. The final saliency map is generated by fusing the Euclidean norm of saliency calculated in three above channels.*
6) *The last model is the spatio-temporal saliency detection through Fourier transforms (Guo et al., 2008). In this model the technique uses the phase spectrum to calculate the saliency map. It is similar in procedure to the amplitude spectrum used by spectral residual model but varies in the component of focus that is phase.*

4 PROPOSED METHODOLOGY

Our proposed methodology consists of 6 saliency computational models which we are applying on videos and computing their saliency maps. The computational models which we are using are Spatio-

temporal Saliency Detection through Fourier transform (Guo et al., 2008), Random Surround model (Vikram et al., 2012), Saliency detection on Wavelets (Nevrez et al., 2013), Graph based Visual Saliency Model (Harel et al., 2007), Context Aware Saliency Model (Goferman et al., 2010) and Achanta et al., (2008) model. These 6 models are applied to 12 video sets each containing specified frames. For instance, the first model is applied to a 12 video which gives us the saliency maps according to number of frames of the video and by using the similar pattern all the models have been applied to all the videos each giving us saliency maps depending on their number of frames.

Now the dataset (Hadizadeh et al., 2012) which we are using have been viewed by 15 observers in a free viewing scenario and their eye tracking data is used to compute fixation map for frames of a video. These fixation maps are computed for all the 12 videos sets. So we have the saliency maps for all the videos along with the fixation maps. These fixation maps are then used for the calculation of AUC score.

We are using the Ali Borji's (Borji et al., 2013a); (Borji et al., 2013b) implementation to compute the AUC score to compare the results of the machine visual system and human visual system. We have calculated the score for each video containing different number of frames for all the 6 models individually. This process gives us the comparison of each model with another and its comparison with the human visual system.

After computing the AUC score we have assigned weights to the models. Mathematically weights are using the following equation.

$$S = \sum_{i=1}^{n} w_i M_i \quad (1)$$

In the above equation S is final saliency map computed, wi is weight assigned to model, Mi is the respective model and n is the total number of models. Here in our case n is equal to 6 as we have used six models. There are two methods of assigning the weights.

i) Automatic Weight Assigning
ii) Manual Assigning(User Assigning)

4.1 Automatic Weight Assigning

In automatic assigning there are three configurations for assigning the weights automatically. In first configuration of automatic weight assigning Model 1 has been assigned 0.8 weights and the rest of other models have been assigned 0.5, then model 2 with 0.8 weight and the others with 0.5 and this process carry on iteratively. Using this arrangement 6 combination has been made. In the second configuration 0.5 and 0.2 are the weights assigned to the models in the similar manner making 6 combinations. In the third configuration 0.8 and 0.2 are the weights assigned to the models again in the similar manner making 6 combinations. After weight assigning AUC score is computed with the weighted saliency map for all the videos. Basically we have chosen three types of weights 0.8 (maximum weight), 0.5(middle weight) and 0.2 (minimum weight).

4.2 Manual Weight Assigning

The second method for weight assigning is the manual weight assigning method. In this procedure the user enters the weights for each computational model depending upon the result of model. The process then computes the saliency map according to weights assigned by the user under the perspective of equation (1). After weight assigning, the AUC is used as a scoring metric to evaluate the saliency maps. Further, as we have used six models yet there is a room to add more models and see which models has fruitful result on videos.

5 FLOWCHART OF PROPOSED METHODOLOGY

The following flow diagrams show the overall procedure of proposed methodology.

6 RESULTS AND DISCUSSION

6.1 Saliency Models Results

The above results shows the saliency maps of six videos which has been obtained by implementing six models on them. By observing these models we can say that the models whose results are clearer and distinct may have better approximation or nearer to human visual system, however this is not the case and can be analysed by the results of AUC.

6.2 Model Performance for Videos

The above table shows the mean AUC score for the 6 models which have been computed on the 12 videos. As it can be seen from the table that in different videos different model has outperformed the other models, but in most of the cases it is GBVS and

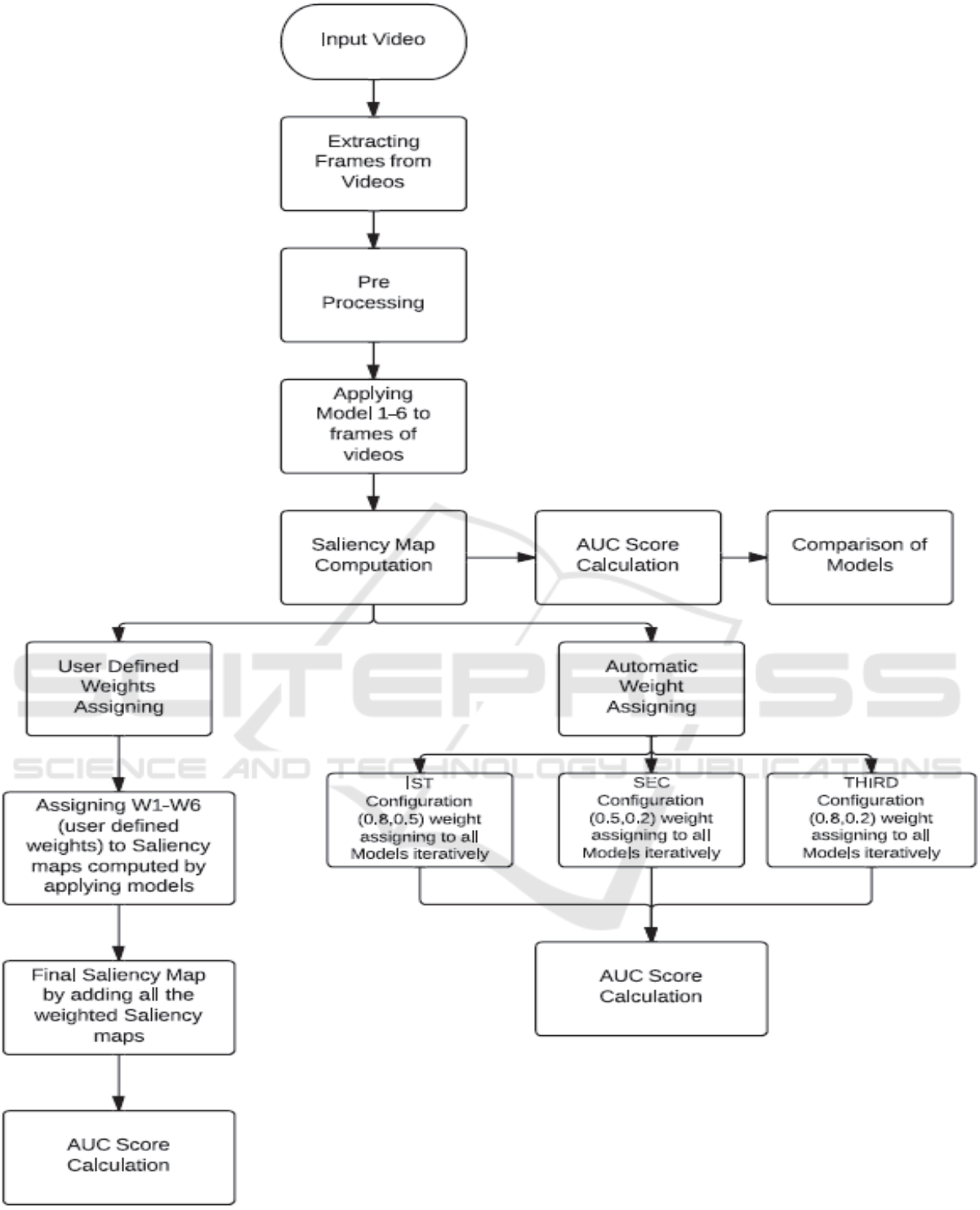

Figure 1: Flowchart of Saliency Map Computation and User Defined and Automatic Weight Assigning.

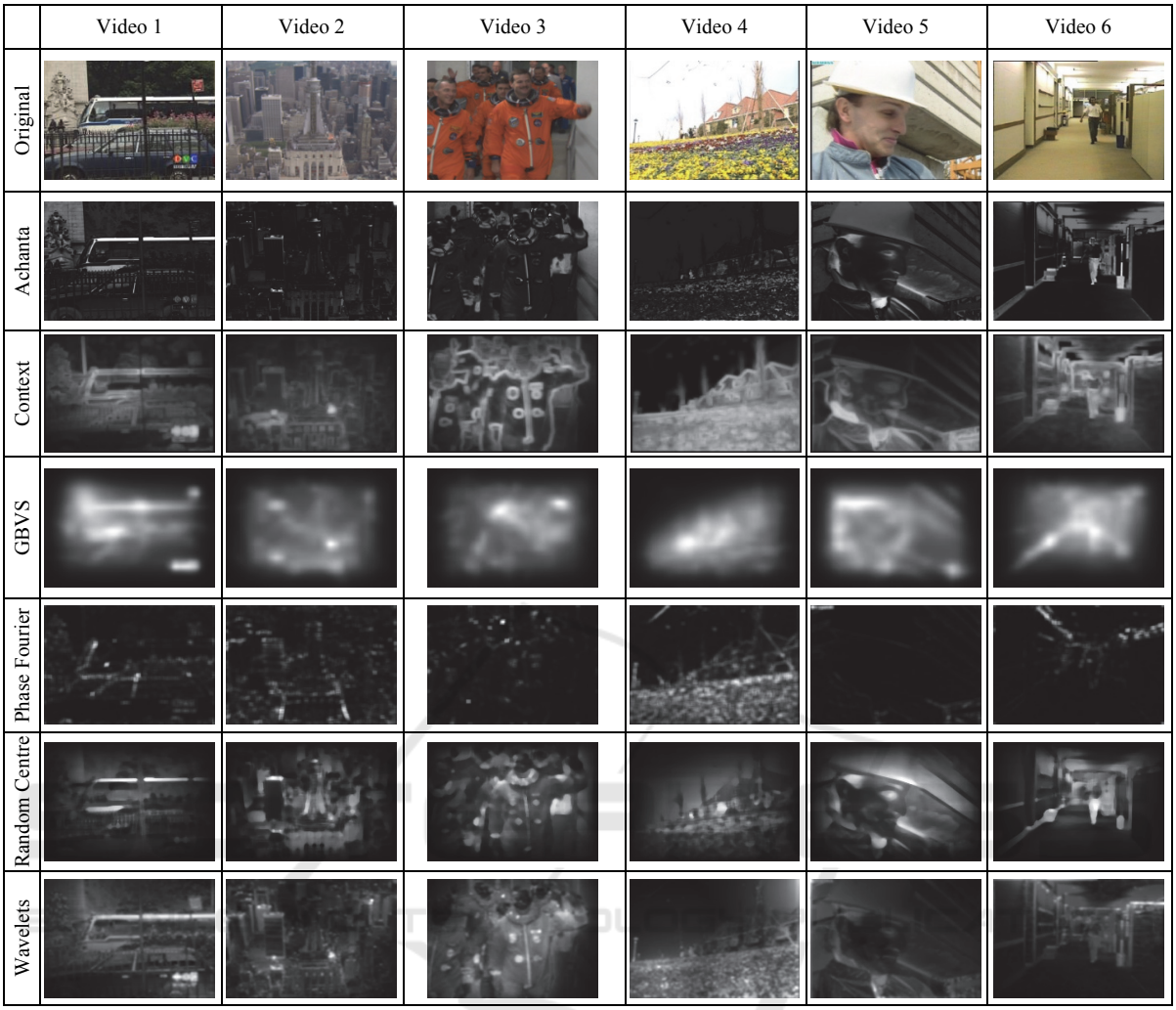

Figure 2: Saliency maps from 6 models of 6 videos.

Table 1: Performance of all six models using AUC implemented on videos. Name of video is also given. For AUC, higher values are better.

AUC Score of Models Implemented on Videos							
S.NO	Video Name	Achanta	Context	GBVS	Phase Fourier	Random Surround	Wavelets
1	Bus	0.4574	0.6700	0.7235	0.4825	**0.7446**	0.5504
2	City	0.6114	0.6460	0.7484	0.5425	**0.7756**	0.6418
3	Crew	0.5085	0.5344	**0.5972**	0.4843	0.5302	0.5407
4	Flower Garden	0.4346	0.5104	**0.8283**	0.4426	0.8156	0.5316
5	Foreman	0.5153	0.6667	**0.7614**	0.5193	0.6072	0.3410
6	Hall Monitor	0.4061	0.4390	0.4864	0.4816	**0.4879**	0.3479
7	Harbor	0.4846	0.3920	**0.8234**	0.4385	0.7339	0.5232
8	Mobile	0.3297	0.3495	**0.6372**	0.4777	0.6004	0.3914
9	Mother	0.4093	0.6467	0.6884	0.4906	0.6713	**0.7111**
10	Soccer	0.6764	0.3749	0.7108	0.4670	**0.7616**	0.5440
11	Stefan	0.5802	0.8743	**0.9269**	0.6349	0.8735	0.6037
12	Tampete	0.5150	0.7362	0.7287	0.6573	**0.7415**	0.6706

Random Surround Model that outperformed the other models. Like in videos (named) BUS and City, Random Surround Model has the better AUC and it is approaching to one. However, in videos (named) Crew, Flower Garden and Foreman, GBVS has better results as compared to other one. There is only one video named Mother, in which Wavelet based model has better results.

6.3 Manual Weight Assigning (User's Defined) Results

In manual weight assigning table, the mean AUC scores of all videos except video 6, 8 and 10 have been increased. This weight assigning is random depending upon the result of models (shown to user) before assigning weights. For this assigning we have selected or assigned the following weights to the models.

Table 2: Performance Score of Manual Weight Assigning (AUC Score).

AUC Score of Manual Weight Assigning		
S.NO	Video Name	AUC Score
1	Bus	0.7237
2	City	0.6877
3	Crew	0.6193
4	Flower Garden	0.7279
5	Foreman	0.7581
6	**Hall Monitor**	**0.3603**
7	Harbor	0.7818
8	**Mobile**	**0.5797**
9	Mother	0.7066
10	**Soccer**	**0.3198**
11	Stefan	0.8391
12	Tampete	0.7291

6.4 Automatic Weight Assigning Results

The above tables show the mean AUC scores of IST, SEC and THIRD configuration of automatic weight assigning. The assigning has been made using the probabilities as 0.8, 0.5 and 0.2 as maximum, medium and minimum to the models in specified configuration. Using this behaviour three configuration has been made. If we see the IST configuration of automatic weight assigning, the weights are assigned as 0.8 and 0.5 in a manner that that in first module (MODULE 1) the weights assigned to models are shown in Table 7. Using the same technique the modules of other two configuration has been achieved.

By considering the AUC scores of IST, SEC and THIRD configuration with the AUC scores of original videos, we can see that there is significant amount of increased. However in some cases, there is decreased in the AUC scores of entire videos due to error in obtaining the fixation data for it. This can be seen in Video 6. So, by viewing it overall there is 20 % improvement in the AUC scores using automatic weight assigning. However the individual increase in some cases is above 50 %. But we are more concerned by overall result. Depending upon this the best configuration which gives this performance is SEC configuration.

7 CONCLUSION

We have performed the comparison between 6 different computational models and their saliency maps. We have also computed the AUC of the saliency map for the accuracy of computational models for videos.

We have designed the weighted algorithm for the saliency maps and using this we have computed their AUC score as well. The weighted method is for both user defined weights and automatic weights assigning, thus make the algorithm more flexible and leaving a room for further enhancement and addition of further models, comparison in future work

8 FUTURE WORK

The future work will include that more videos dataset can be used for both parts. Further video dataset where observer is explicitly asked to view a certain object will be included, so that the performance of computational model is being analysed in not free viewing videos. More saliency computational models will be used to predict the human fixation in videos.

ACKNOWLEDGEMENTS

This work was supported by Institute for Information & communications Technology Promotion (IITP) grant funded by the Korea Government (MSIP) (No.B0101-15-0525, Development of global multi-target tracking and event prediction techniques based on real-time large-scale video analysis) and Centre for Integrated Smart Sensors as Global Frontier (CISS-2013M3A6A6073718).

Table 3: Weight Assigned to Models in Manual Weight Assigning.

Weight Assigned to Models						
Models	Achanta	Context	GBVS	Phase Fourier	Random Surround	Wavelets
Weights(0-1)	0.5	0.6	0.8	0.4	0.7	0.8

Table 4: Performance Score of Automatic Weight Assigning of IST Configuration (AUC Score).

AUC Score of IST Configuration of Automatic Weight Assigning							
S.NO	Video Name	Module 1	Module 2	Module 3	Module 4	Module 5	Module 6
1	Bus	0.7043	0.6922	0.6953	0.7097	0.6977	0.6968
2	City	0.6853	0.6848	0.6878	0.6898	0.6867	0.6799
3	Crew	0.6448	0.6409	0.6284	0.6300	0.6428	0.6239
4	Flower Garden	0.7378	0.7387	0.7252	0.7246	0.7335	0.7192
5	Foreman	0.7770	0.7726	0.7664	0.7791	0.7712	0.7690
6	Hall Monitor	0.3649	0.3573	0.3655	0.3728	0.3574	0.3543
7	Harbor	0.7698	0.7673	0.7671	0.7783	0.7778	0.7871
8	Mobile	0.5749	0.5706	0.5678	0.5817	0.5744	0.5636
9	Mother	0.6712	0.6790	0.6810	0.6834	0.7047	0.6896
10	Soccer	0.5451	0.5454	0.5449	0.5449	0.5451	0.5451
11	Stefan	0.8332	0.8339	0.8572	0.8411	0.8494	0.8734
12	Tampete	0.7235	0.7180	0.7189	0.7211	0.7315	0.7185

Table 5: Performance Score of Automatic Weight Assigning of SEC Configuration (AUC Score).

AUC Score of SEC Configuration of Automatic Weight Assigning							
S.NO	Video Name	Module 1	Module 2	Module 3	Module 4	Module 5	Module 6
1	Bus	0.6948	0.7044	0.7063	0.7280	0.7139	0.6967
2	City	0.7015	0.6966	0.6996	0.7186	0.6927	0.6977
3	Crew	0.6510	0.6373	0.6208	0.6277	0.6374	0.6199
4	Flower Garden	0.7107	0.7428	0.7199	0.7274	0.7051	0.7278
5	Foreman	0.7522	0.7747	0.7779	0.7760	0.7821	0.7788
6	Hall Monitor	0.3683	0.3685	0.3631	0.3730	0.3746	0.3578
7	Harbor	0.7700	0.7581	0.7756	0.7490	0.7588	0.7801
8	Mobile	0.5873	0.6111	0.6023	0.5784	0.5880	0.6095
9	Mother	0.6850	0.7022	0.6817	0.6821	0.6812	0.6787
10	Soccer	0.5442	0.5449	0.5437	0.5447	0.5449	0.5444
11	Stefan	0.8483	0.8664	0.8631	0.8689	0.8693	0.8545
12	Tampete	0.7273	0.7291	0.7167	0.7396	0.7266	0.7270

Table 6: Performance Score of Automatic Weight Assigning of THIRD Configuration (AUC Score).

AUC Score of THIRD Configuration of Automatic Weight Assigning							
S.NO	Video Name	Module 1	Module 2	Module 3	Module 4	Module 5	Module 6
1	Bus	0.6936	0.7098	0.7001	0.6916	0.7021	0.7088
2	City	0.7089	0.6850	0.6999	0.6848	0.6833	0.6686
3	Crew	0.6284	0.6408	0.6341	0.6318	0.6306	0.6353
4	Flower Garden	0.7042	0.7296	0.7271	0.7224	0.7247	0.7222
5	Foreman	0.7648	0.7685	0.7710	0.7745	0.7641	0.7558
6	Hall Monitor	0.3546	0.3629	0.3626	0.3560	0.3660	0.3717
7	Harbor	0.7702	0.7717	0.7604	0.7797	0.7802	0.7688
8	Mobile	0.5756	0.5996	0.5880	0.6086	0.5867	0.5996
9	Mother	0.7121	0.6977	0.6958	0.7003	0.6602	0.6801
10	Soccer	0.5448	0.5451	0.5453	0.5455	0.5453	0.5456
11	Stefan	0.8635	0.8659	0.8738	0.8556	0.8628	0.8780
12	Tampete	0.7232	0.7255	0.7254	0.7094	0.7133	0.7198

Table 7: Complete Weight assigning in automatic weight assigning configuration. Example of IST configuration.

S. No	Achanta	Context	GBVS	Phase Fourier	Random Surround	Wavelets
Module 1	0.8	0.5	0.5	0.5	0.5	0.5
Module 2	0.5	0.8	0.5	0.5	0.5	0.5
Module 3	0.5	0.5	0.8	0.5	0.5	0.5
Module 4	0.5	0.5	0.5	0.8	0.5	0.5
Module 5	0.5	0.5	0.5	0.5	0.8	0.5
Module 6	0.5	0.5	0.5	0.5	0.5	0.8

REFERENCES

Ali Borji, Member, IEEE, Dicky N. Sihite, and Laurent Itti, Member, IEEE, 2013(a). *Quantitative Analysis of Human-Model Agreement in Visual Saliency Modeling: A Comparative Study*, *In IEEE Trans. Image Processing, VOL. 22, NO. 1.2013.*

Ali Borji, Hamed R. Tavakoli, Dicky N. Sihite, Laurent Itti, 2013(b). *Analysis of scores, datasets, and models in visual saliency prediction, In ICCV 2013*.

Ali Borji, Member, IEEE, and Laurent Itti, Member, IEEE, 2010. *State-of-the-art in Visual Attention Modeling, In IEEE Trans. Pattern Anal. Mach. Intell, 2010.*

C. Guo, Qi Ma and Liming Zhang, 2008.*Spatio-temporal Saliency detection using phase spectrum of quaternion Fourier transforms. In CVPR, 2008.*

C. Koch and S. Ullman, 1985. *Shifts in selective visual attention: towards the underlying neural circuitry. Human Neurobiology, 4:219–227, 1985.*

C. M. Privitera and Lawrence W. Stark, 2000. *Algorithms for defining visual regions-of-interest: Comparison with eye fixations. IEEE Trans. Pattern Anal. Mach. Intell., 22:970–982, Sep' 2000.*

D. Parkhurst, Klinton Law, and Ernst Niebur, 2002. *Modeling the role of salience in the allocation of overt visual attention. Vision Research, 42(1):107 – 123, 2002.*

H. Hadizadeh, M. J. Enriquez, and I. V. Bajić, 2012.*Eye-tracking database for a set of standard video sequences," IEEE Trans. Image Processing, vol. 21, no. 2, pp. 898-903, Feb. 2012.*

J. Harel, C. Koch, and P. Perona 2007. *Graph based visual saliency, In Advances in Neural Information Processing Systems*. MIT Press.

J. M. Henderson, J. R. Brockmole, M. S. Castelhano, and M. Mack, 2007. *Visual saliency does not account for eye movements during visual search in real-world scenes. Eye Movement Research: Insights into Mind and Brain, 2007*.

L. Elazary and L. Itti, 2008. *Interesting objects are visually salient. J. Vis., 8(3):1–15, 3 2008.*

L. Itti, 2005. *Quantifying the contribution of low-level saliency to human eye movements in dynamic scenes. Visual Cognition, 12:1093–1123, 2005*.

Nevrez Imamoglu, Weisi Lin, and Yuming Fang, 2013. *A Saliency Detection Model Using Low-Level Features Based on Wavelet Transform. IEEE Trans. Multimedia 15(1): 96-105 (January 2013).*

N. Bruce and J. Tsotsos, 2006. *Saliency based on information maximization. In Y. Weiss, B. Sch¨olkopf, and J. Platt, editors, Advances in Neural Information*

Processing Systems 18, pages 155–162. MIT Press, Cambridge, MA, 2006.

R. Achanta, F. Estrada, P. Wils and S. Süsstrunk, 2008. *Salient Region Detection and Segmentation, In ICVS, 2008, Vol. 5008*, Springer Lecture Notes in Computer Science, pp. 66-75, 2008.

Stas Goferman, Lihi Zelnik-Manor, and Ayellet Tal, 2010.*Context aware saliency detection. In CVPR'10, pages 2376–2383, 2010.*

Tilke Judd, FrØdo Durand, and Antonio Torralba, 2012. *A Benchmark of Computational Models of Saliency to Predict Human Fixations*. MIT Computer Science and Artificial Intelligence Laboratory Technical Report, January 13, 2012.

Vikram T. N., Tscherepanow M. and Wrede B., 2012. *A saliency map based on sampling an image into random rectangular regions of interest, In Pattern Recognition (2012).*

Yubing Tong, Faouzi Alaya Cheikh, Fahad Fazal Elahi Guraya, Hubert Konik and Alain Tremeau, 2011. *A Spatiotemporal Saliency Model for Video Surveillance. 3(1):241-263, Journal of Cognitive Computing*. Springer.

Yubing Tong, Faouzi Alaya Cheikh, Hubert Konik and Alain Tremeau, 2010. *Full reference image quality assessment based on saliency map analysis (Journal of Imaging Science and Technology, 54(3):030503-030514, 2010.*

Young, Richard (1987). *The Gaussian derivative model for spatial vision: I. Retinal mechanisms. Spatial Vision 2 (4): 273–293(21).*