To appear in the 2014 IEEE International Conference on Intelligent Transportation Systems (ITSC2014)
The published version of this paper is available in IEEE Xplore

Top-down based saliency model in traffic driving environment*

Tao Deng, Andong Chen, Min Gao, Hongmei Yan

***Abstract*—Traffic driving environment is a complex and dynamically changing scene. During driving, drivers always focus their attention on the most important and saliency areas or targets. Traffic saliency detection is an important application area of computer vision, which could be used to support autonomous driving, traffic sign detection, driving training, and car collision warning, etc. At present, most saliency approaches are based on bottom-up computation which does not consider the top-down control and cannot match the actual traffic saliency in drivers' eyes. In this paper, by carefully analyzing the eye tracking data of 40 subjects who were non-drivers and drivers when viewing 100 traffic images, we found that the drivers' attention was mostly concentrated on the front of road. We proposed that the vanishing point of road can be regarded as top-down guidance in the traffic saliency model. Subsequently, we gave the framework of a bottom-up and top-down combined traffic saliency model, and the results showed that our method can effectively simulate the attentive areas in driving environment.**

I. Introduction

Traffic environment is a complex and tridimensional scene of multiple information sources, which changes dynamically and requires being processed instantly, especially in the urban road. While driving a car, a person navigates to a desired destination (e.g. grocery store) while paying attention to different types of objects in the environment (roads, cars, people, street, traffic signs, etc.) and obeying traffic laws (speed limit, stop signs, etc.). Humans manage these competing tasks by selectively fixating their eyes to the most important or salient areas targets instantaneously according to driving demanding. Namely, human brain filters the irrelevant visual information and computes the momentary traffic saliency of environment very quickly via the deployment of a foveated visual system, although we are still unclear how this is done so effortlessly, yet so reliably.

Starting from the Feature Integration Theory of Treismanand Gelade [1] and the bottom-up attention model by Koch and Ullman[2], a series of ever refined algorithms has been designed to predict where subjects will fixate in synthetic or natural scenes [3-8].

Generally speaking, there are two different factors that influence visual saliency. One is the bottom-up, task-independent factor, which is driven by the low-level attributes of input image, such as color, intensity, and orientation, etc. The other is the top-down, task-dependent factor, which is driven by the goals and experiences and so on. Most now available models of saliency [5, 9-14]are biologically inspired and based on a bottom-up computational model. For example, Itti et al.[5, 9] proposed an amazing method, and multiple low-level visual features such as intensity, color, orientation, texture and motion were extracted from the image at multiple scales. They computed each features saliency map, then normalized and combined in a linear or non-linear fashion into a master saliency map that represented the saliency of each pixel. Finally, the winner-take-all and inhibition of return operations were adopted to identify every significant area. Harel et al.[10]computed saliency map using global information. They proposed a graph-based solution to obtain a saliency map, which is dependent on global information. Hou and Zhang[11, 12]proposed a simple and fast algorithm, called the spectrum residual (SR), which was based on the Fourier Transform. They proposed that the spectrum residual corresponds to image saliency. Bruce and Neil also proposed a bottom-up model called AIM[13], and Zhang Lingyun proposed the SUN model[14]. All the above bottom-up saliency models should be able to simulate humans' visual saliency to some extent. However, they are lack of top-down control. Therefore, the existing models have limited uses in specific or task-relevant conditions, such as traffic environment.

Traffic saliency detection is an important application area of computer vision, which computes the salient and prior area targets in driving environment, and could be used to support autonomous driving, traffic sign detection, driving training, and car collision warning, etc. However, as described previously, driving is a dynamic, task-oriented behavior. It is possible that the real traffic saliency might be completely different from the salient maps computed by traditional saliency algorithms or models. There is lack of experimental research and saliency model in this specific area currently. Nowadays, nearly all of the saliency models about traffic are about the traffic sign detection, few models estimated driver's real attention and gazes during driving.

Fig.1 showed that the classic bottom-up based saliency maps (GBVS and Itti saliency maps) don't match the actual humans' attentive or most fixational areas (eye tracking results in free-viewing and simulating driving tasks).Although

*This work was supported by the 973 project (2013CB329401), the Natural Science Foundations of China (91120013，61375115, 31300912), the fundamental research funds for the Central Universities (ZYGX2013J098). The work was also partially supported by the 111 Project (#B12027).

Tao Deng Andong Chen and Min Gao are with Biomedical Engineering, University of Electronic Science and Technology of China, Chengdu, China. (e-mail: tinydao@163.com)

Hongmei Yan is professor of Biomedical Engineering, University of Electronic Science and Technology of China, Chengdu, China. (e-mail: hmyan@uestc.edu.cn)

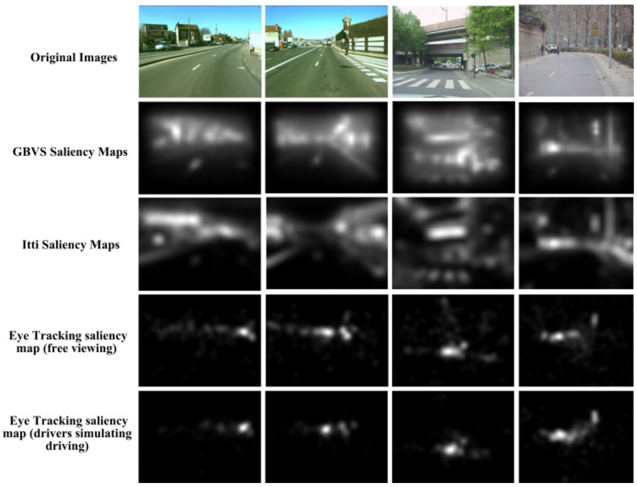
Original Images
GBVS Saliency Maps
Itti Saliency Maps
Eye Tracking saliency map (free viewing)
Eye Tracking saliency map (drivers simulating driving)

Figure 1.Pre-existing saliency model cannot accurately estimate human fixations in traffic environment. The first row shows the original images. The second and third row shows the saliency maps computed with GBVS and Itti model. The forth row shows the non-experience subjects' free-viewing eye tracking data. And the fifth row shows the drivers' simulating driving eye tracking data in our experiment.

we only showed GBVS and Itti based saliency maps here, the AIM, SR and SUN models were also considered, and all these models cannot accurately estimate the actual humans' attention areas in driving condition either.

From above, a top-down guide should be added in the traffic saliency computing models. Previous behavior studies about driving analyzed drivers' daytime eye movements and found that drivers looked straight ahead at the road 59 percent of the time, to the right side of the road 15 percent of the time, and to the left side of the road 25 percent of the time[15].Our results of eye tracking experiments also showed that the viewers' gazes mostly concentrated on the vanishing point of the road. This is a very important top-down control when driving. Therefore, the focus of this paper is to make use of the top-down mechanism to build up a top-down traffic saliency model. A vanishing point detection algorithm proposed by Hui Kong et al was mainly adopted as a top-down guidance in this paper[16, 17]. The results showed that the top-down traffic saliency model has an amazing improvement on traffic saliency detection compared to the classic models.

In this paper, a top-down saliency model about the road traffic environment was proposed, and we built a model framework in traffic saliency detection. The contribution of this paper is two folds: 1) a human traffic database with eye movement data is built. 2) Based on the combination of bottom-up and top-down selective attention mechanism, a top-down traffic saliency model is proposed.

II. Method

In this section, we state the human traffic database with eye movement tracking in the behavior experiment.

A. *Data gathering protocol*

We collected 100 traffic driving images, and all images were about the urban road (Fig.1). Eye movement were recorded from 40 subjects (18male and 22 female) aged 21-45 (average age28).Two groups were enrolled separately, and each comprised of 20 subjects. One group has no driving experience (Group I) and the others are drivers who have at least two years of driving experience (Group II). The group I viewed these images freely and group II viewed these images assuming they were driving a car. Each image was presented at full resolution for 10seconds separated by 20 seconds of resting with a gray screen. Eye movements were recorded with an infrared eye tracker (Eyelink2000, SR Research Ltd.) and sampled at 1000 Hz. Head movements were restricted by a forehead and chin rest. The pupil of the left eye was tracked at a sample rate of 1000 Hz and a spatial resolution of 0.1º.

In order to obtain a continuous saliency map of an image, we convolve a Gaussian filter across the user's fixation locations. The average saliency maps across all viewers were exampled as Fig.2, where the red and yellow area overlapped on the image indicates the area is more observably fixated, the green area indicates comparatively observed, and the rest mean fewer fixations on these parts.

The results of the experiment show that most of the attention of subjects focused on the front of the road in traffic environment (Fig 2), although there are some differences between the eye movements' data of the two groups. For instance, the attentional area of non-experience viewers is sparser, but that of the driving-experience group is more intensive. We proposed the fact that attentions always focus on the front of road is a top-down control or guidance in traffic environment. In Hui Kong's work, they proposed that the front of road exists one point what is called Vanishing Point (VP). So, the Vanishing-Point should be very important information in traffic driving environment.

B. *Top-down attentional area and vanishing point of road*

According to this experiment, we can conclude the following result: in the traffic environment, the attention of the participants is mainly concentrated on the front of the road. The result is consistent with Higgins's research. That is because traffic environment is quite different from other natural scenes. The latter usually has a significant target to attract human's attention. The traffic driving environment can be regarded as a special selected attention task condition. So we can consider this attention based on task-driven as top-down attention.

For vanishing point of road, researchers have already studied the detection algorithms. For example, in Hui Kong's study[16], they came up with a good algorithm to estimate the vanishing point of road. The main idea of the algorithm is that: through the analysis of the texture towards of the road toward, they found that all the texture is in one direction that they orient one point what is called vanishing point.

In this paper, we used Kong's algorithm as reference. First, we calculated all the vanishing points of our images. Then we analyzed their distribution. The Fig 3.a shows all

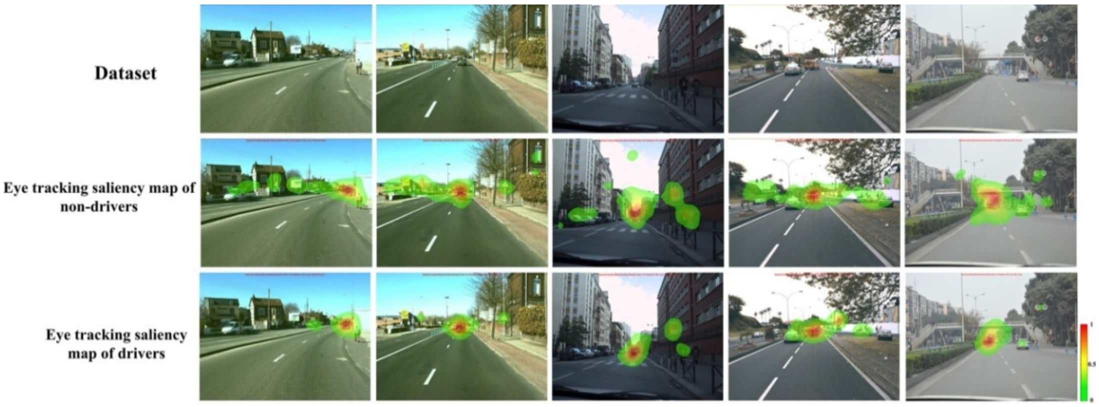
Dataset
Eye tracking saliency map of non-drivers
Eye tracking saliency map of drivers

Figure 2.Saliency map driven by different selective attention in the experiment. The first rows show the original images. The second rows show the free-viewing saliency maps of subject without driving experience. The third rows show the saliency maps of drivers with 2 years driving experience.

the vanishing points of our dataset, and Fig 3.b shows all vanishing points with a 2-D Gaussian at their locations.

Nevertheless, not all vanishing points can be computed successfully in our dataset. There are 4(total 100) vanishing points estimate abortively. We can see that most of the vanishing points centralize on the front of the traffic road. However, some points are out of this area when curve road exists.

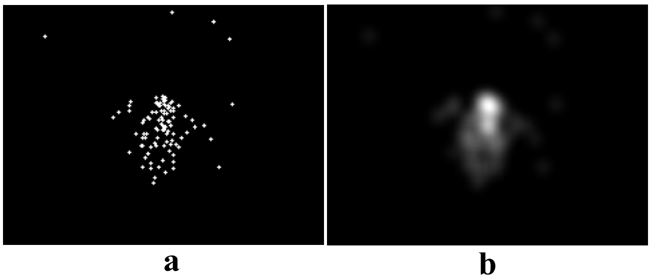

Figure 3.The distribution of all vanishing points. A shows all vanishing points of our dataset, and b shows the vanishing points with a 2-D Gaussian.

We compared the vanishing points with the eye tracking saliency maps in the dataset, and found that vanishing points overlapped with the positions drivers fixated most frequently in most cases, shown in Fig 4. Therefore, we proposed that the vanishing point of road can be regarded as top-down guidance in the traffic saliency model.

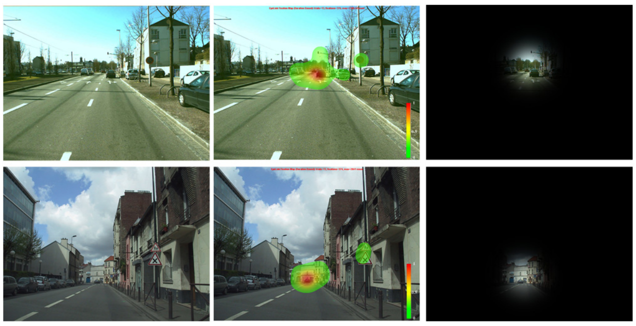

Figure 4. The first columns show the original images. The second columns show the eye tracking saliency maps of subjects with 2 years driving experience. The last columns show the vanishing points placed a 2-D Gaussian with $\sigma = 60$.

III. MODEL

Based on the vanishing point information, we propose a computing framework of a top-down based traffic saliency model (Fig.5), which is composed of classical bottom-up saliency model and top-down constraint. The main methodology of the model is to find the top-down constraint based on the eye movement experiment and then combine it with classical bottom-up model in a linear fashion. Finally, the model creates the saliency map based on formula5.

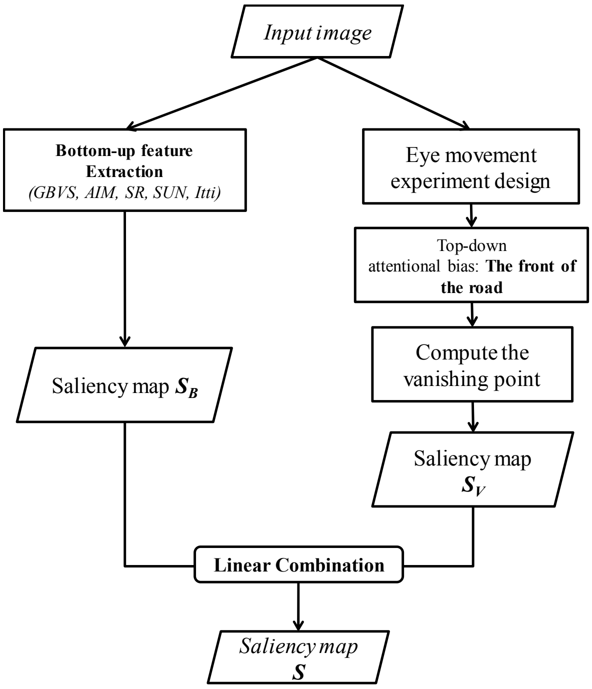
Input image
Bottom-up feature Extraction
(GBVS, AIM, SR, SUN, Itti)
Eye movement experiment design
Top-down attentional bias: The front of the road
Compute the vanishing point
Saliency map S_B
Saliency map S_V
Linear Combination
Saliency map S

Figure 5.The framework of our method.

A. *Vanishing point detection*

In the texture-based vanishing-point detection methods, the vanishing-point was usually estimated by analyzing texture orientation at each image pixel. Recently, Hui Kong proposes a new generalized Laplacian of Gaussian (gLoG) filter[18] to estimate the texture orientation. In Hui Kong's early work, he estimated the texture orientation with the Gabor-based method. The gLoG filter can estimate texture

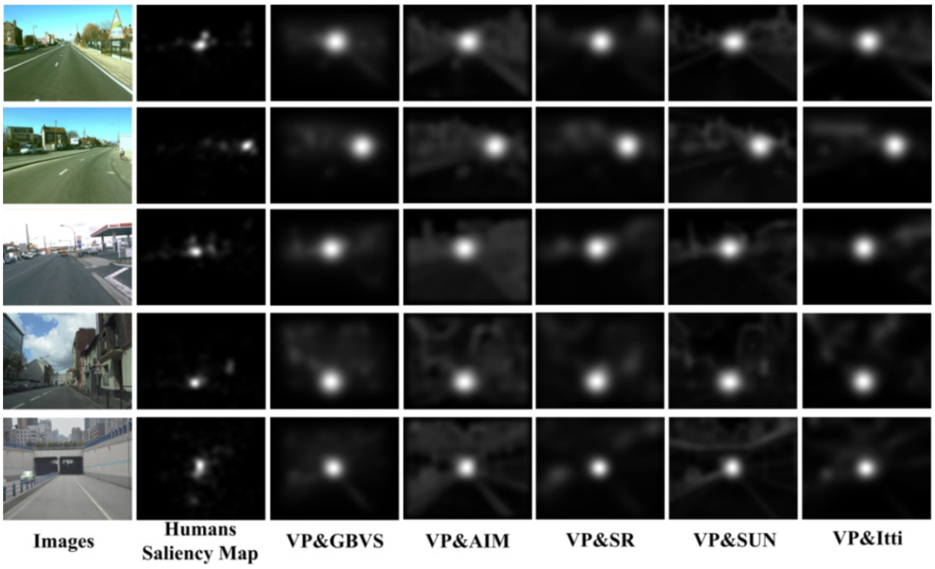

Figure 6.Comparisons of the vanishing point based top-down saliency maps.

orientation more accurately than the Gabor-based approach [16, 17], but more costly. The gLoG filter is applied to estimate the texture orientation at each pixel of an image. Then the vanishing point is detected based on the estimated texture orientations. The 2-D Gaussian function is defined as

$$G(x,y;\sigma)=\frac{1}{2\pi\sigma^2}\exp(-\frac{x^2+y^2}{2\sigma^2}) \quad (1)$$

where x and y represent spatial coordinates, and σ is the standard derivation of the Gaussian function.

The Gaussian scale-space representation $L(x,y;\sigma)$ of the image $f(x,y)$ is

$$L(x,y;\sigma)=f(x,y)*G(x,y;\sigma) \quad (2)$$

where $*$ is the convolution operator. Then, the Laplacian operator ∇^2 can be defined as

$$\nabla^2=\frac{\partial^2 L}{\partial x^2}+\frac{\partial^2 L}{\partial y^2}. \quad (3)$$

The ∇^2 can be applied to the Gaussian scale-space representation of an image to get its LoG (4) scale-space representation.

$$\nabla^2 G(x,y,\sigma)=\frac{x^2+y^2-2\sigma^2}{\pi\sigma^4}\exp(-\frac{x^2+y^2}{2\sigma^2}) \quad (4)$$

According to the above texture computation theory, the vanishing point can be voted based on the texture orientation at each image pixel using the locally adaptive soft-voting principle. The experiment result shows that Hui Kong's vanishing point detection methods can also estimate the vanishing point in our urban road image dataset.

B. Bottom-up and Top-down combined saliency model

As described in introduction, the classical bottom-up saliency models included GBVS, AIM, SR, SUN and Itti models. All of these saliency models have been applied to the natural scenes or object detection, but did not applied to the special scene such as traffic driving environment. In the natural scenes, usually there are one or more targets which have salient features to attract people's attention. Hence, the classical bottom-up saliency models could find out the targets by simulating the humans' vision system.

In 2001, Itti and Koch first proposed the idea of top-down influence to better estimate the saliency in specific tasks[19]. They addressed that there is a link between visual attention and eye movement. So, it's necessary to combine the eye movement with computation model to research humans' visual system. In recent years, some top-down saliency models are proposed by learning method. Judd et al.[20]considered the top-down information in their work by designing the eye tracking experiment to collect eye tracking data of their dataset and built a saliency learning model. This model produces a saliency map by analyzing the low-, mid- and high-level features of the input image, then combining them after training the features for every pixel of the image. Qi Zhao et al.[21]also proposed a saliency model based on learning method. They computed the weights of the features such as color, intensity, orientation and face by statistically analyzing the eye movement data. In their work, the weight of face is the most maximum than other features. The performances of the above two top-down models are greatly improved.

However, the above bottoms-up and top-down saliency models are not suitable for the traffic road environment. Therefore, we proposed a vanishing-point based bottom-up and top-down combined traffic saliency model. In order to make use of the vanishing point information in traffic environment, we combine it with classical bottom-up saliency model such as GBVS [10], SR[11], AIM [13], SUN[14]and Itti[5, 9].

We execute the vanishing point in a convolution with the Gaussian filter, and then combine the result with classical bottom-up saliency map in a linear additive. In the end, we get the final saliency map (Fig 6). The new saliency map $S(x,y)$ is defined as

$$S(x,y) = wS_V(x,y) + (1-w)S_B(x,y) \quad (5)$$

where w is the selected weight and it is defined as $w = 0.8$, $S_V(x,y)$ represents the saliency map of the vanishing point convolved Gaussian filter, $S_B(x,y)$ represents the saliency map of classical bottom-up saliency model. The reason why we define $w = 0.8$ is that we find that the subjects' most attention is focused on the vanishing point of traffic road and few attention is focused on the else scene. Practice has proved that the performance of this algorithm is optimal when $w = 0.8$.

Therefore, we realized the integration of the two visual attention mechanisms: the bottom-up, feature-based attention and top-down, task-dependent attention. The experimental results show that the algorithm performance has greatly improved after the bottom-up model joined with VP which represented the top-down guidance.

IV. Result and discussions

After integrating the current classic saliency models (GBVS, SR, SUN, AIM and Itti) with the aforementioned top-down information, we got the saliency images and then made quantitative analysis about the saliency data of human eyes obtained in our experiment.

We use an ROC[22] metric to evaluate the performance of human saliency maps to predict eye fixations. Using this method, the saliency map from the fixation locations of one user is treated as a binary classifier on every pixel in the image. Here, we rewrite the ROC evaluation algorithm that we only consider the true positive rate because the saliency model's true detection rate has greatly improved after adding the VP information. In Judd's paper[20], saliency maps are thresholded such that a given percent of the image pixels are classified as fixated and the rest are classified as not fixated. Then this saliency map is threshold at n =1, 3, 5, 10, 15, 20, 25, and 30 percent of the image for binary saliency maps which are typically relevant for applications.

We make the following observations from the ROC curves (Fig.7): (1) the model with VP information reaches 0.8 of the way to human performance. For example, when images are thresholded at 15% salient, our model performs at 0.65 while humans are at 0.8. (2) The models with VP feature perform much better than themselves. For example, at the 10% salient location threshold, the Itti model with VP feature performs at 0.48 while Itti performs at 0.2 for a 28% jump in performance. (3) The true positive rate of the saliency model with VP information has improved at 5% to 20% so that it means our model can simulate humans' attention areas quickly and precisely.

In the meantime, we also analyzed the change of the AUC value after adding top-down information to each algorithm. It can be seen from table.1 that after adding top-down information to the saliency algorithm of bottom-up, the AUC value which matches with eye-tracking data is greatly improved, which means that the new algorithm was much better and could better simulate human eye's attention mechanism. We can conclude that the vanishing point is the effective top-down guidance in traffic saliency models.

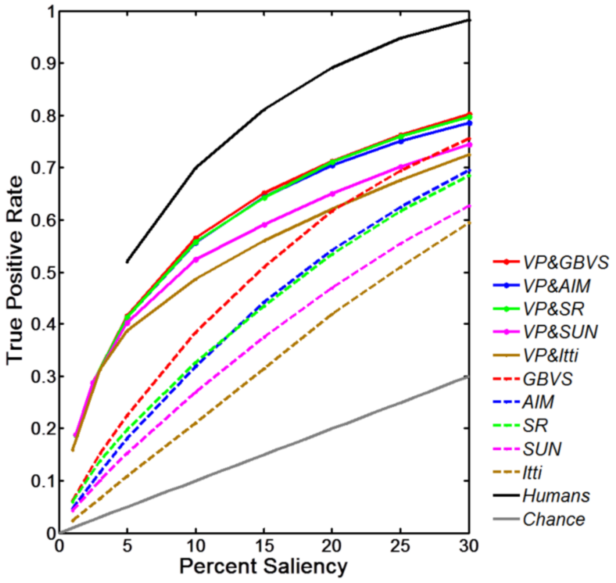

Figure 7.The ROC curve of these algorithms. The solid lines show the ROC of each algorithm who has combined with the VP (vanishing point) information, the dashed lines show the ROC of the classical saliency algorithms in our dataset.

Table 1.
The compare of AUC value of each saliency model.

AUC value	*GBVS*	*AIM*	*SR*	*SUN*	*Itti*
Without VP	0.7874	0.7581	0.7558	0.7210	0.7041
With VP	0.8286	0.8188	0.8181	0.7965	0.7809
↑	**0.0412**	**0.0607**	**0.0623**	**0.0755**	**0.0768**

Recently, the center bias method has been proposed by several researchers[20, 21]. Although center bias model is very similar to our model, but it can't be work when the road is curved or the front of road is out of the center of image. Fig.8 illustrated the comparison of the center bias model and our proposed model in the curved road environment. We can see that the center bias model is not suitable for the curved road and other situations, such as the road is not in the center

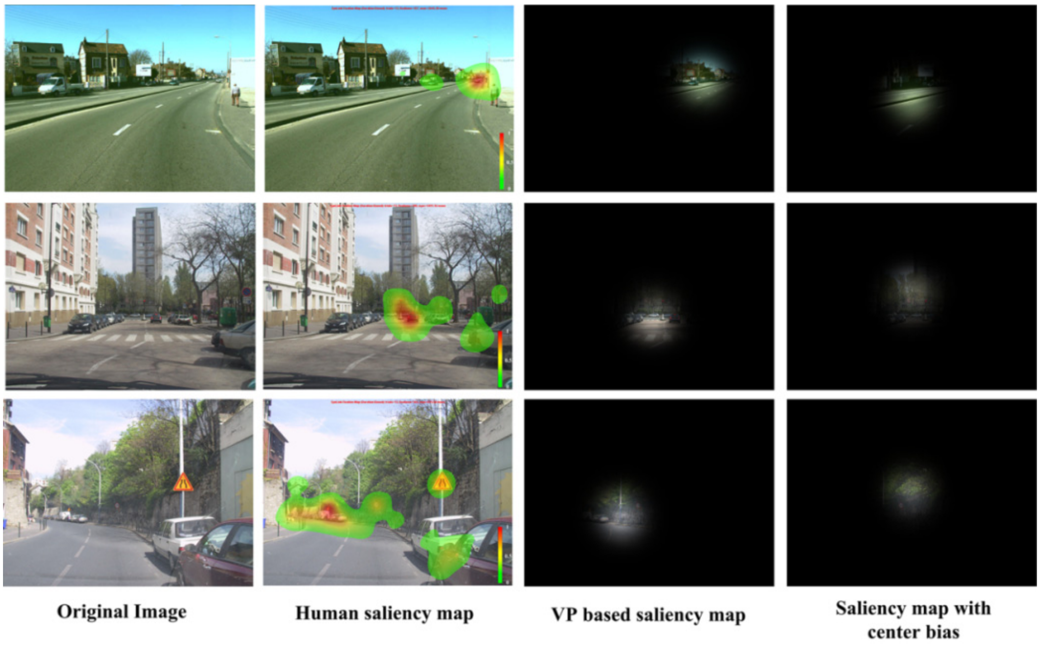

Figure 8.Comparison of the center bias model and our proposed model in the curved road environment.

of image. The saliency map of bias model cannot match the humans' attentive area in these situations. However, the vanishing points based model can always effectively estimate the area of human's attention.

In conclusion, in this paper, based on previous studies of eye movement in driving and our behavior experimental results, we found that the drivers' attention mostly focused on the front of road. Then we proposed that the vanishing point of road can be regarded as the top-down guidance in the traffic saliency model. Subsequently, we gave the framework of a bottom-up and top-down combined traffic saliency model and the results showed that our method can effectively simulate the attentive areas in traffic environment than classic models.

REFERENCES

[1] A. M. Treisman and G. Gelade, "A feature-integration theory of attention," *Cognitive psychology,* vol. 12, pp. 97-136, 1980.

[2] C. Koch and S. Ullman, "Shifts in selective visual attention: towards the underlying neural circuitry," in *Matters of Intelligence*, ed: Springer, 1987, pp. 115-141.

[3] W. Einhäuser, M. Spain, and P. Perona, "Objects predict fixations better than early saliency," *Journal of Vision,* vol. 8, p. 18, 2008.

[4] T. Foulsham and G. Underwood, "What can saliency models predict about eye movements? Spatial and sequential aspects of fixations during encoding and recognition," *Journal of Vision,* vol. 8, p. 6, 2008.

[5] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on pattern analysis and machine intelligence,* vol. 20, pp. 1254-1259, 1998.

[6] A. Oliva, A. Torralba, M. S. Castelhano, and J. M. Henderson, "Top-down control of visual attention in object detection," in *Image Processing, 2003. ICIP 2003. Proceedings. 2003 International Conference on*, 2003, pp. I-253-6 vol. 1.

[7] D. Parkhurst, K. Law, and E. Niebur, "Modeling the role of salience in the allocation of overt visual attention," *Vision research,* vol. 42, pp. 107-123, 2002.

[8] D. Walther, T. Serre, T. Poggio, and C. Koch, "Modeling feature sharing between object detection and top-down attention," *Journal of Vision,* vol. 5, p. 1041a, 2005.

[9] L. Itti and C. Koch, "A saliency-based search mechanism for overt and covert shifts of visual attention," *Vision research,* vol. 40, pp. 1489-1506, 2000.

[10] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," *Advances in neural information processing systems,* vol. 19, p. 545, 2007.

[11] X. Hou and L. Zhang, "Saliency detection: A spectral residual approach," in *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, 2007, pp. 1-8.

[12] X. Hou and L. Zhang, "Dynamic visual attention: searching for coding length increments," in *NIPS*, 2008, p. 7.

[13] N. Bruce and J. Tsotsos, "Saliency based on information maximization," *Advances in neural information processing systems,* vol. 18, p. 155, 2006.

[14] L. Zhang, M. H. Tong, T. K. Marks, H. Shan, and G. W. Cottrell, "SUN: A Bayesian framework for saliency using natural statistics," *Journal of vision,* vol. 8, p. 32, 2008.

[15] M. Ko, L. Higgins, S. T. Chrysler, and D. Lord, "Effect of Driving Environment on Drivers' Eye Movements: Re-Analyzing Previously Collected Eye-Tracker Data," in *Transportation Research Board 89th Annual Meeting*, 2010.

[16] H. Kong, J.-Y. Audibert, and J. Ponce, "Vanishing point detection for road detection," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, 2009, pp. 96-103.

[17] H. Kong, J.-Y. Audibert, and J. Ponce, "General road detection from a single image," *Image Processing, IEEE Transactions on,* vol. 19, pp. 2211-2220, 2010.

[18] H. Kong, S. E. Sarma, and F. Tang, "Generalizing Laplacian of Gaussian filters for vanishing-point detection," *Intelligent Transportation Systems, IEEE Transactions on,* vol. 14, pp. 408-418, 2013.

[19] L. Itti and C. Koch, "Computational modelling of visual attention," *Nature reviews neuroscience,* vol. 2, pp. 194-203, 2001.

[20] T. Judd, K. Ehinger, F. Durand, and A. Torralba, "Learning to predict where humans look," in *Computer Vision, 2009 IEEE 12th international conference on*, 2009, pp. 2106-2113.

[21] Q. Zhao and C. Koch, "Learning a saliency map using fixated locations in natural scenes," *Journal of vision,* vol. 11, p. 9, 2011.

[22] D. M. Green and J. A. Swets, *Signal detection theory and psychophysics* vol. 1974: Wiley New York, 1966.