Saliency Detection and Tracking Of Stereoscopic Images and Videos

Jeena John
Electronics and Communication
Mar Baselios Institute of Technology and Science
Ernakulam, India
Jeenajohn.jj@gmail.com

Abstract— **Saliency detection is believed to be an important precursor for a vast number of multimedia processing applications. Besides 2D saliency detection methods, depth feature is also considered to detect saliency in stereoscopic images as well as for videos. This paper presents an enhanced saliency detection framework, depending on the feature contrast of luminance, color, texture and depth, which in turn are extracted from discrete cosine transform coefficients for the purpose of feature contrast calculation. Here, in order to consider the calculation of local and global contrast, a Gaussian model of spatial distance between image patches is adopted. Human vision system actively seeks salient regions and movements in video sequences to reduce the search effort. The motion saliency model detects the moving objects whose motion is salient to its background. In this paper, we propose a novel stereoscopic images and video saliency detection model for detecting the attended regions that correspond to both interesting objects and dominant motions in video sequences. And the salient region on the video is tracked.**

Keywords—stereoscopic images; video tracking; 3D images.

I. INTRODUCTION

Saliency detection means detecting visually attracted regions in images. It is an aspect of exploring visual attention from a computer vision viewpoint. The human pays unequal attention to what is seen in the world. This ability of withdraw from some things in order to deal effectively with others is called attention. Detecting these areas of attention is called saliency detection. The three-dimensional (3D) display devices such as 3D-TV has experienced a rapid advancement in the past few years, which brings the 3D technology closer to our daily life than ever before. It is a hot research topic in both academia and industry nowadays.In these years, image attention detection has been long studied, while not much work has been extended to video sequences where motion plays an important role. Neuro physiological experiments have proved that neurons in the middle temporal visual area compute local motion contrast. And such neurons underlie the perception of motion pop out and figure-ground segmentation which influences the attention allocation. After realizing the importance of motion information in video attention, the motion feature has been added into the saliency models.

In this project, we propose a saliency detection framework for stereoscopic images and a video saliency detection model for detecting the attended regions that correspond to both interesting objects and dominant motions in video sequences and tracking that salient region based on the feature contrast. Four types of features, namely color, luminance, texture and depth are extracted from discrete cosine transform coefficients for feature contrast calculation. A Gaussian model of the spatial distance between image patches is adopted for consideration of local and global contrast calculation. Then, a new fusion method is designed to combine the feature maps to obtain the final saliency map for stereoscopic images. In addition, we adopt the center bias factor and human visual acuity, the important characteristics of the human visual system, to enhance the final saliency map for stereoscopic images. The obtained salient region of the video is successfully tracked in this project. This is done by using Regionprops algorithm.

The remaining of this paper is organized as follows. Section II, the proposed model is described in detail. In Section III, provides the experimental results on eye tracking databases. Section IV provides the final section concludes the paper.

II. PROPOSED MODEL

The main working that is detecting the saliency of 3D images and video is same. In case of video we first extract n frames separately from the video. Then taking the first frame we are applying the same procedure to detect the saliency of 3D images and its working is as follows. Firstly, the color, luminance, texture, and depth features are extracted from the input stereoscopic image. Based on these features, the feature contrast is calculated for the feature map calculation. A fusion method is designed to combine the feature maps into the saliency map. Additionally, we use the centre bias factor and a model of human visual acuity to enhance the saliency map based on the characteristics of the HVS. We will describe each step in detail in the following subsections.

(A) Feature Extraction

In this study, the input image is divided into small image patches and then the DCT coefficients are adopted to represent the energy for each image patch. In this paper, we use the patch size of 8 × 8 for the saliency calculation. The used image patch size is also the same as DCT block size in JPEG compressed images. The input RGB image is converted to YCbCr color space. In YCbCr color space, we use the DC coefficient of Y component to represent the luminance feature for the image patch while the DC coefficients of Cb and Cr components are adopted to represent the color features. We use AC coefficients from only Y component to represent the texture feature of the image patch. We use several first AC coefficients to represent the texture feature of image patches.

In a stereoscopic display system, depth information is usually represented by a disparity map which shows the parallax of each pixel between the left-view and the right-view images. In this study, the depth map M of perceived depth information is computed based on the disparity, where *V* represents the viewing distance of the observer; *d* denotes the interocular distance; *P* is the disparity between pixels; *W* and *H* represent the width (in cm) and horizontal resolution of the display screen, respectively. We set the parameters based on the experimental studies in. Similar with feature extraction for color and luminance, we adopt the DC coefficients of patches in depth map calculated as in Equation the study [1]

I.

(B) Feature Map Calculation

A direct method to extract salient regions in visual scenes is to calculate the feature contrast between image patches and their surrounding patches in visual scenes. In this study, we estimate the saliency value of each image patch based on the feature contrast between this image path and all the other patches in the image. Here, we use a Gaussian model of spatial distance between image patches to weight the feature contrast for saliency calculation. The saliency value F_i^k of image patch *i* from feature *k* can be calculated as: where *k* represents the feature; l_{ij} denotes the spatial distance between image patches *i* and *j* ; U_{ij}^k represents the feature difference between image patches *i* and *j* from feature *k*; σ is the parameter of the Gaussian model and it determines the degree of local and global contrast for the saliency estimation.

$$F_i^k = \sum_{\mathrm{j \neq i}} \frac{1}{\sigma\sqrt{2\pi}} e^{\mathrm{l}_{\mathrm{ij}}^2 \frac{1}{2\sigma^2}} \mathrm{U}_{\mathrm{ij}}^{\mathrm{k}}$$

For any image patch *i*, its saliency value is calculated based on the center-surround differences between this patch and all other patches in the image. The weighting for the center-surround differences is determined by the spatial distances (within the Gaussian model) between image patches. The differences from nearer image patches will contribute more to the saliency value of patch *I* than those from farther image patches. Thus, we consider both local and global contrast from different features in the proposed saliency detection model. The feature difference U_{ij}^l between image patches *i* and *j* is computed differently from features *k* due to the different feature representation method.Since the colour color, luminance and depth features are represented by one DC coefficient for each image patch, the feature contrast from these features between two image patches *i* and *j* can be calculated as the difference between two DC coefficients of two corresponding image patches as follows in study [1]

(C) Saliency Estimation From Feature Map Fusion

After calculating feature maps indicated in Equation, we fuse the feature maps to compute the final saliency map. The different visual dimensions in natural scenes are competing with each other during the combination of final saliency map. All the visual features interact and contribute simultaneously to the saliency of visual scenes. To avoid the drawbacks from ad-hoc weighting of linear combination of feature maps, we propose an adaptive weighting for the fusion of feature maps in this study.

Generally, the salient regions in a good saliency map should be small and compact, since the HVS always focus on some specific interesting regions in images. During the fusion, we assign more weighting for those feature maps with small and compact salient regions and less weighting for others with more spread salient regions. The spatial variance$\upsilon_k$of feature map $F_k$can be computed as follows as in study [1]

We use the normalized υk values to represent the compactness property for feature maps. With larger spatial variance values, the feature map is supposed to be less compact. We calculate the compactness *βk* of the feature map *Fk* as follows.

$$\beta_k = {}^{1}/_{e^{v_k}}$$

Based on compactness property of feature maps calculated in equation, we fuse the feature maps for the saliency map as follows.

$$S_f = \sum_{k} \beta_k \,.\, F_k + \sum_{p \neq q} \beta_{p\,.}\ \ \beta_q \,.\, F_p\, F_q$$

Different from existing studies using the constant weighting values for different images, the proposed fusion method assign different weighting values for different images based on their compactness properties provides an image sample for the feature map fusion. Experimental results in the next section show that the proposed fusion method can obtain promising performance.

(D) Saliency Enhancement

The already existing Eye tracking experiments have shown that the bias towards the screen center exists during human fixation, which is called centre bias. In this paper, we have used the centre bias factor to enhance the saliency map from the proposed 3D saliency detection model. We use a Gaussian function with kernel width as one degree (foveal size) to model the centre bias factor. A CBM *Sc* can be obtained by the Gaussian function. The experimental results in the study shows the centre bias is irrespective to the distribution of image features, which means that the centre bias is independent on the saliency map *S f* calculated from image features. The saliency map by considering the center bias factor can be calculated as follows.

$$S = \gamma_1\ S_f\ {}_{+}\gamma_2\ S_c$$

It is well accepted that the HVS is highly space-variant due to the different densities of cone photoreceptor cells in the retina. On the retina, the fovea owns the highest density of cone photoreceptor cells. Thus, the focused region has to be projected on the foveal to be perceived at the highest resolution.

The density of the cone photoreceptor cells becomes lower with larger retinal eccentricity. The visual acuity decreases with the increased eccentricity from the fixation point. We use this property to enhance the saliency map of 3D images. In the saliency map, the pixels whose saliency value is larger than certain threshold are considered as salient regions. The human eyes would focus on these salient regions when observing the natural scenes and they are also most sensitive to these regions. In this study, we use a model of human visual sensitivity in to weight the saliency map. The retina eccentricity *e* between the salient pixel and non-salient pixel can be computed according to its relationship with spatial distance between image pixels. For any pixel position *(i, j)*, its eccentricity *e* can be calculated by the spatial distance between this pixel and the nearest salient pixel*(i*0*, j*0*)* as:

$$\boldsymbol{e = \ tan^{-1}\,(\frac{d^l}{v})}$$

The final saliency map *S* enhanced by the normalized visual sensitivity *Cs (f, e)* can be calculated as:

$$\boldsymbol{s^l = s * c_s(f, e)}$$

With the enhancement operation by the centre bias factor, the saliency values of center regions in images would increase, while with the enhancement operation by human visual acuity, the saliency values of non-salient regions in natural scenes would decrease and the saliency map would get visually better with the enhancement operation by the centre bias factor and human visual acuity, the saliency map can predict the saliency more accurately,

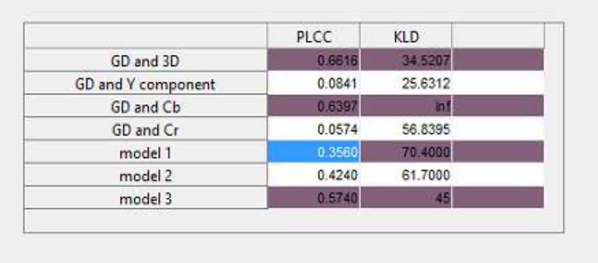
Now the saliency of an image is obtained or the saliency of the 1st frame is obtained. For tracking the salient region now we are applying the region props algorithm to this frame. Similarly we are obtaining the saliency of all other n frames of the video separately and applying the tracking algorithm to it. After completing the process in all the frames, we are combining all the frames together to obtain the tracked salient region of the video as well.

III. EXPERIMENTAL EVALUATION

In this section, we conduct the experiments to demonstrate the performance of the proposed 3D saliency detection model. We first present the evaluation methodology and quantitative evaluation metrics. Following this, the performance comparison between different feature maps is given in subsection. Then we provide the performance evaluation between the proposed method with other existing ones.

Table I

Experimental evaluation of comparison between different feature channels and comaprison between the proposed model and the other existing ones



	PLCC	KLD
GD and 3D	0.6616	34.5207
GD and Y component	0.0841	25.6312
GD and Cb	0.6397	Inf
GD and Cr	0.0574	56.8395
model 1	0.3560	70.4000
model 2	0.4240	61.7000
model 3	0.5740	45

(A) Evaluation Methodology

In the experiment, we adopt the eye tracking database to evaluate the performance of the proposed model. Currently, there are few available eye tracking database for 3D visual attention modelling in the research community. This database includes 18 stereoscopic images with various types such as outdoor scenes, indoor scenes, scenes including objects, scenes without any various object, etc. Some images in the database were collected from the Middlebury 2005/2006 dataset, while others were produced from videos recorded by using a Panasonic AG-3DA1 3D camera.

The performance of the proposed model is measured by comparing the ground-truth and the saliency map from the saliency detection model. As there are left and right images for any stereoscopic image pair, we use the saliency result of the left image to do the comparison, similar with the study. The PLCC (Pearson Linear Correlation Coefficient), KLD (Kullback-Leibler Divergence), and AUC (Area Under the Receiver Operating Characteristics Curve) are used to evaluate the quantitative performance of the proposed stereoscopic saliency detection model. Among these measures, PLCC and KLD are calculated directly from the comparison between the fixation density map and the predicted saliency map, while AUC is computed from the comparison between the actual gaze points and the predicted saliency Equations

(B) Compariosn Between Different Feature Channels

In this experiment, we compare the performance of different feature maps. The table provides the quantitative comparison results. In this table, $C1$ and $C2$ colour represent the colour feature from Cb and Cr components respectively. From this table, we can see that the performance of saliency estimation from $C1$ colour feature is similar with that from $C2$ colour feature, while the feature map from Luminance feature can obtain better performance than that of colour feature map from $C1$ or $C2$ component. Compared with colour and luminance features, the depth feature can estimate better saliency result. For the texture feature, it gets the lowest PLCC and AUC values among these used features. Its KLD value is also higher than those from other features. Thus, the saliency estimation from texture feature is poorest among the used features. Compared with feature maps from these low-level features of colour, luminance, texture and depth, the final saliency map calculated from the proposed fusion method can get much better performance for saliency estimation for 3D images, as shown by the PLCC, KLD and AUC values.

(C) Comparison Between The Proposed Model And The Other Existing Ones

In this experiment, we compare the proposed 3D saliency detection model with other existing ones. The quantitative comparison results are given in Table. From this table, we can see that the PLCC and AUC values from the proposed model is larger than those from other models, while KLD value from the proposed model is lower than those from other models. The statistical test results show the performance of the proposed model is significantly different from that from other existing ones. Thus, the proposed model can obtain a significantly higher performance than other existing models. The reason for this is that the 2D saliency detection model1 calculates saliency map mainly by local contrast. Similarly, there is the same drawback for the saliency maps. In the proposed model of Table, we use the average combination for the feature maps from color, luminance, and texture features to obtain the proposed 2D model and combine it with the proposed DSM (Depth Saliency Map) to obtain the experimental results. From this table, we can see that the3D saliency detection model with the depth information always obtains better performance than 2D saliency detection model, which demonstrates that the depth information is helpful in designing 3D saliency detection models.

CONCLUSION

In this study of images, we propose a stereoscopic saliency detection model for videos and 3D images. The features of color, luminance, texture and depth are extracted from DCT coefficients to represent the energy for small image patches. The saliency is estimated based on the energy contrast weighted by a Gaussian model of spatial distances between image patches for the consideration of both local and global contrast. A new fusion method is designed to combine the feature maps for the final saliency map. Additionally, we adopts the characteristics of the HVS (the Centre bias factor and human visual acuity) to enhance the saliency map. The obtained salient region of the video is successfully tracked in this project. This is done by using Region props algorithm. Experimental results show the promising performance of the proposed saliency detection model for stereoscopic images based on the recent eye tracking databases.

An approach to automatically track salient regions in generic videos is presented in this paper. Specifically, given a starting video frame, we first detect its salient regions based on extracted color and orientation maps; then we track these regions for the rest of frames within a short period of time. The extracted salient region information could be used to evaluate the frame importance so as to provide guidance in key frame extraction for video summarization purpose.

As a future scope, to reduce cognitive overload in CCTV monitoring, it is critical to have an automated way to focus the attention of operators on interesting events taking place in crowded public scenes. The attention of CCTV operators tends to deteriorate after prolonged surveillance video monitoring, especially when they are overwhelmed with unpredictable and complex motion patterns observed in a crowded scene. To increase surveillance effectiveness and reduce operator cognitive overload, it is critical to filter the displayed input streams and focus operator attention on interesting/salient events. So a saliency detection method that extracts unique regions from an unknown background, which is regarded as a pre-attentive process, is required to be developed.

REFERENCES

[1] Yuming Fang, Member, IEEE, Junle Wang, Manish Narwaria, Patrick Le Callet," Saliency Detection for Stereoscopic Images", IEEE Transactions On Image Processing, Vol. 23, No. 6, June 2014

[2] D. Zhong, and S. F. Chang, "An integrated approach for content-based video object segmentation and retrieval", IEEE Trans. on Circuits and Systems for Video Technology, 9(8), December 1999.

[3] D. Wang, "Unsupervised video segmentation based on watersheds and temporal tracking", IEEE Trnns. on Circuits and Systems for Video Technoloby, 8(5), 1998.

[4] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," IEEE Trans. Pattern Anal. Mach. Intell., vol. 20, no. 11, pp. 1254–1259, Nov. 1998.

[5] N. Bruce and J. Tsotsos, "An attentional framework for stereo vision," in Proc. 2nd IEEE Canadian Conf. Comput. Robot Vis., May 2005, pp. 88–95.

[6] I. van der Linde, "Multi-resolution image compression using image foveation and simulated depth of field for stereoscopic displays," Proc. SPIE, vol. 5291, Stereoscopic Displays and Virtual Reality Systems XI, 71, May 2004.

[7] MJ Black and DJ Fleet, "Probabilistic detection and tracking of motion discontinuities", IEEE Conf. ICCV, pp.551-558, 1999

[8] S. Zhang and F.W.M. Stentiford, "Motion detection using a model of visual attention," in Proc. of ICIP, San Antonio, USA, pp. 513-516, Sept. 16-19, 2007.