International Journal of Engineering Research and General Science Volume 4, Issue 3, May-June, 2016
ISSN 2091-2730

Salient Region Extraction for 3D-Stereoscopic Images

Leena Chacko, Priyadarsini S, Arya Jayakumar

Department of Computer Engineering, College of Engineering, Chengannur, leenachacko52492@gmail.com, 9048182219

Abstract— Detecting region of interest in images is an interesting topic. It has wide range of applications in Image Processing. With the emerging application of stereoscopic display, it became necessary to design saliency detection for stereoscopic images. The field of optical flow estimation is making steady progress now a days. In this paper, a new optical flow method is designed to estimate the relative motion of the images captured by left eye and right eye of the observer from the scene. DCT feature map promises energy compactness property. The method promisingly obtained efficient result over the existing methods.

Keywords— visual attention, optical flow, 3D saliency, motion vector, median filtering, DCT feature map, temporal saliency.

INTRODUCTION

In the Human Visual System (HVS), Visual attention is an important characteristic in visual information processing. Attention helps us to decide where to move our eyes next. It would selectively process the important part by filtering out others. Thus it helps to reduce the complexity of scene analysis. These visually important information is also termed as Salient regions or Region of Interest. Two mechanisms of visual attention are usually distinguished as: bottom-up and top-down. Bottom-up mechanism, which is image-driven and task-independent, is a perception process for automatic salient region selection for natural scenes while top-down mechanism is a knowledge driven and task-dependent cognitive processing. These two mechanisms interact with each other and affect the human visual behavior.

Visual attention models have many image processing applications, such as classification, visual retargeting, visual coding, watermarking, image segmentation, image retrieval etc. Many bottom up saliency detection models have been proposed for 2D images/videos. With the emergence of stereoscopic display, which is capable of conveying depth perception to the viewer, the requirement of designing computational saliency detection models for 3D multimedia applications increased. Different from that of 2D images, it is important to consider the depth factor, in Saliency detection for 3D images.

In this paper, a new technique is designed for 3D saliency detection. Inspired from the concepts of video which is a collection of frames, Optical flow algorithm is performed on the input stereoscopic image (comprises of left and right eye view images). The resultant image clearly shows the relation motion of the scene in two images. Thereafter, DCT feature map of the obtained result are taken into account to obtain a pre-final saliency map. The final saliency map is calculated based on the feature contrast.

RELATED WORK

Visual attention is a set of cognitive operations that allow us to efficiently selecting relevant information and by filtering out others. Attention is a highly flexible mechanism, because it can operate on regions of space, specific features of an object, or on entire objects. Corresponding to bottom up and top-down approaches in perception process, the attention mechanism may be stimulus-driven or goal-driven.

Many studies have been done in the field of visual attention. Itti *et al.* proposed the earliest computational saliency detection models based on the neuronal architecture of the primates' early visual system [19]. The saliency map is obtained by calculating multiscale center-surround differences using features like intensity, color, and orientation. Harel et al. extended Itti's model [18]. In that study, saliency from feature contrast is measured using the graph- based theory, a more accurate measure of dissimilarity. Hou et al. proposed a saliency detection method using the concept of Spectral Residual [17]. The saliency map is computed by log spectra representation of images from Fourier Transform. Now a days, some saliency detection models have been proposed by patch-based contrast and obtained promising performance [5]–[7]. A saliency detection model in compressed domain is designed by Fang et al. for the application of image retargeting [5]. A context-aware saliency detection model is proposed by Goferman based on feature contrast from color and intensity using image patches [7]. Achanta *et al.* [6] tried to obtain more frequency information to get a better saliency measure.

Besides 2D saliency detection models, in [2], Bruce at al. proposed a stereo attention framework by extending an existing attention structure to the binocular domain. However, there is no computational model proposed in that study [2]. The key of 3 Dimensional

saliency detection model is how to adopt the depth cues besides the traditional 2 Dimensional features such as color, intensity and orientation. Studies from neuroscience indicate that the depth feature as well as other low level features would cause human beings' attention focusing on the salient regions [3]. Therefore, an accurate 3 Dimensional saliency detection model should take depth contrast into account as well as contrast from other common low-level features. Based on multiple perceptual stimuli, Zhang et al. designed a stereoscopic visual attention algorithm for 3D video [4]. Chamaret et al. built a Region of Interest extraction method for adaptive 3D rendering [8]. Both studies [4] and [8] adopt depth map to weight the two dimensional saliency map to calculate the final saliency map for 3D images. Another method of 3 Dimensional saliency detection model is built by incorporating depth saliency map into the traditional 2D saliency detection methods.

In [9], Ouerhani et al. extended a 2 Dimensional saliency detection model to 3 Dimensional saliency detection by taking depth cues into account. Potapova introduced a 3 Dimensional saliency detection model for robotics tasks by incorporating the top-down cues with the bottom-up saliency detection [10]. Recently, Wang et al. proposed a computational model of visual attention for 3 Dimensional images by extending the traditional 2 Dimensional saliency detection methods. In [2], a public database with ground-truth of eye-tracking data is provided.

The concepts of some video saliency detection models were very helpful in designing the new method. In [13], a phase-based saliency detection model for video was proposed. The saliency map is obtained through inverse Fourier transform on a constant amplitude and the original phase spectrum of input video frames. Itti *et al.* [11] developed a model for detecting the surprising events in video, considering that surprising events attracts human beings' attention. Studies were also done in combining spatial saliency maps with temporal saliency map [12]. Based on the rarity of features, the authors of [14] designed a dynamic visual attention model. The main idea to the proposed technique is contributed by [13].

PROPOSED FRAMEWORK

Depth perception is highly correlated with visual attention, especially in the case of stereoscopic images. Therefore, saliency detection in stereoscopic images definitely considers the factor of depth. Slightly different from that, in this paper, a new saliency detection technique is designed by considering the relative motion of the scene captures by the left eye and right eye of the observer. That is, the visual perception information is the factor considering here. It can be termed as temporal saliency. This temporal saliency evaluation algorithm starts with Optical flow based motion estimation. The new method directly incorporates motion information by modeling the visual perception process in an information communication framework. The proposed framework is shown in figure 1.

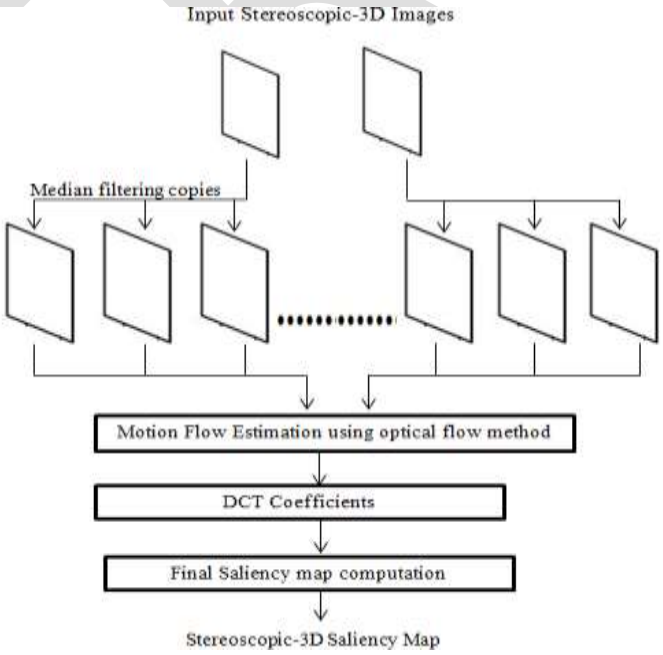

Figure 1: The framework of the proposed model

From the figure itself, it is very well understood that the proposed system is consisting of mainly 4 steps.

1. Create filtered version of Input Image

2. Motion vector calculation using optical flow
3. DCT feature map of image blocks
4. Saliency map calculation

1. Create filtered version of Input Image

In this step, first the input Stereoscopic Image is selected. In order to estimate the optical flow, it is needed to have multiple frames. Due to the lack of multiple frames, it is necessary to make the copies of the input Images. For that, filtered versions of the Input Stereoscopic Images are created using median filtering.

The median filter is used to reduce noise of image or simply the outliers. It is like the mean filter/average filter. It is better than the mean/average filter. Like mean filter, the median filter considers each pixel in the image to find whether or not it is representative of its surroundings. Instead of replacing pixel value with the mean of neighboring pixels, it replaces with median of those values. The median can be calculated by sorting all the pixel values from the neighborhood and then replacing the pixel with the middle pixel value. The median filter allows high spatial frequency detail to pass and it is very effective at removing noise on images where less than half of the pixels have been affected.

Advantages of using Median filter

- Preserves useful detail in the image.
- Single unrepresentative pixel in a neighborhood will not affect median value.
- No degradation to the underlying image.

2. Motion vector calculation using optical flow

Optical flow can be explained as the pattern of apparent motion of objects or surfaces. From a sequence of images, it is possible to find local image motion. The field of optical flow estimation is making steady progress now a days. The basics of today's methods are just the resemblance of Horn and Schunck (HS).
First of all, the left view mage and right view image are selected. After creating filtered copies of input Image, in order to compute the motion flow, the Classical optical flow method is used.

The classical optical flow algorithm is a direct descendant of the original **HS** formulation. Applying a median filter to intermediate flow values increases the accuracy of the recovered flow fields. It is found that the classical flow formulation is the best method in optical flow methods. Classical Optical Flow is a widely used differential method for optical flow estimation. Here, the assumption is that, the displacement of the image contents between two nearby frames is approximately constant and small within a neighborhood of the point p under consideration, and solves the basic optical flow equations for all pixels in that neighborhood. The classical optical flow objective function in its spatially discrete form as:

$$E(u, v)=\sum_{i,j}\left\{\rho D\left(I_1(i,j)-I_2\left(i+u_{i.j}, j+v_{i,j}\right)\right)+\lambda[\rho s(u_{i,j}-u_{i+1.j})+\rho s\left(u_{i,j}-u_{i.j+1}\right)+\rho s\left(v_{i,j}-v_{i+1.j}\right)+\rho s(v_{i,j}-v_{i.j+1})]\right\}$$

where u and v are the horizontal and vertical components of the optical flow field to be estimated from images I_1 and I_2, λ is a regularization parameter and ρD and ρS are the data and spatial penalty functions.

There are three different penalty functions: (1) the quadratic HS penalty $\rho(x) = x^2$; (2) the Charbonnier penalty (Classic C) $\rho(x) = \sqrt{x^2 + \epsilon^2}$ [13], a differentiable variant of the absolute value, the most robust convex function and (3) the Lorentzian (Classic L) $\rho(x) = \log\left(1+\frac{x^2}{2\sigma^2}\right)$, which is a non-convex robust penalty. Another recent classical flow method is classic NL, which is faster than the other two.

The main equation we are implementing to calculate the flow is

$$V = MV_P + (-1) * MV_f$$

where MV is the original motion vector. MV_f and MV_P are the past reference and the future reference frames respectively [15].

In this paper, the Rudin-Osher-Fatemi (ROF) structure texture decomposition method is followed to pre-process the input sequences and linearly combine the texture and structure components. The optical flow estimated at a coarse level is used to warp the second

image toward the first at the next finer level, and a flow increment is calculated between the first image and the warped second image. The standard deviation of the Gaussian anti-aliasing filter is set to be $\frac{1}{\sqrt{2d}}$, where d denotes the down-sampling factor. Each level is recursively down sampled from its nearest lower level.

At each warping step, data term is linearized which involves computing terms of the type $\frac{\partial}{\partial x} I_2\left(i + {u_{i,j}}^k, j + {v_{i,j}}^k\right)$, where $\frac{\partial}{\partial x}$ denotes the partial derivative in the horizontal direction, u^k and v^k denote the current flow estimate at iteration *k*. Then the derivative of the second image is taken and warped the second image and its derivatives towards the first using the current flow estimate by bi-cubic interpolation. Then the spatial derivatives of the first image is taken and average with the warped derivatives of the second image.

Graduated non-convexity (GNC) scheme is a technique that attempts to solve a difficult optimization problem by initially solving a greatly simplified problem, and progressively transforming that problem until it is equivalent to the difficult optimization problem. For the Charbonnier (Classic C) and Lorentzian (Classic L) methods, the Graduated non-convexity (GNC) scheme that linearly combines a quadratic objective with a robust objective in varying proportions, from fully quadratic to fully robust is used.

3. DCT feature map of image blocks

As the image is divided into patches, it is essential to find the relative motion of i^{th} patch. Salient regions in visual scenes have feature contrast from their surrounding regions. Thus, calculating the feature contrast between the image patches and their surrounding patches is the direct method to extract salient regions in visual scenes. We calculate the saliency value of each image patch based on the feature contrast between this image patch and all the other patches in the image. Here, we use a Gaussian model of spatial distance between image patches to weight the feature which determines the impact of neighboring patches based on their distances to the current patch.

$$v_i = \sum_{j \neq i} \frac{1}{\sigma\sqrt{2\pi}} e^{l_{ij}^2} / (2\sigma^2) D_{ij}^v$$

where D_{ij}^v is the length of the vector difference between the mean absolute motion vectors of patches i and j. l_{ij}^2 represents the spatial distance between image patches *i* and *j*. σ is the parameter of the Gaussian model.

4. Final Saliency Map Calculation

It would be a surprising event to HVS, if an object is of strong motion with respect to the background. HVS would pay more attention to such events. So, visual attention of motion can be calculated on the basis of the perceptual prior probability distribution about the speed of motion. Based on the results of [16], the perceptual prior probability distribution of motion speed can be defined with a power-law function:

$$p(v) = \kappa / v^{\alpha}$$

where *v* is the motion speed; and κ and α are two positive constants. That is, with increasing object speed, the probability decreases and hence the visual surprise increases. This also helps to compute the motion speed based temporal saliency using its self-information as

$$S_t = -\log p(v) = \alpha \log v + \beta$$

where $\beta = -\log k$ is a constant. The value of parameters α and β are based on the study [16]. Removing the GNC step for the Charbonnier penalty function causes high EPE on most sequences and higher energy on all sequences.

EXPERIMENTS & RESULTS

A publicly available Stereoscopic image dataset Middlebury 2005/2006/2014 is used to evaluate the performance of the proposed model. The database contains more than 30 stereoscopic images with various types such as outdoor scenes, indoor scenes, scenes including objects, scenes without any various object, etc. Some of them are shown below.

Figure 2: Sample Images of Middlebury dataset

Stimuli were displayed on a 26-inch Panasonic BT-3DL2550 LCD screen with a resolution of 1920 × 1200 pixels and refresh rate of 60 Hz. The stereoscopic stimuli was viewed by participants with a pair of passive polarized glasses at a distance of 93 cm. The environment luminance was adjusted for each observer and thus the pupil had an appropriate size for eye-tracking. The data was collected by SMI RED 500 remote eye-tracker and a chin-rest was used to stabilize the observer's head.

Generally, an efficient saliency detection model would have high response at the most attractive region and no response at random locations. The performance of the proposed model is measured by comparing the ground-truth and the saliency map from the saliency detection model. Some of the performance evaluation measures for Saliency detection models are KLD (Kullback Leibler Distance), PLCC (Pearson Linear Correlation Coefficient) and AUC (Area under the Receiver Operating Characteristics Curve).

Among this, Kullback Leibler Distance is used to measure the similarity between these two distributions. Here saliency distributions at the most attractive region and random locations are calculated over the saliency map as:

$$KL\ (H, R) = \frac{1}{2}\left(\sum_n h_n \log \frac{h_n}{r_n} + \sum_n r_n \log \frac{r_n}{h_n}\right)$$

where H and R are saliency distribution at the most attractive region and random locations with probability density function h_n and r_n respectively. The saliency detection model with larger KL distance value gives better performance.

PLCC (Pearson Linear Correlation Coefficient) is another measure, with its low value gives better performance. It is calculated directly from the comparison between the fixation density map and the predicted saliency map. Table 1 shows the performance evaluation measure values of PLCC and KL distance for existing methods and the proposed one.

$$r = \frac{\sum XY - \frac{(\sum X)(\sum y)}{n}}{\sqrt{(\sum X^2 - \frac{(\sum X^2)}{n})\ (\sum Y^2 - \frac{(\sum Y^2)}{n})}}$$

where X and Y are the pixel points in the saliency map.

Table I: KL Distance & PLCC value for existing methods and the proposed system.

Models	KL Distance	PLCC
Model [19]	0.364	0.6
Model [5]	0.346	0.58
Model [14]	0.438	0.46
Model [15]	0.301	0.424
Model[2]	0.443	0.4
Model[1]	0.6487	0.389
Proposed system	0.7030	0.301

From the table it is clear that the proposed method have the maximum KL Distance value over the previous methods, which means that the most salient region will be away from other random locations. This gives an accurate saliency map. Figure 3 shows the visual comparison of stereoscopic saliency detection models. It shows the final output obtained from different existing methods and the proposed one.

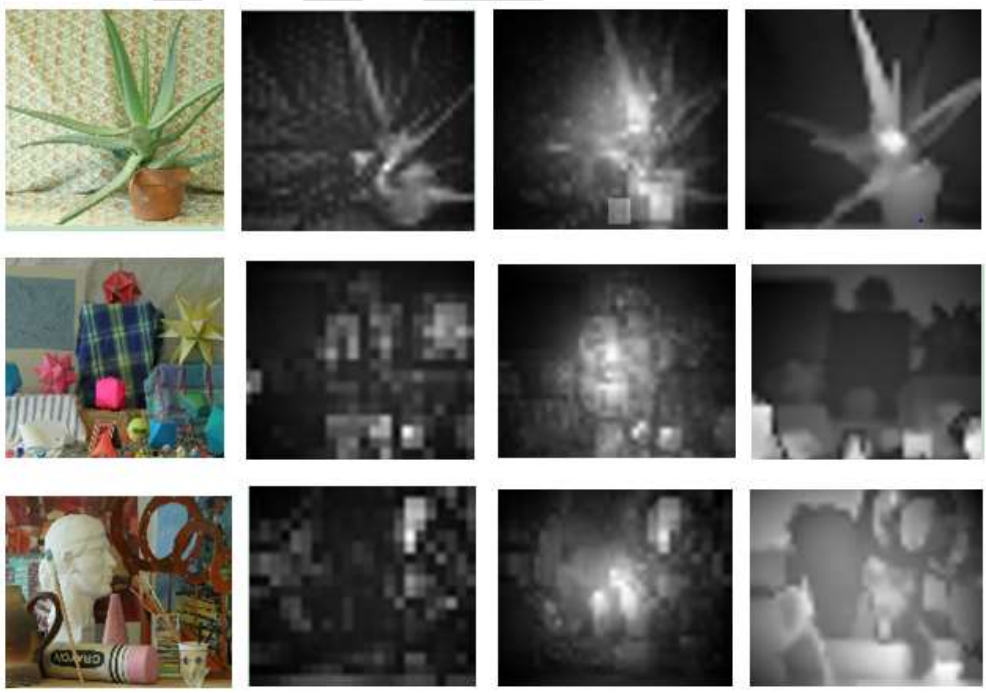

Figure 3. Visual comparison of stereoscopic saliency detection models. (a) Input Image (b) Saliency map using model [2] (c) Saliency map obtained using Adaptive feature map fusion[1] (d) Proposed model.

From the figure, we can clearly see that the output obtained from the optical flow method is clear over the other two. In [1] and [2], the salient region is spread and so that a particular region cannot point out. They destroys the overall information and makes the image

unclear. The proposed method is far better than the existing stereoscopic saliency detection methods [2], [1] and it almost preserves the scene as it is.

ACKNOWLEDGMENT

We wish to thank HOD of CS department, Project Coordinator and all friends of College of Engineering, Chengannur, for their immense support and encouragement for the fulfillment of this work.

CONCLUSION

Visual saliency is the distinct subjective perceptual quality which immediate attracts humans' attention. Saliency detection models finds this most attractive region. Different from all existing methods, we have proposed a new stereoscopic saliency detection using Optical flow method. It takes the concepts of relative motion of the scene in the Image captured by the left eye and right eye. The limitation is the availability of stereoscopic image pairs having good resolution. The method gives a promising and efficient result over all existing methods. In future, it can be enhanced for the salient region/object detection in videos.

REFERENCES:

[1] Yuming Fang, Junle Wang, Manish Narwaria, Patrick Le Callet, and Weisi Lin, "Saliency Detection for Stereoscopic Images", IEEE Transactions on Image Processing, Vol. 23, no. 6, June 2014, pp. 2625-2635.

[2] J. Wang, M. Perreira Da Silva, P. Le Callet, and V. Ricordel, "Computational model of stereoscopic 3D visual saliency," *IEEE Trans. Image Process.*, vol. 22, no. 6, Jun. 2013, pp. 2151–2165.

[3] J. M. Wolfe and T. S. Horowitz, "What attributes guide the deployment of visual attention and how do they do it?" *Nature Rev., Neurosci.*, vol. 5, no. 6, 2004, pp. 495–501.

[4] Y. Zhang, G. Jiang, M. Yu, and K. Chen, "Stereoscopic visual attention model for 3d video," in *Proc. 16th Int. Conf. Adv. Multimedia Model.*, 2010, pp. 314–324.

[5] Y. Fang, Z. Chen, W. Lin, and C.-W. Lin, "Saliency detection in the compressed domain for adaptive image retargeting," *IEEE Trans. Image Process.*, vol. 21, no. 9, Sep. 2012, pp. 3888–3901.

[6] V. Gopalakrishnan, Y. Hu, and D. Rajan, "Salient region detection by modeling distributions of color and orientation," *IEEE Trans. Multimedia*, vol. 11, no. 5, Aug. 2009, pp. 892–905.

[7] S. Goferman, L. Zelnik-Manor, and A. Tal, "Context-aware saliency detection," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 2376–2383.

[8] 8. C. Chamaret, S. Godeffroy, P. Lopez, and O. Le Meur, "Adaptive 3D rendering based on region-of-interest," *Proc. SPIE*, vol. 7524, Stereoscopic Displays and Applications XXI, 75240V, Feb. 2010.

[9] Ouerhani and H. Hugli, "Computing visual attention from scene depth," in *Proc. IEEE 15th Int. Conf. Pattern Recognit.*, Sep. 2000, pp. 375–378.

[10] E. Potapova, M. Zillich, and M. Vincze, "Learning what matters: Combining probabilistic models of 2D and 3D saliency cues," in *Proc. 8th Int. Comput. Vis. Syst.*, 2011, pp. 132–142.

[11] L. Itti and P. Baldi, "Bayesian surprise attracts human attention," *Adv. Neural Inform. Process. Syst.*, vol. 46, nos. 8–9, Apr. 2006, pp. 1194–1209.

[12] Y. Zhai and M. Shah, "Visual attention detection in video sequences using spatiotemporal cues," in *Proc. ACM Int. Conf. Multimedia*, 2006, pp. 815–824.

[13] Deqing Sun, Stefan Roth, Michael J. Black "Secrets of Optical Flow Estimation and Their Principles", in Computer Vision and Pattern recognition (CVPR),IEEE conference, 2010, pp. 2432-2439.

[14] Yuming Fang, Zhou Wang, Weisi Lin, "Video saliency incorporating spatiotemporal cues and uncertainty waiting", on International Conference on multimedia and expo(ICME),2013, pp.362-368.

[15] Yuming Fang, Weisi Lin, Zhenzhong Chen, "A Video Saliency Detection Model in Compressed Domain", IEEE trans. On circuits and systems for video technology, vol. 24, no. 1, Jan. 2014, pp. 27-38

[16] Zhou Wang and Qiang Li, "Video Quality Assessment Using a Statistical Model of Human Visual Speed Perception", *J. Opt. Soc. Amer. A*, vol. 24, no. 12, 2007, pp. B61–B69.

[17] X. Hou and L. Zhang, "Saliency detection: A spectral residual approach," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2007, pp. 1–8.

[18] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," in *Proc. Adv. NIPS*, 2006, pp. 545–552.

[19] 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, Nov. 1998, pp. 1254–1259.