VISUAL SALIENCY DRIVEN ERROR PROTECTION FOR 3D VIDEO

[*]Chaminda T.E.R. Hewage, [‡]Junle Wang, [*]Maria G. Martini, [‡]Patrick Le Callet
[*]Wireless Multimedia & Networking (WMN) Research Group, Kingston University-London, UK.
[‡]IVC Team, IRCCyN Lab, University of Nantes, France.

ABSTRACT

Viewers tend to focus into specific regions of interest in an image. Therefore visual attention is one of the major aspects to understand the overall Quality of Experience (QoE) and user perception. Visual attention models have emerged in the recent past to predict user attention in images, videos and 3D video. However, the usage of these models in quality assessment and quality improvement has not been thoroughly investigated to date. This paper investigates 3D visual attention model driven quality assessment and improvement methods for 3D video services. Moreover, a visual saliency driven error protection mechanism is proposed and evaluated in this paper. Both objective and subjective results show that the proposed method has significant potential to provide improved 3D QoE for end users.

Index Terms— Visual saliency, visual attention, visual attention model, 3D visual attention model

1. INTRODUCTION

Viewers tend to focus into specific regions of interest in an image, hence visual attention is one of the major aspects to understand Quality of Experience (QoE) and user perception. Eye tracking experiments are widely used to investigate user eye gaze positions during consumption of visual information. The collected eye movement data are then post-processed to obtain Fixation Density Maps (FDM) or saliency maps. There are two major approaches to analyze user visual attention, namely: free viewing task (i.e., bottom-up approach) and task oriented (top-bottom approach). The former approach is driven by low level image features such as spatial and temporal frequencies. The top-bottom approach is driven by the task. Several other factors influence visual attention such as sociocultural background, context, duration, etc. Visual attention models have emerged in the recent past to predict user attention in images, videos and 3D video [1-3]. These attention models predict user eye movements based on low level image features such as spatial frequency, edge information, etc. Visual attention models can therefore be used in image processing applications (e.g. post processing, image quality evaluation, image retargeting).

The attention of users during 3D viewing can be influenced by several factors including spatial/temporal frequencies, depth cues, conflicting depth cues, etc. A comprehensive analysis of visual attention in 3D, and of the weaknesses of existing models and their usage is discussed in [4]. The studies on visual attention in 2D/3D images found out that the behaviors of viewers during 2D viewing and 3D viewing are not always identical. For instance, the study in [5] for 2D/3D images has shown that added depth information increases the number of fixations, eye movement throughout the image and shorter and faster saccades. This observation is also complemented by the investigation carried out by Häkkinen et al. [6], which showed that eye movement during 3D viewing is more distributed. In contrast to these observations, Ramasamy et al.'s study in [7] found out that the spread of fixation points are more confined in 3D viewing than 2D viewing. These observations have direct influences in how we perceive 3D video. Therefore, effective 3D video quality evaluation and 3D QoE enhancement schemes could be designed based on these observations. The proposed image processing methods in the literature exploit these visual attention patterns and models to measure and improve 3D QoE.

Modeling visual attention in 2D viewing is driven by spatial and temporal frequencies of the image as suggested by many studies [8][9]. However, for 3D images/video, depth cues need to be added to the existing image features in order to generate a robust 3D saliency map. Most of the reported 3D visual attention models in the literature [10][11] are therefore based on scene depth information in addition to motion and spatial characteristics. 3D visual attention models can be divided into two main categories, as shown below:

- Depth weighted 3D saliency model (see Figure 1(a));
- Depth saliency based 3D saliency model (see Figure 1(b)).

The depth weighted model weighs the generated 2D saliency model based on the depth information in order to obtain the 3D visual saliency map. The second method generates two visual saliency maps: the first one for 2D image information and the second one for the corresponding depth map of the scene (see Figure 1(b)). Then both saliency maps are combined into one 3D saliency map based on the selected weights as described in (1).

In this paper, the 3D saliency model developed based on the depth saliency model is employed to identify visually salient areas [1]. This model considers both current image information and prior knowledge. However, this 3D saliency model does not take into account the temporal activity of the scene.

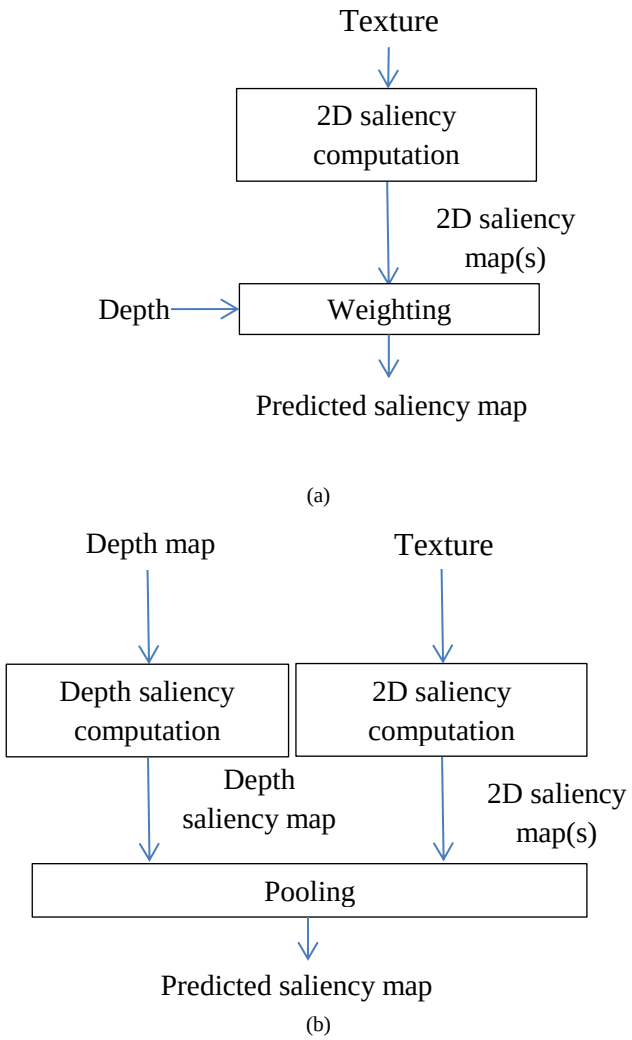

Figure 1. 3D saliency models, (a) Depth weighted 3D saliency model and (b) Depth saliency based 3D saliency model

$$SM_{3D} = w_1 \times SM_{Depth} + w_2 \times SM_{2D} \quad (1)$$

where w_1 and w_2 are weights assigned for the depth saliency model (i.e., SM_{Depth}) and 2D saliency model (i.e., SM_{2D}) respectively.

In this paper, we discuss how we could exploit 3D visual attention models to measure and improve 3D video quality. Moreover, a visual saliency based error protection mechanism is proposed and tested. The following two subsections briefly discuss how we could exploit 3D visual attention models to measure and improve 3D video perception and 3D QoE in general.

1.1. Visual attention models for quality evaluation

There are still unanswered questions such as whether quality assessment is analogous to attentional quality assessment and also how we could integrate attention mechanisms into the design of QoE assessment methodologies. 2D image/video quality assessment presented in [12], investigated the impact of different regions of interest on image quality evaluation. However, a thorough study has not been conducted to date in order to identify the relationship between 3D image/video attention models and 3D image/video quality evaluation. The COST action presentation in [13] identifies three main approaches to integrate visual attention into image/video quality evaluation (see Figure 2). Similar to the integrated model described above, attentive areas identified by visual attention studies can be utilized to extract image features which can be used to design No-Reference (NR) and Reduced-Reference (RR) quality metrics for real-time 3D video application. The use of extracted features to design RR 3D image/video quality metrics have been undertaken in previous research [14][15][16][17]. Furthermore, the use of 3D visual saliency information could be used to further reduce the amount of side-information for real-time quality evaluation.

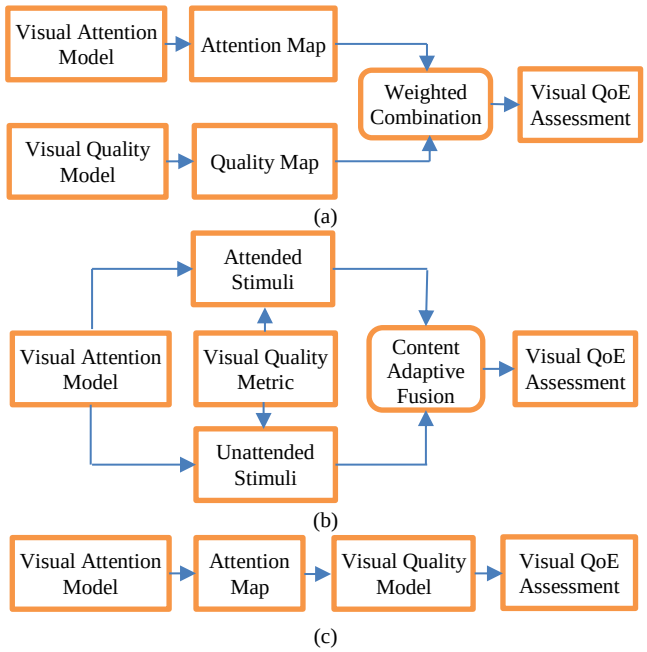

Figure 2. Integration of visual attention model in quality evaluation [13]; (a) Direct combination (b) Divided integration and (c) Integrated combination

1.2. Visual attention models for quality improvement

Since visual attention models can predict the highly attentive areas of an image or video, these can be integrated into video coding at the source-end. The proposed ROI-based encoding methods for 2D/3D video have shown improved quality at a given bitrate compared to

conventional encoding methods [18][19]. For instance the ROI based encoding method proposed and evaluated in [19] shows that by protecting combined edges of colour plus depth based 3D video, the overall quality of the rendered views can be improved. This study also incorporates an Unequal Error Protection (UEP) mechanism to protect different image regions. However, visual attention based ROI encoding methods have not been reported for 3D video applications to date. Therefore in this paper we investigate how we could incorporate 3D visual attention models to efficiently encode 3D video based on ROI coding and protect it over unreliable wireless channels. The proposed 3D visual saliency based UEP mechanism is simulated and tested.

This paper is organized as follows. Section 2 describes the proposed error protection mechanism. Results and discussion are presented in Section 3. Section 4 concludes the paper.

2. PROPOSED ERROR PROTECTION METHOD

The proposed error protection mechanism starts at the source end before encoding. The graphical illustration of the proposed mechanism is shown in Figure 3 at the end of the paper. The *barrier* sequence from the NAMA3DS1-COSPAD1 3D HD dataset is used for this illustration [20]. The 3D visual saliency model described in [1] is employed to identify the visual saliency region for given left and right stereoscopic image sequence. This 3D saliency model generates two saliency maps:

- 2D saliency map (Figure 3 (c))
- Depth saliency map (Figure 3 (d))

These two saliency maps are then converted into binary images by thresholding the original saliency maps (see Figure 3 (e) and (f)). Subsequently, these two saliency maps are combined to form a 3D saliency map (g) using chosen weights. In this case w_1 and w_2 of (1) are selected as 0.5. However, optimum weights could be found based on further experimentation. Figure 3 (h) shows the visually salient region identified in the original left image.

Once the visually salient region is identified as described above, a Region of Interest (ROI) map is generated and passed to the encoder. Sample identified ROIs are shown in Figure 4. ROI 0 represents the attentive area identified by the 3D visual attention model. The visually less salient region is represented as ROI 1 in Figure 4. These ROI maps will be used to encode left and right video with region of interest encoding of H.264/AVC. The ROI map is transmitted to the decoder via the Picture Parameter Set (PPS) of H.264/AVC bit-stream. The update frequency of the ROI map can be selected based on the available bandwidth and sequence characteristics. For instance, if the motion activity of the video is significant, we could send frequent ROI updates based on the 3D saliency model.

Once ROIs are encoded, application layer channel coding (error correction codes) is used in the proposed method to protect the ROIs. An Unequal Error Protection (UEP) mechanism is employed to protect ROI 0 and ROI 1 unequally. For instance, ROI 0 (highly attentive area) is protected by a lower channel code rate (higher redundancy and protection) and the ROI 1 is protected using a higher channel code rate (lower redundancy and protection). When this 3D video is sent over an unreliable communication channel, information of attentive region (ROI 0) will be recovered with a high probability of success due to stronger error correction codes whereas information of the less attentive region (ROI 1) will be subjected to more uncorrectable errors. As a result, the attentive area will be reconstructed with less errors compared to the less attentive area identified by the 3D visual attention model. This would result in improved quality compared to equally protected 3D video.

The performance of the proposed method is compared with an Equally Error Protection (EEP) mechanism at the same bitrate. The results and discussions are presented in Section 3.

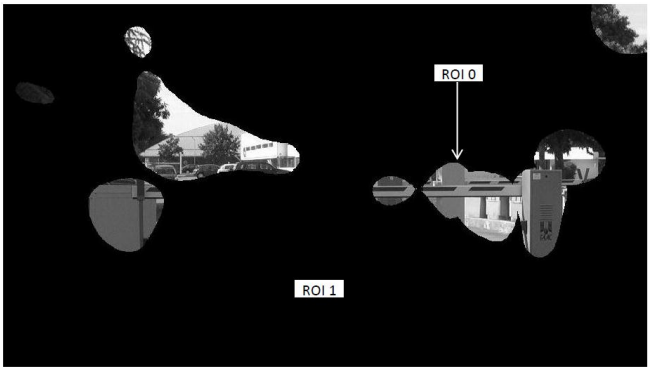

Figure 4. Region of Interest (ROI) based on visual saliency information

3. RESULTS AND DISCUSSIONS

The *barrier* 3D HD video sequence from NAMA3DS1-COSPAD1 database [20] is compressed using the H.264/AVC compression standard with Quantization Parameter (QP) = 30 to obtain good image quality at a reasonable bitrate. A sixteen-seconds sequence (400 frames at 25 *fps*) is considered for the experiment. Simulcast encoding approach (i.e., two parallel encoders) is employed to encode both left and right image sequences. They were encoded using the *IPPP…IPPP…* frame sequence format to provide a high quality 3D video stream. An *I* frame is encoded at every 1 *second* interval. In this study, the ROI selection frequency is set to 1 *second* (i.e., a separate ROI map is generated by every 1 *second* using the 3D saliency model). Therefore, a 3D saliency map is generated for every *second* using the 3D saliency model described in [1]. However, during the initial experiments, it is observed that the ROI update frequency should be increased when the

motion activity is high in the scene. Hence, the ROI update frequency is set to every 10 frames for the video segments which have high motion activity. A 3D saliency map is used to identify the saliency region of each image (i.e., ROI 0). The rest of the surrounding area (less attentive area) is marked as ROI 1. ROI 0 and ROI 1 of a sample image are illustrated in figure 3. For the 3D video sequence considered for the experiment, the average amount of visually attentive area (saliency region) is about 15% of the whole image. However, this increases up to 25% of the whole image after encoding, since most of the activities happen within this attentive region (ROI 0).

We assume that our 3D video source is affected, due to transmission, by random packet losses, where losses occur with the same probability p_L in the different portions of the stream. We assume p_L = 20% when FEC is not applied. We adopt Reed Solomon (RS) coding at the application layer, where packets are inserted in a matrix in rows and coding is performed in columns.

The adopted packet size is 1000 Bytes. In the case of equal error protection (EEP), we adopt the RS code (23,31), whereas in the case of unequal error protection (UEP) the ROI 0 containing visually saliency region is protected with an RS code (21,31) and the packets associated to the remaining part of the image (i.e., less attentive region, ROI 1) are protected with an RS code (24,31). The channel coding rates are selected such that the overall bitrate after FEC is the same for both the Equal Error Protection (EEP) and UEP methods. With the aid of the considered codes, the packet loss ratio is reduced to 5% for the EEP case, whereas for the UEP case we achieve 1% packet loss rate for ROI 0 (i.e., saliency information) and 7% PLR for ROI 1 (non-salient region). Packets of the first and subsequent *I* frames are not subject to losses in order to decode all the frames of the bit-streams smoothly. In order to obtain average results, simulations are run for several times. At the decoder, the missing packets are copied from the corresponding frames of the previous time instance.

Table 1: Average left and right image quality

Video Segment	Average Left and Right Image quality (PSNR/dB)			
	Encoded	No Error Protection	EEP	UEP
Segment 1	36.23	29.42	35.49	35.33
Segment 2	36.23	26.72	35.16	35.30
Segment 3	36.20	25.67	32.48	33.02
Segment 4	36.47	23.81	30.09	31.67
Overall	36.31	26.17	33.31	33.83

The objective quality results achieved with the proposed UEP and reference methods (i.e., EEP, no protection) are listed in Table 1 and Figure 5. In addition, the quality of the encoded sequences is also shown. Results are presented for four segments of 100 frames (where a separate ROI map is sent for every 25 frames) and for the long sequence of 400 frames in general. It is evident that, when no error protection is deployed, the resultant average quality is significantly low. However, when UEP or EEP methods are used the resultant average stereo image quality is high. The proposed UEP method outperforms the reference EEP method in all the cases. This suggests that the proposed UEP method can achieve improved results with a high quality saliency region when 3D video is transmitted over unreliable networks. The quality improvement is significant towards the latter part of the sequence where high motion activity occurs. This shows that the proposed method is effective when the highly dynamic objects come within the visually saliency region of the video.

In order to evaluate the true user perception, subjective quality tests are performed using 15 expert subjects. Subjective experiments are carried out in both IRCCyN Lab, University of Nantes, France and WMN Research Group, Kingston University-London, UK. HD 3D displays with polarized and active shutter glasses are used for subjective experiments. The DSIS method is used to record subject's opinions. Results are presented in Table 2. The results show a clear improvement with the proposed UEP method compared to the EEP method. Therefore, it is evident that by protecting visually attentive region compared to visually less attentive region, we could improve the 3D QoE for 3D video transmission over unreliable networks.

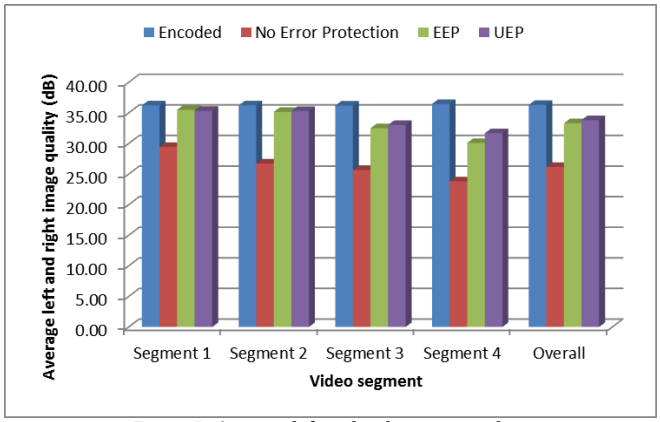

Figure 5. Average left and right image quality

Table 2: Subjective results

Method/Sequence	MOS
Original sequence	4.75
Encoded sequence	4.50
No error protection	2.75
EEP	3.50
UEP	3.75

Even though preliminary results are promising, further experiments will be carried out to improve the performance of the proposed method. For instance, we did the tests with high ROI update frequency. However, this may not be enough for highly dynamic sequences where objects are moving rapidly. Furthermore, the used 3D visual attention

model does not incorporate the motion activity of the scene. Therefore, it fails to identify attentive area based on the temporal activity and this may cause the loss of more packets in highly dynamic regions of the image. Integrating temporal cues into the 3D visual attention model will enable us to obtain an effective 3D saliency map and improved QoE with the proposed method.

4. CONCLUSION

The paper elaborates on the use of 3D visual attention details in quality measurements and improvements. The proposed visual saliency driven UEP method achieve better quality compared to EEP method at the same bitrate. Therefore, visual attention driven error protection mechanisms as described in this paper will enable us to deliver 3D video over unreliable communication channels with significantly improved 3D QoE.

5. ACKNOWLEDGEMENT

This work was supported in part by the EU FP7 Programme (CONCERTO project) and QUALINET EU COST Action .

6. REFERENCES

[1] J. Wang, M.P. Da Silva, P. Le Callet, V. Ricordel, "Computational Model of Stereoscopic 3D Visual Saliency", *IEEE Transaction on Image Processing*, vol. 22, no. 6, pp. 2151–2165, June 2013.

[2] X. Hou and L. Zhang, "Saliency detection: A spectral residual approach," *in Proc. of IEEE Inter. Conference on Computer Vision and Pattern Recognition*, June 2007, pp. 1–8.

[3] Y. Zhang, Gangyi Jiang, Mei Yu, Ken Chen, "Stereoscopic Visual Attention Model for 3D Video", *Advances in Multimedia Modelling: Lecture Notes in Computer Science*, vol. 5916, pp. 314-324, 2010.

[4] Q. Huynh-Thu, M. Barkowsky, P. Le Callet, "The Importance of Visual Attention in Improving the 3D-TV Viewing Experience: Overview and New Perspectives", *IEEE Transactions on Broadcasting*, vol. 57, no. 2, pp. 421-431, 2011.

[5] L. Jansen, S. Onat, and P. König,, "Influence of disparity on fixation and saccades in free viewing of natural scenes", *Journal of Vision*, vol. 9, no. 1, pp. 1–19, Jan. 2009.

[6] J. Häkkinen, T. Kawai, J. Takatalo, R. Mitsuya, and G. Nyman, "What do people look at when they watch stereoscopic movies?", *in Proc. SPIE Conf. Stereoscopic Displays and Applications XXI*, vol. 7524, San Jose, January 2010.

[7] C. Ramasamy, D. House, A. Duchowski, and B. Daugherty, "Using eye tracking to analyze stereoscopic filmmaking", *in Proc. SIGGRAPH 2009:Posters*, 2010.

[8] M. Treisman and G. Gelade, "Feature integration theory of attention", *Cognitive Psychology*, vol. 12, pp. 97–136, January 1980.

[9] O. Le Meur and P. Le Callet, "What we see is most likely to be what matters: visual attention and applications", *in Proc. IEEE International Conference on Image Processing*, Cairo, Nov 2009, pp. 3085–3088.

[10] A. Maki, P. Nordlund, and J. O. Eklundh, "Attentional scene segmentation: Integrating depth and motion from phase", *Computer Vision and Image Understanding*, vol. 78, pp. 351–373, 2000.

[11] N. Ouerhani and H. Hügli, "Computing visual attention from scene depth", *in Proc. International Conference on Pattern Recognition*, Barcelona, 2000, pp. 375–378.

[12] O. Le Meur, A. Ninassi, P. Le Callet and D. Barba, "Overt visual attention for free-viewing and quality assessment tasks. Impact of the regions of interest on a video quality metric", *Elsevier, Signal Processing: Image Communication*, vol. 25, no. 7, August 2010.

[13] J. You, Visual Attention Driven QoE: Towards Integrated Research, Norwegian University of Science and Technology, *Qualinet COST Action: Doc-Qi0179.*, WG2, Prague, Feb 2012

[14] C.T.E.R. Hewage, and M.G. Martini, "Edge based reduced-reference quality metric for 3D video compression and transmission", *IEEE Journal of Selected Topics in Signal Processing* vol. 6, no. 5, pp. 471-482, Sept 2012.

[15] C.T.E.R. Hewage, and M.G. Martini, "Reduced-reference quality assessment for 3D video compression and transmission", *IEEE Transactions on Consumer Electronics*, vol. 57, no. 3, pp. 1185-1193, Aug. 2011.

[16] C.T.E.R. Hewage, S.T. Worrall, et al, "Quality evaluation of color plus depth map-based stereoscopic video", *IEEE Journal of Selected Topics in Signal Processing*, vol. 3, no. 2, pp. 304-318, 2009.

[17] C.T.E.R. Hewage, and M.G. Martini, "Quality evaluation for Real-time 3D video services", *In: 2nd International Workshop on Hot Topics in 3D*, 15 July 2011, Barcelona, Spain.

[18] M.G. Martini, and C.T.E.R. Hewage, "Flexible Macroblock Ordering for context-aware ultrasound video transmission over mobile WiMAX", *International Journal of Telemedicine and Applications*, 2010, ISSN (print) 1687-6415.

[19] C.T.E.R. Hewage, and M.G. Martini, "ROI-based transmission method for stereoscopic video to maximize rendered 3D video quality", *In: Stereoscopic Displays and Applications XXIII*; Jan 2012, California, U.S.A. (Proceedings of SPIE, no. 8288).

[20] M. Urvoy, M. Barkowsky, et al, "NAMA3DS1-COSPAD1: Subjective video quality assessment database on coding conditions introducing freely available high quality 3D stereoscopic sequences," *4th International Workshop on Quality of Multimedia Experience (QoMEX)*, July 2012, pp.109-114.

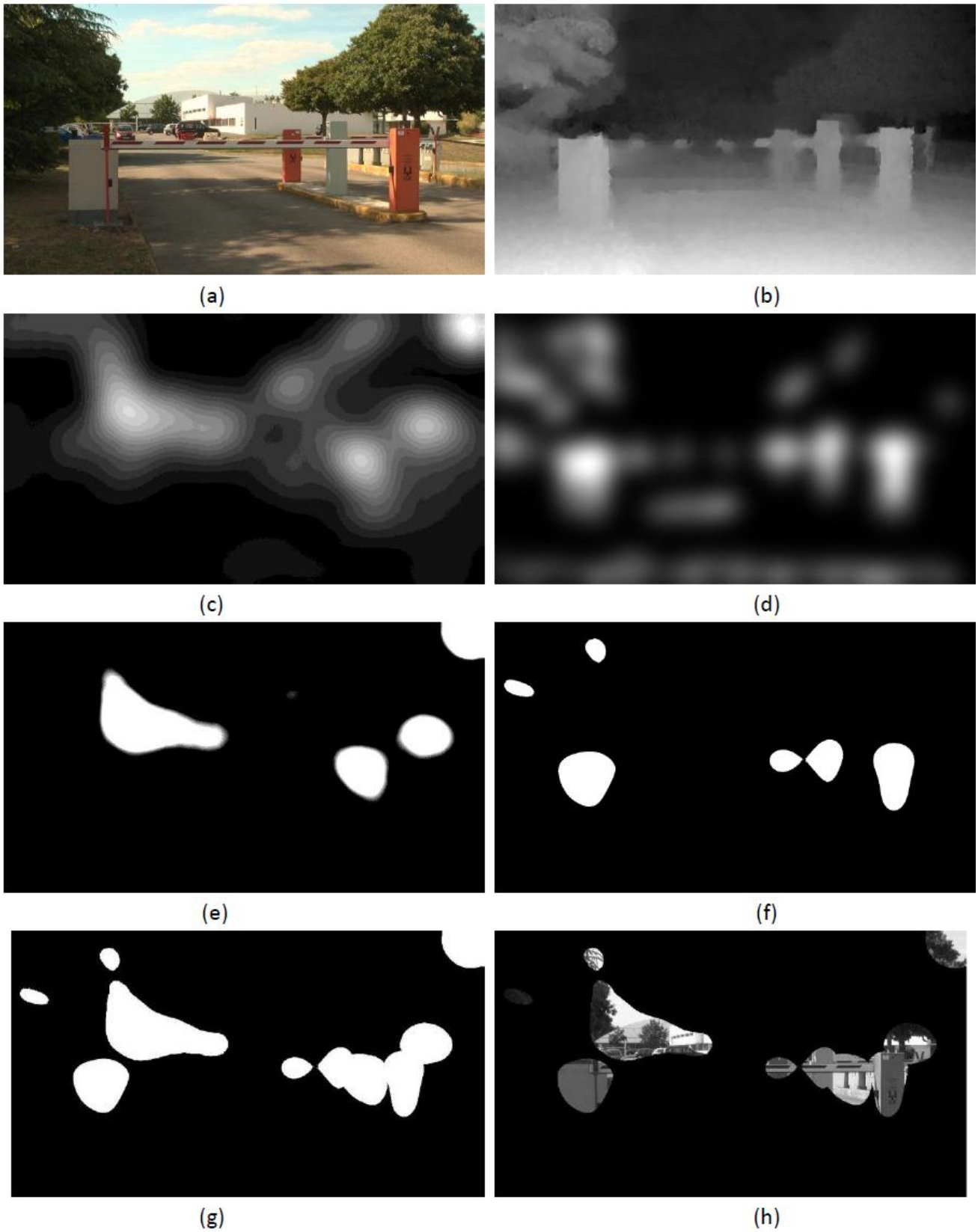

Figure 3. Sample image from the *Barrie*r sequence, (a) The original left image; (b) corresponding depth map (generated based in optical flow analysis); (c) 2D saliency map based on the left image; (d)Depth saliency map based on the depth image; (e) binary 2D saliency map (after applying a threshold); (f) binary depth saliency map (after applying a threshold); (g), predicted 3D saliency map (binary); and (h) the identified saliency area of the left image