Research Article

Anna Rogalska* and Piotr Napieralski

The visual attention saliency map for movie retrospection

https://doi.org/10.1515/phys-2018-0027
Received Nov 04, 2017; accepted Nov 24, 2017

Abstract: The visual saliency map is becoming important and challenging for many scientific disciplines (robotic systems, psychophysics, cognitive neuroscience and computer science). Map created by the model indicates possible salient regions by taking into consideration face presence and motion which is essential in motion pictures. By combining we can obtain credible saliency map with a low computational cost.

Keywords: saliency prediction, visual attention, gaze tracking, computer graphics

PACS: 42.30.-d, 87.19.lt

1 Introduction

The film and motion picture industry is revenue of the forecast to increase nearly to 50 billion U.S. dollar in 2020 [1]. While the US system is profitable, other film industries tend to be financed through government subsidies. It is important not to exceed the movie budget, while maintaining desired picture quality. That is why it is important to review pre-production scenes. The proposed solution allows for the verification if a certain image region visually stands out, compared to its surrounding area, important for the audience (visual saliency). Saliency-extraction techniques of this region generate a saliency map initially presented by Itti *et al.* [2]. Saliency map indicates regions that are likely to attract human attention. Red regions indicate high saliency level, blue regions indicate low saliency level. Figure 1 shows an example saliency map (presented as a heatmap for legibility).

Nowadays domain of visual attenuation has evolved in the many scientific disciplines (robotic systems, psychophysics, cognitive neuroscience and computer science).

Visual saliency detection is widely used in video compression, image retargeting, video retargeting or object detection [3]. Saliency-detection methods, combine a probability of saliency with spatial variance of color of a superpixel, global contrast map or spatially-aware (non-global) color quantization [4]. The final saliency map is calculated by linearly interpolating the product of these components. Choice of these components is made by the fact that some cues take more attention in the observed scene. Human Visual System (HVS) is sensitive to features like changes in color, specific shapes, light intensity or human faces [5]. These features indicate Regions Of Interest (ROI). Such regions, with certain probability, are likely to attract attention. According to Yarbus [6] the task that viewers have will also influence ROI. For instance, when a viewer was asked to simply look at the picture (figure 2) his gaze track looked different compared to the gaze track after questions like “How old are people in the room?” or “How are they dressed?”.

Figure 1: Left: original frame, right: saliency map of frame

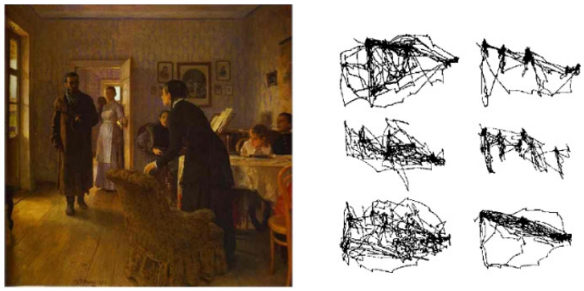

Figure 2: Different gaze path for the same picture and different tasks [6]

***Corresponding Author: Anna Rogalska:** Institute of Information Technology, Lodz University of Technology ul. Wólczańska 215, 90-924 Lodz, Poland; Email: 800667@edu.p.lodz.pl
Piotr Napieralski: Institute of Information Technology, Lodz University of Technology ul. Wólczańska 215, 90-924 Lodz, Poland; Email: piotr.napieralski@p.lodz.pl

These differences in gaze track are due to HVS being able to pay attention to only limited number of ROI at certain point of time.

When attention cues are accumulated in the picture, some of them are more likely to attract attention than the others. For example when there is a girl dressed in yellow raincoat what attracts attention first is her face and later – the yellow raincoat (contrast, intensity). This becomes more complicated where motion picture is analyzed instead of a static one. Duration of shot influences ROI, because some attention cues attract gaze earlier than the other. Referring to previous example, when the shot is only two seconds long, only girl's face will attract attention. If the shot is longer, then after a while viewers will look at the rest of the scene (for example at yellow raincoat).

Popularity of stereoscopic movies, bring metrics of visual comfort for stereoscopic content [7–9]. These methods are far more complicated as they take into account disparity (which can itself lead to loss in visual comfort [10]), planar and depth velocities, as well as the spatial frequency of luminance contrast and motion on the screen plane. The saliency stereoscopy models take into account more components than an ordinary picture. This model is made by fusing depth perception and visual comfort factors, like image and visual comfort based saliency, and above that depth saliency.

2 Existing methods

Saliency prediction models can be divided into two basic categories: bottom-up and top-down. Bottom-up means that we analyze an image but not what is in the image. For example color, intensity and orientation are bottom-up cues, while face presence, movement, interesting or emotional objects are top-down cues. There already exist several bottom-up models such as Itti *et al.* [2] which is based on color, intensity and orientation and combines multiscale image features into a single topographical saliency map, Levine *et al.* [11] combines local contrast with global contrast in order to detect both small and large saliency regions, detect saliency in cluttered scenes and inhibit repeating patterns. Yildrim *et al.* [4] combines global contrast with probability of saliency which is based on spatial center and variances of colors. These models include only bottom-up features and ignore top-down ones. This can lead to building inaccurate saliency maps. Several top-down models, such as: Xu *et al.* [12] proposed not only face-based saliency map, but also facial features-based saliency map. He verified that the fixations tend to cluster around facial features, when viewing images with large faces. Yu *et al.* [15] predicted saliency maps by indicating interesting objects using a novel object extraction approach by integrating sketch-like and envelope-like maps.

Unfortunately bottom-up features cannot be ignored as they attract attention during first 500 milliseconds. Because of that, there were developed models that combine both bottom-up and top-down features. For example Xu et at. [16] combined Ittis model with movement by using adaptive fusion method that merges the spatial and temporal saliency maps of each frame into its spatiotemporal saliency map. Kucerowa *et al.* [5] chose texture and faces as attention cues and combined them using local context suppression of multiple cues.

Most existing models take into consideration static images. However motion picture plays an essential role in film industry. It may turn out that some attention cues are far more important when it comes to motion picture. For example it is arguable if face will attract more attention than contrast or movement.

Another thing is that some of presented models are not versatile. As an example: method presented by Kucerowa *et al.* [5] will not work for images without text or face, as saliency map is based only on face and text presence.

3 Method description

We propose a model that is versatile and can be used in motion pictures. We work on videos with natural scenes, with or without faces. Proposed model, compared to state-of-the-art methods was improved by adding more attention cues. Based on the eye tracking measurements [13], proposed model evaluates the level of importance regions associated to viewing movie or motion picture [14]. We create saliency map by combining four attention cues:

- Global contrast
- Distance from the center of an image
- Face
- Movement

We combine these features into a single saliency map using following equation:

$$S(i) = w_1 \star K(i) + w_2 \star G(i) + w_3 \star F(i) + w_4 \star M(i)$$

$$w_1 + w_2 + w_3 + w_4 = 1$$

w_1, w_2, w_3, w_4 – *map weights*

i – *image pixel*

K – *contrast*

G – *distance* from *the center of an image*

F – *face presence*

M – *movement*

S – *final saliency map*

Using this equation we can indicate which attention cues are more important than others, so the final saliency map can look differently depending on weight parameters. In the following section we elaborate on the importance of weight parameters. We also test how different weight parameters influence saliency map accuracy.

3.1 Global contrast

We use global contrast instead of local contrast for improved efficiency. We construct histogram and calculate difference for each color.

$$K(i) = \sum_{j=1}^{m} f_j \left| C_i - C_j \right|$$

$K(i)$ – *saliency value for* color

m – *number of colors*

C – color *value*

f_j – *frequency of* color C_j

3.2 Distance from the center of an image

Research has showed that the probability of salient object appearance decreases with the distance from the center of an image. We use Gaussian function to model this.

$$G(x, y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2+y^2}{2\sigma^2}}$$

x, y – *coordinates of pixel with respect* to *image center*

3.3 Face

In this case, research has shown that face is first to attract attention during first 500 milliseconds of viewing (as observed in 80% of cases). For face detection we use Haar clasificatior.

3.4 Movement

In order to detect motion we use optical flow, using Gunnar Farnebacks [19] algorithm. We assume that pixel intensity does not change in consecutive frame. We calculate the distance between a pixel in the first and in the second frame.

$$M(x, y, t) = M(x + \Delta x, y + \Delta y, t + \Delta t)$$

i – *image pixel*

x, y – *pixel coordinates*

4 Experimental results

We compared our saliency map with ground truth map, which is comprised of human eye fixation map. The above methods were tested on $1920 \times 1080px$, $1m42s$ long video which is *Isyrius* company marketing material. The video consists of different types of scenes with variety of color schemes, different dynamics and shot length. This particular video was chosen because of its dynamic montage, as well as, shot richness and variety. We obtained eyetracking data by using *Tobii* eyetracker and tested 18 persons aged from 22 to 37. On computer with *Intel Core i7* processor and 8GB RAM the saliency map was calculated after 210 minutes. That gives 0.2 fps processing speed. Figure 3 shows a schematic view of our setup.

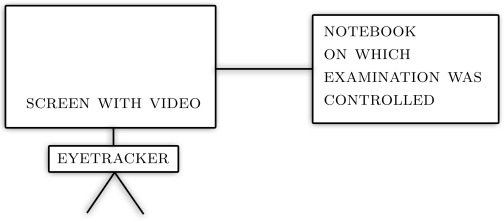

Figure 3: Schematic view of research stand

We tested accuracy of proposed method by comparing calculated saliency map S_A (its black and white version) with ground truth map S_B, using Linear Correlation Coefficient (LCC):

$$LCC = \frac{\sum_{i=1}^{n} \left(S_A(i) - \overline{S_A}\right)\left(S_B(i) - \overline{S_B}\right)}{\sqrt{\sum_{i=1}^{n} \left(S_A(i) - \overline{S_A}\right)^2} \sqrt{\sum_{i=1}^{n} \left(S_B(i) - \overline{S_B}\right)^2}}$$

n – *number of image pixels*

i – *pixel*

S_A – *calculated saliency map*

S_B – *ground truth map*

Figure 4 shows scalculated saliency map S_A and ground truth map (eyetracking map) S_B for an example frame.

Figure 4: Left: calculated saliency map S_A, right: ground truth map S_B

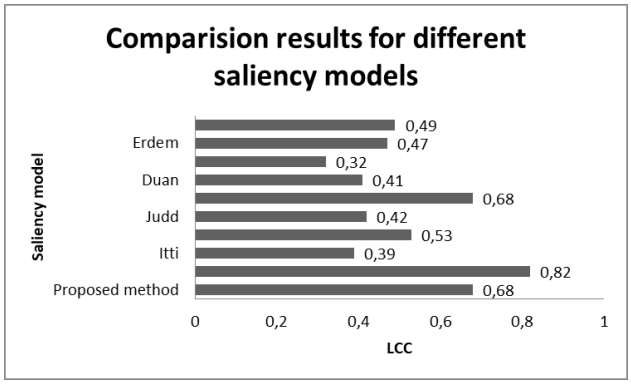

Figure 5: Comparison results for different saliency models

Figure 6: Right column (from top):. global contrast based saliency map, distance from center based saliency map, face presence based saliency map, motion based saliency map

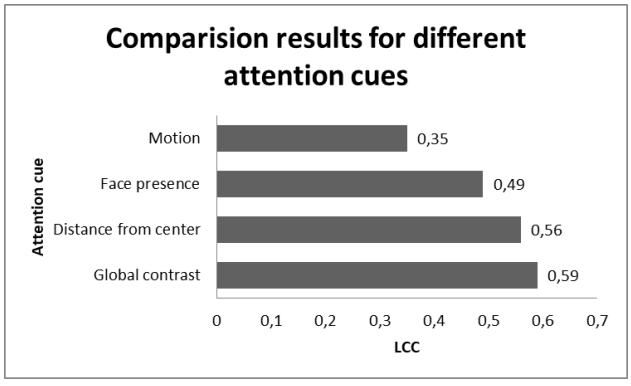

Figure 7: Comparison results for different attention cues

LCC was calculated for each frame and averaged for the whole video clip. Figure 5 shows LCC for our method compared to the other existing methods.

Figure 6 shows four different saliency maps (right column). The first map from the top is a global contrast based saliency map. Another map is generated by distance from center attention cue. Third map represents face presence based saliency map. The last map is motion based saliency map. In the bottom-left corner eyetracking map (overlayed on original frame) is shown. In the top-left corner saliency map is presented as a heatmap. This heatmap was generated using weights: contrast 30%, distance from center 20%, face 20%, motion 30%. Following the above, there is another saliency map (generated with different weights: contrast 10%, distance from center 20%, face 20%, motion 50%), and further down to the right is the original footage. Such presentation shows how each of attention cues influences final saliency map. Moreover, it can be compared at the same time with actual fixation points.

As can be seen in figure 6, different attention cues generate completely different saliency maps. Looking at figure 7 we can also see that they are also responsible for different LCC values. Figure 7 depicts how each attention cue influence LCC (each saliency map was calculated for every frame of the video). For video used for this research, LCC with use of three maps has value of 0.68. As we can see in figure 7 each of maps alone provides less LCC value. That indicates that combing them in a right way (with optimal weights) will generate credible saliency map.

5 Conclusion

Our method can be very useful in the film industry. It indicates possible salient regions by taking into consideration global contrast, distance from the center of an image, face presence and motion. Our method is versatile and suitable for fullHD videos. In contrast, previous methods considered mostly robust static features and were computationally complex so they are not suitable for videos. Also, many state-of-the-art methods can be completely misleading in motion picture as they neglect the importance of different attention cues, which we take into consideration.

We carried out several tests that gave us eyetracking data which we could compare with generated saliency maps. These data showed the deficiency in our method and at the same time provided us with interesting observations concerning relations between different attention cues.

Future work considers designating optimal weight parameters when combining obtained four saliency maps into final saliency map, adding more attention cues (specific for film industry) as horizon line and golden proportion, and improving processing speed.

References

[1] The Statistics Portal, https://www.statista.com, 20.04.2017

[2] Itti L., Koch C., Niebur E., A model of saliency-based visual attention for rapid scene analysis., IEEE Transactions on PAMI 20, 1998

[3] Puchala D., Yatsymirskyy M., Joint Compression And Encryption of Visual Data Using Orthogonal Parametric Transforms, Bulletin of the Polish Academy of Sciences-Technical Sciences, 2016, 64, 2, 373-382.

[4] Yildirim G., Susstrunk S., FASA: Fast, Accurate, and Size-Aware Salient Object Detection, Computer Vision, 2015

[5] Kucerova J., Sikudova E., Saliency map augmentation with facial detection, Proceedings of CESCG, 2011

[6] Yarbus A.L., Eye Movements and Vision, Plenum, New York, 1967.

[7] Du S., Masia B., Hu S., Gutierrez D., A metric of visual comfort for stereoscopic motion, ACM Trans. Graph., 2013, 32, 6, 222

[8] Pollock B.T., Burton M., Kelly J.W., Gilbert S., Winer E., The Right View from the Wrong Location: Depth Perception in Stereoscopic Multi-User Virtual Environments, IEEE Transactions on Visualization and Computer Graphics, 2012, 18, 581-588.

[9] Qiuping J., Feng Sh., Gangyi J., Mei Y., Zongju P., Changhong Y., A depth perception and visual comfort guided computational model for stereoscopic 3D visual saliency, Signal Processing-Image Communication, 2015

[10] Kowalczyk M., Napieralski P., Noise resistant method enabling detection of vertical disparity in three-dimensional visualizations, 2017 18th International Symposium on Electromagnetic Fields in Mechatronics, Electrical and Electronic Engineering (ISEF) Book of Abstracts, Lodz, Poland, 2017, 1-2, DOI: 10.1109/ISEF.2017.8090694

[11] Levine M., An X., He H., Saliency detection based on frequency and spatial domain analyses, Proceedings of the BMVC, 2011

[12] Xu M., Ren Y., Wang Z., Learning to Predict Saliency on Face Images, IEEE International Conference on Computer Vision (ICCV), 2015, 3907-3915

[13] Wojciechowski A., Fornalczyk K., Single web camera robust interactive eye-gaze tracking method, Bulletin of the Polish Academy of Sciences-Technical Sciences 2015, 63, 879-886.

[14] Rogalska A., Napieralski P., A model of saliency-based visual attention for movie retrospection, 2017 18th International Symposium on Electromagnetic Fields in Mechatronics, Electrical and Electronic Engineering (ISEF) Book of Abstracts, Lodz, Poland, 2017, 1-2, DOI: 10.1109/ISEF.2017.8090692

[15] Yu H., Li J., Tian Y., Huang T., Automatic interesting object extraction from images using complementary saliency maps, Proceedings of the 18th ACM international conference on Multimedia, 2010

[16] Xu J., Tu Q., Li C., Gao R., Men A., Video saliency map detection based on global motion estimation, IEEE International Conference on Multimedia and Expo Workshops (ICMEW), 2015, 1-6.

[17] Farneback G., Two-frame motion estimation based on polynomial expansion, Lecture Notes in Computer Science, 2003, 363-370.

[18] Cheng M.M., Mitra N.J., Huang X., Torr P.H.S., Hu S.M., Global Contrast Based Salient Region Detection, IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37, 3, 569-582.

[19] Borji A., Boosting bottom-up and top-down visual features for saliency estimation, IEEE Conference on Computer Vision and Pattern Recognition, 2012

[20] Ko B., Nam J., Object-of-interest image segmentation based on human attention and semantic region clustering, J. Opt. Soc. Am., 2006, 23(10), 2462

[21] Liu T., Yuan Z., Sun J., Wang J., Zheng N., Tang X., Shum H., Learning to detect a salient object, CVPR, 2007, 1-8.

[22] Judd T., Ehinger K., Durand F., Torralba A., Learning to predict where humans look, ICCV, 2009

[23] Subramanian R., Katti H., Sebe N., Kankanhalli M., Chua T.S., An eye fixation database for saliency detection in images, ECCV, 2010