Image and video saliency models improvement by blur identification

Yoann BAVEYE, Fabrice URBAN, Christel CHAMARET

Technicolor, 1 av. de belle fontaine, CS 17616 35576 Cesson Sevigne, France
fabrice.urban@technicolor.com

Abstract. Visual saliency models aim at predicting where people look. In free viewing conditions, people look at relevant objects that are in focus. Assuming blurred or out-of-focus objects do not belong to the region of interest, this paper proposes a significant improvement and the validation of a saliency model by taking blur into account. Blur identification is associated to a spatio-temporal saliency model. Bottom-up models are designed to mimic the low-level processing of the human visual system and can thus detect out-of-focus objects as salient. The blur identification allows decreasing saliency values on blurred areas while increasing values on sharp areas. In order to validate our new saliency model we conducted eye-tracking experiments to record ground truth of observer's fixations on images and videos. Blur identification significantly improves fixation prediction for natural images and videos.

Keywords: Visual attention, Saliency, Blur, image processing

1 Introduction

In image or video content, focal blur or out-of-focus blur occurs when objects in the scene are placed out of the focal range imposed by the focal length of the camera. This limitation is also used as an artistic effect by photographers to reduce the depth of field to emphasize an object in focus. In the human visual system objects need additionally to be projected onto the fovea, the part of the retina that contains most of the visual cells, to appear sharp. Objects that are not fixated appear blurred on the retina. Nevertheless, it has been shown that blurring affects the way an observer will look at an image [1]. People will look at objects in focus, and neglect blurred background.

Visual attention allows the human visual system to understand complex scenes by successively focusing on interesting features or objects. Saliency models [2, 3] aim at predicting where an observer will look. They are often based on a frequency analysis of images and extraction of features that contrast with their surroundings. [4, 5] experimentally determine functions where contrast sensitivity is measured for an average observer. Those functions perform as band-pass filtering for the contrast identification. Thus visual attention modeling already introduces frequency selection that can be seen as blur differentiation. However, in [6] the authors refine previous assumptions by demonstrating that the

most attractive signal appears in medium frequencies. Consequently, blurred areas can still be salient. An interpretation of this is that the notion of saliency and sharpness might not be processed at the same level in the human brain, sharpness being identified in a second step. Indeed, because of the physiology of the eye, and in particular the small size of the fovea, the observer does not know if the object attended after next saccade is sharp or blurred on the screen. Bottom-up saliency models do not currently detect blurred areas and will mark out-of-focus objects as Region of Interest (RoI). Integrating blur detection to refine the bottom-up saliency map and remove out-of-focus objects from RoI is thus biologically plausible. An attempt of using blur detection to enhance saliency models can be found in [1]. The authors use different state-of-the-art saliency models and a blur identification algorithm based on an edge map, combined with machine learning. They do not provide validation using ground truth eye-tracking data. Besides, they present results for still images only.

This paper proposes two contributions: (i) a new saliency model that integrates blur identification in order to more precisely detect RoI in images and videos and (ii) a validation of such prediction improvement via eye-tracking on a dedicated database with blurred images. One can notice that the blur detection algorithm uses the same frequency analysis stage as the saliency model and thus introduces very little computation cost overhead. The remainder of the paper is organized as follows: section 2 describes the saliency model and the blur identification algorithm, then results are presented in section 3. Finally, conclusions and perspectives are drawn in section 4

2 Visual attention model

The saliency model used is a spatio-temporal model of the bottom-up selective visual attention, derived from [7]. In this paper, we present an improvement of an existing spatial saliency model thanks to the identification of blur.

2.1 Bottom-up spatial saliency model

The spatial saliency model is described in [6]. It is based on the plausible neural architecture of Koch and Ullman [8] and designed to be simple and computationally efficient. Its performances are similar or even higher than state-of-the-art models in terms of prediction.

The visual attention model uses a hierarchical decomposition of the visual signal. Its synoptic is described in Figure 1. The YUV 4:2:0 color space is used. It separates achromatic (Y) and chromatic (U: green-magenta and V: orange-cyan) perceptual signals. This color space has been chosen because it takes the human visual system into consideration and is commonly used in image and video processing.

The first step of the model extracts early visual features from the image. A 9/7 Cohen-Daubechies-Feauveau (CDF) wavelet transform is used to separate frequency bands and orientation ranges. The resulting multi-scale pyramid is

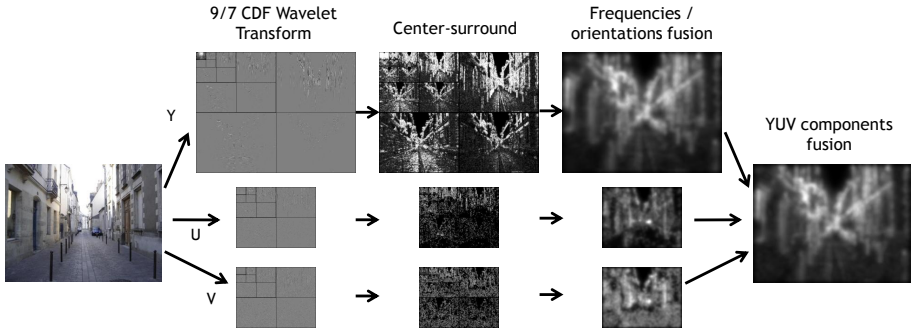


Fig. 1. Bottom-up saliency model

composed of oriented contrast maps with limited frequency range and a low-resolution image.

In the second step of the model, a Difference of Gaussian (DoG) modeling the center-surround response of visual cells is applied on each oriented contrast map (wavelet sub-band). Next, the orientation maps from each level are summed together.

The final step is the fusion of these early feature maps. Two fusions are successively applied: levels fusion and color channels fusion. Levels fusion operation is an across-scale addition using successive bi-linear up-sampling and additions of the per-pixel level maps. YUV components fusion keeps the maximum saliency value between Y, U, and V for each pixel after normalizing with an empirical maximum value taking into account the difference of amplitude between the three channels. The output map is finally mapped in the range of 0 to 1.

2.2 Blur identification

The implemented blur identification method is a modified version of the method proposed by Tong *et al.* [9] which uses the ability of wavelet transform in both discriminating different types of edges and identifying sharpness from blur. In our implementation, the same wavelet decomposition as the saliency model is used in order to avoid additional computations. Moreover, the CDF wavelet transform leads to a more precise frequency analysis than the Harr wavelets used in [9]. From this decomposition, block-based blur values are computed, leading to a map discriminating blurred from sharp areas (see Figure 2). For each decomposition level, an edge map is computed as:

$$E_{i,l} = \max_{k \in D_i} \left(\sqrt{LH_{k,l}^2 + HL_{k,l}^2 + HH_{k,l}^2} \right), \tag{1}$$

l being the current wavelet level (with the highest resolution level denoted $l = 0$), i the current pixel, LH, HL and HH the wavelet sub-bands and D the squared

non-overlapping neighborhood such that it corresponds to a 2×2 block in the smallest resolution level, 4×4 in the next, ... Then the final block-based blur value is defined as:

$$\begin{cases} 0\,(= blurred) & if \min\limits_{l \in [0, L-1]} E_{i,l} < Eblur \\ \min\left(255, 4 \times \sqrt{\sum\limits_{l \in [0, L-1]} \left(\frac{E_{i,l}^2}{2^l}\right)}\right) & otherwise \end{cases} \tag{2}$$

L is the number of decomposition levels, and $Eblur$ an experimental threshold set to 5 in our implementation. The blur map values are then mapped in the range of 0 (respectively blurred) to 1 (resp. sharp).

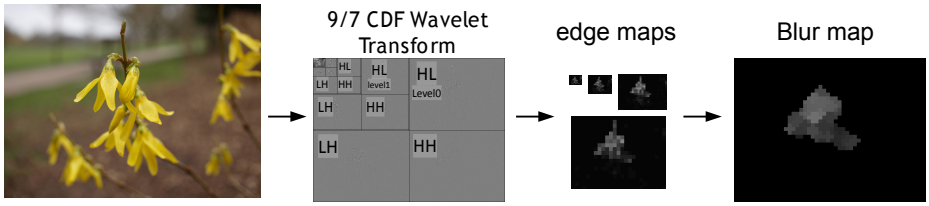


Fig. 2. Wavelet sub-bands decomposition

Blur identification is used to improve RoI detection by removing saliency in blurred areas, as these are considered to be visually unattractive areas. Thus, saliency values of blurred areas need to be reduced, while sharp objects' saliency need to be emphasized. The final spatial saliency map is defined as:

$$Saliency_{blur} = Saliency \times (.5 + Blur) \tag{3}$$

where $Saliency$ is the saliency map as described in section 2.1 and $Blur$ is the blur map as defined in section 2.2. With such a fusion, blur identification has an impact on the final saliency map while the blur map is not totally trusted.

2.3 Spatio-temporal saliency model

The temporal model assumes the visual attention is attracted by motion contrast. Such contrast is deduced from the difference between local and global motion. A fast hierarchical block-based motion estimator [10] is used to compute local motion. Then, using a weighted least square optimization approach, the global motion is deduced from block-based vectors with a parametric model. The temporal saliency map highlights blocks that have a different motion relative to the global motion.

The spatio-temporal saliency map is computed as the average of the spatial and the temporal saliency maps. For still images, the saliency map is directly

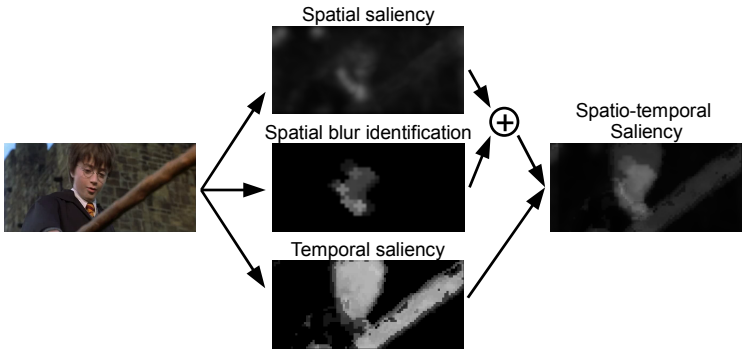


Fig. 3. Spatio-temporal saliency model

the output of the spatial saliency model as there is no temporal analysis. The global synopsis of the spatio-temporal attention model is depicted in figure 3.

The next section presents results of the application of this algorithm to still images and videos, and its validation on eye-tracking data.

3 Results

The presented model has been confronted to a ground truth of eye tracking data with still images and video contents. Saliency maps and fixation positions have been compared using both the NSS (Normalized Scan-path Saliency) metric [11] and the AUC. The AUC (Area Under Curve) is a classification indicator stemming from the ROC (Receiver Operating Characteristic) analysis. An AUC value tending to 1 means a good agreement between the predicted and the experimental saliency maps. The AUC results highly depend on the smoothness of the computed saliency map and the chosen threshold used to compute the binary ground truth. The NSS has the advantage to normalize the salience per scanpath: scanpaths with different number of fixations have the same weight. In other words, every observer has the same impact on salience. Moreover, the NSS gives more weight to areas more often fixated. We use both metrics anyway to show that the results are not metric-biased.

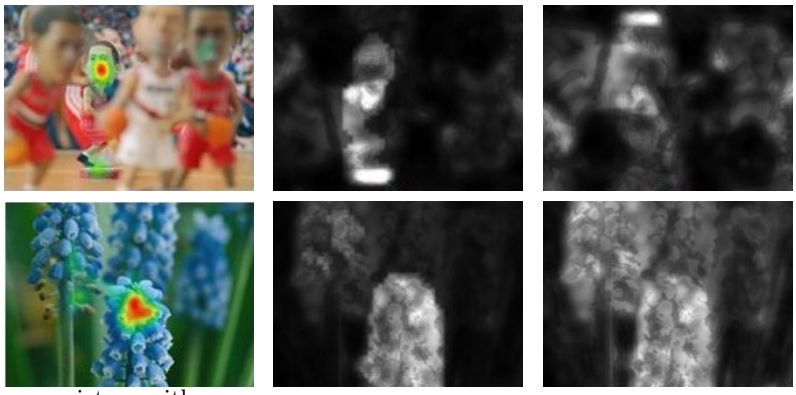
Eye-movements of observers were recorded in free-viewing conditions using an SMI RED 50 IView X eye-tracker with a 50Hz sampling frequency. Two different experiments were conducted, one with still images and one with videos, in order to validate the model in both configurations. Details are given below.

3.1 Still image database

25 volunteer subjects (11 females and 14 males) viewed 50 images[1] with a reduced depth of field thus naturally containing blur. All the subjects had normal or

[1] Part of the photos were by courtesy of Nicolas Le Goff (http://dishio.eu). All rights reserved.

corrected-to-normal vision. They were all naive to the purpose of the experiment. Each image was presented during 5 seconds, in a random order, interleaved with a neutral gray image containing a randomly placed black cross to reduce the center bias of the first fixation. The resolution of the images was 800x600 pixels. They have been manually selected on the internet from various topics. Figure 4 shows example images with fixation positions overlayed as heat map and their corresponding saliency map computed with and without blur detection.



picture with superimposed fixation heat map | spatial saliency map with blur detection | spatial saliency map

Fig. 4. Example image stimuli with corresponding fixation heat map and saliency maps

The proposed model improves fixation prediction. An important improvement is seen on highly blurred pictures where the model without blur detection gave too much importance to blurred areas. On pictures containing smooth or homogeneous areas, the improvement is less visible because such areas are detected as blurred. Table 1 presents average NSS and AUC values of the 50 images, for the model without blur identification and the augmented model. Blur identification significantly improves the prediction performances (paired t-test with $p<0.001$).

	Average NSS values			Average AUC values		
	no blur	with blur	paired t-test	no blur	with blur	paired t-test
image database	0,83	0,99	4,4E-08 (***)	0.72	0.75	8.3E-08(***)
video database	0,98	1,04	0,005 (**)	0.70	0.71	0.001(**)

**t significant at p<.01 ***t significant at p<.001

Table 1. Average NSS and AUC values comparison of saliency model without and with blur identification for image and video databases

3.2 Video database

The efficiency of the model has also been tested on 21 videos and compared with gaze positions recorded from 30 volunteer subjects during free-viewing task. Selected videos were mainly extracted from movie trailers. They did not contain highly reduced depth-of-field content such as for the experiment with still pictures. The proposed saliency model with blur detection significantly improves fixation prediction in terms of NSS and AUC (paired t-test with $p<0.01$) (see Table 1). The impact of blur identification on the video database is reduced compared to the still image database because videos contain less blurred areas than selected pictures. Moreover more significant improvements are averaged in time with other frames where less improvement is obtained, thus reducing the measured improvement.

Blur detection is efficient on close-up scenes where the background is usually out-of-focus. It works also well on wide shots where homogeneous areas are detected as blurred leading to concentrate more salience on the RoI. Screenshots of selected videos with their corresponding fixations and saliency maps are presented on Figure 5.

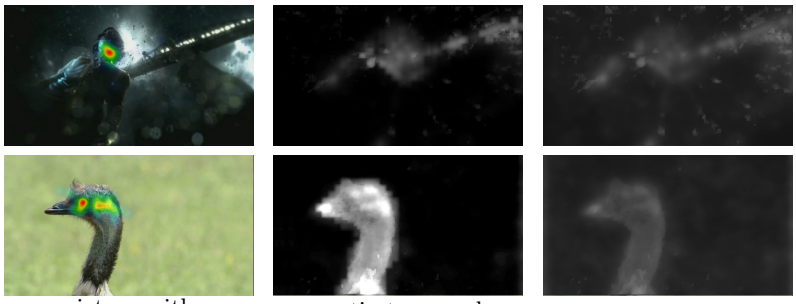


Fig. 5. Video examples with corresponding fixation heat map and spatio-temporal saliency maps

While blur identification improves fixation prediction on images or videos where blurred areas are detected correctly, one must be careful with homogeneous areas. Indeed, those regions have no high frequencies, being detected as blurred areas, but they might be part of RoI where fixations are located (i.e. interior of an object). Detection of homogeneous areas associated with a special treatment is thus necessary to prevent erroneous results.

In this paper, the detection of blur is performed under the assumption of focal blur. Motion blur generates blur in the direction of the motion. This type of blur may not be detected by our blur identification method because it affects only one

direction of the wavelet transform coefficients. A motion blur detection taking advantage of already computed wavelet coefficients and motion information in the temporal branch may thus be of interest in a fixation prediction context.

4 Conclusions

In this paper, we presented a new saliency model that uses blur detection to improve fixation predictions. Low-level saliency models can detect blurred areas as salient. This is consistent with the low-level models of the human visual system, but blurred areas are rarely of interest. Identifying sharp from blurred areas and lowering the saliency value on blurred areas improves saliency model performances for images and videos. Eye-tracking experiments have been conducted to create a ground truth of human fixations on images and videos naturally containing blur. A significant improvement has been obtained with the proposed algorithm. As expected, the biggest improvements are obtained for content with a small depth-of-field, where the background is highly blurred. Currently, the blur identification algorithm assumes focal blur only and does not detect motion blur. Further improvement of performances could be achieved with motion blur detection.

References

1. R. A. Khan, H. Konik, and E. Dinet, "Enhanced image saliency model based on blur identification," in *25th International Conference of IVCNZ*, Dec. 2010.
2. O. Le Meur, P. Le Callet, and D. Barba, "Predicting visual fixations on video based on low-level visual features," *Vision Res*, vol. 47, no. 19, pp. 2483–98, 2007.
3. L. Itti, C. Koch, and E. Niebur, "Model of saliency-based visual attention for rapid scene analysis," *IEEE PAMI*, vol. 20, no. 11, pp. 1254–1259, 1998.
4. S. Daly, "The visible differences predictor: An algorithm of image fidelity," *Digital Images and Human Vision*, pp. 179–206, 1993.
5. J. L. Mannos and D. J. Sakrison, "The effects of a visual fidelity criterion on the encoding of images," *IEEE TIT*, vol. 20, no. 4, pp. 525–536, 1974.
6. F. Urban, B. Follet, C. Chamaret, O. Le Meur, and T. Baccino, "Medium Spatial Frequencies, a Strong Predictor of Salience," *Cogn Comput*, vol. 3, pp. 37–47, 2011.
7. O. Le Meur, D. Thoreau, P. Le Callet, and D. Barba, "A spatio-temporal model of the selective human visual attention," *IEEE ICIP*, vol. 3, pp. 1188–1191, 2005.
8. C. Koch and S. Ullman, "Shifts in selective visual attention: towards the underlying neural circuitry," *Hum Neurobiol*, vol. 4, no. 4, pp. 219–27, 1985.
9. H. Tong, "Blur detection for digital images using wavelet transform," in *Proceedings of IEEE ICME*, 2004, pp. 17–20.
10. Fabrice Urban, Jean François Nezan, and Mickael Raulet, "HDS, a real-time multi-DSP motion estimator for MPEG-4 H.264 AVC high definition video encoding," *Springer Journal of real-Time Image Processing*, vol. 4, no. 1, pp. 23–31, 2009.
11. R. Peters, A. Iyer, L. Itti, and C. Koch, "Components of bottom-up gaze allocation in natural images," *Vision Research*, vol. 45, pp. 2397–2416, 2005.