Color Information in a Model of Saliency

Shahrbanoo Hamel, Nathalie Guyader, Denis Pellerin, Dominique Houzet

COLOR INFORMATION IN A MODEL OF SALIENCY

Shahrbanoo Hamel, Nathalie Guyader, Denis Pellerin and Dominique Houzet

GIPSA-lab, UMR 5216, Grenoble, France

Abstract

Bottom-up saliency models have been developed to predict the location of gaze according to the low level features of visual scenes, such as intensity, color, frequency and motion. We investigate in this paper the contribution of color features in computing the bottom-up saliency. We incorporated a chrominance pathway to a luminance-based model (Marat et al. [1]). We evaluated the performance of the model with and without chrominance pathway. We added an efficient multi-GPU implementation of the chrominance pathway to the parallel implementation of the luminance-based model proposed by Rahman et al. [2], preserving real time solution. Results show that color information improves the performance of the saliency model in predicting eye positions.

Index Terms— color information, visual saliency, video, GPU

1. INTRODUCTION

When exploring our visual environment we frequently move our eyes to select and focus on specific regions of the scene. The selection is driven by the properties of the visual stimulus through bottom-up processes, as well as by the goal of observer through top-down processes [3], [4]. Visual attention models tend to predict the parts of the scene that are likely to guide visual attention and hence, the gaze of observers [5], [6], [7], [1]. Most of the models are bottom-up models based on the Feature Integration and Guided Search theories [8], [9]. These theories stipulate that some elementary salient visual features such as intensity, color, depth and motion, are processed in parallel at a pre-attentive stage, subsequently combined to drive the focus of attention. This approach is in accordance with the physiology of the visual system. Hence, in almost all the models of visual attention, low level features like intensity, color, spatial frequency are considered to determine the visual saliency of regions in static images, whereas motion and flicker are also considered in the case of dynamic scenes [5], [7], [1]. The contribution of different features like color in guiding eye movements when viewing natural scenes has been debated. Some studies suggested that color has little effect on fixating locations [10], [11], [12], which brings to question the necessity of the inclusion of color features in the saliency models [13].

On the other hand, most of biologically-inspired saliency models such as Marat et al. [1] involve many computationally intensive operations. Hence, real-time solutions seem impossible or only achievable by the simplification of whole pathway, as proposed by Itti [4]. Yet, the evolution of computer graphics hardware and programming models like CUDA [14] makes the graphics devices a suitable platform to implement the real-time solutions for visual attention algorithms.

In this article, we investigate the contribution of color information in predictive power of saliency model by incorporating a chrominance pathway to the luminance-based model of saliency proposed by Marat et al. [1]. We evaluated the performance of the model on two datasets of eye movements: a public dataset for saliency available on CRCNS (Collaborative Research in Computational Neuroscience) [15], and a new dataset that we provided through an eye-tracking experiment using new video stimuli [16]. Then we include a parallel adaptation of the chrominance-pathway onto the GPU implementation of the luminance-based model proposed by Rahman et al. [2].

2. METHOD

2.1. Saliency model

The luminance-based saliency model of Marat et al. [1] draws inspiration from human visual system. The model consists of two pathways: static and dynamic. Both pathways are only based on luminance information of visual scene, processed in two steps: the first step simulates some basic pre-processing done by the retina cells through a cascade of three linear filters: a band pass filter for luminance pre-processing and two low pass filters for chrominance. The retina separates the input signal into low and high spatial frequencies that schematically represent the magno- and parvo- cellular outputs of the retina. At second step each signal is decomposed into elementary features by a bank of cortical-like filters. These filters, according to their frequency selectivity, orientation and motion amplitude, provide two saliency maps: static map M_{ls} and dynamic map M_{ld}, Figure 1.

This research was supported by Rhone-Alpes region (France) under the CIBLE project No. 2136. Thanks to D. Alleysson and D. Meary for providing us with spectrometer measurements.

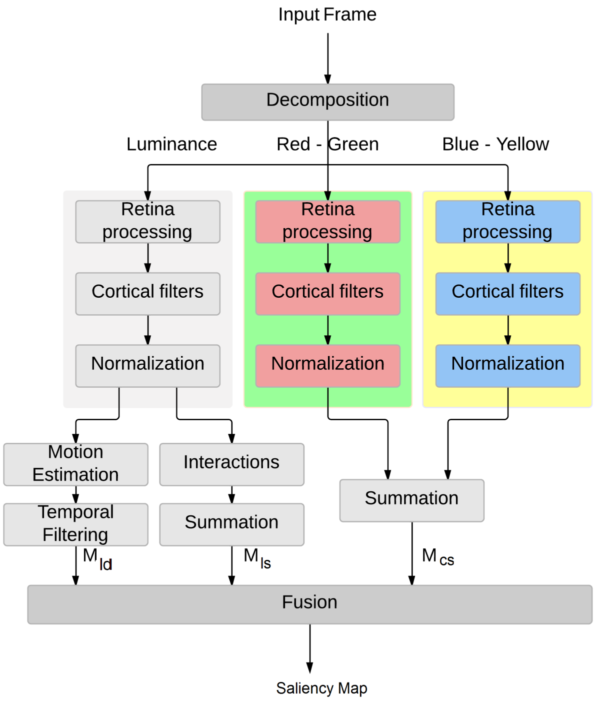

Fig. 1: The spatio-temporal saliency model.

The model proposed by Marat et al. is only based on the luminance information. We incorporated the color information to the model. According to psychological and physiological experiments the early transformation of the Long, Medium and Short wavelength signals, absorbed by cones, provides an opponent-color space in which signals are less correlated.

There are several color spaces proposing different combination of cone responses to define the principal components of luminance and opponent colors, red-green (RG) as well as blue-yellow (BY) [17]. One space might be preferred to another according to application needs. Here, we selected Krauskpof et al. [18] color space to encode color information, because it has been validated for encoding images that are observed in *I.T.U* recommended conditions [19]. This color space was also used in the saliency model proposed by Le Meur et al. [7]. A color image is encoded into orthogonal directions, A, Cr1 and Cr2, which represent luminance, chromatic opponent red-green and chromatic opponent yellow-blue respectively. The following equation is used to compute A, Cr1 and Cr2.

$$\begin{pmatrix} A \\ Cr1 \\ Cr2 \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 \\ 1 & -1 & 0 \\ -0.5 & -0.5 & 1 \end{pmatrix} \begin{pmatrix} L \\ M \\ S \end{pmatrix}$$

L, M and S signals of the three cone cells are obtained from the *CIE XYZ* components according to the following transformation:

$$\begin{pmatrix} L \\ M \\ S \end{pmatrix} = \begin{pmatrix} 0.4002 & 0.7076 & -0.0808 \\ -0.2263 & 1.1653 & 0.0457 \\ 0 & 0 & 0.9182 \end{pmatrix} \begin{pmatrix} X \\ Y \\ Z \end{pmatrix}$$

The human visual system is sensitive to the high spatial frequencies of luminance and the low spatial frequencies of chrominance [20], [21]. The amplitude spectra of the two color-opponent Cr1 and Cr2 images do not have as many specific orientations as the amplitude spectra of the luminance image [22]. Hence the retinal and cortical processing of chrominance information is different from luminance information. We integrated to the Marat et al. [1] spatio-temporal saliency model, the chrominance processing steps first introduced by Ho-Phuoc et al. [23]. The retinal processing step of chrominance information starts with low pass filtering illustrated by the contrast sensitivity functions (CSFs) for chrominance information [7]. Following these CSFs, the two color opponents are processed by two low-pass filters. Then the cortical like filters extract the spatial information of Cr1 and Cr2 color opponents according to 4 orientations $(0, 45, 90,$ and 135 degrees) and 2 spatial frequencies($\frac{1}{16}, \frac{1}{32}$ 1/pixel), providing a chrominance-based static saliency map M_{cs}. Luminance-based static saliency map M_{ls}, luminance-based dynamic saliency map, M_{ld} and chrominance saliency map M_{cs}, after normalizing, are combined, according to the following equation, to obtain a master spatio-temporal saliency map per video frame. This map predicts the salient regions i.e. the regions that stand out in a visual scene.

$$Saliency\ \ map = \alpha M_{ls} + \beta M_{ld} + M_{cs} + \alpha\beta(M_{ls} \times M_{ld})$$

Where, α and β are the max of M_{ls} and skewness of M_{ld} respectively. Figure 3, extracted from [16], shows an example frame and its intermediate and final saliency maps.

In addition, we compared the performance of the model with one of the reference saliency models, Itti and Koch saliency model [24], [5]. The performance of Itti and Koch saliency model was also evaluated once using all features with equal weights (intensity, orientation, contrast, flicker and motion and DKL color features), and once using all features except color.

GPU implementation. The saliency model presented above with luminance-based static, chrominance-based static and luminance-based dynamic pathways is compute-intensive. Rahman et al. [2] have proposed a parallel adaptation of two luminance-based pathways onto GPU. The sequential code has been decomposed into two sub-programs: one running on the host (the CPU) and kernel code running on the device (the GPU). The host code is responsible for all data read/write operations from disk and data copy operations to/from the device, while the device codes perform the tasks in parallel. The kernel code is compiled by the nvcc compiler supplied

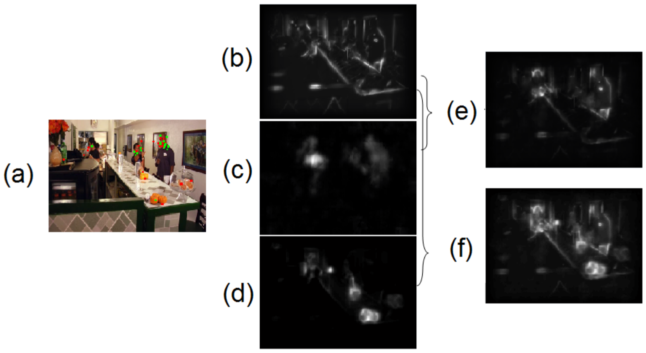

Fig. 2: An example frame and its corresponding saliency maps,(a) Original frame (b) luminance saliency map M_{ls}, (c) dynamic saliency map M_{ld}, (d) chrominance M_{cs}, (e) fusion of M_{ls} and M_{ld}, (f) fusion of M_{ls}, M_{cs} and M_{ld}.

by NVIDIA. We included the parallel adaptation of chrominance pathway to this GPU implementation maintaining the real time solution. Algorithm 1 shows the sequentially executed kernels providing chrominance saliency map for each of opponent-color components.

```
Input: An image Im of size $w \times l$
Output: A saliency map
map ← $RetinalFilter$ (Im) ;
map ← $FFT$ (map) ;
for $i \leftarrow 1$ to $orientations$ do
    for $j \leftarrow 1$ to $frequencies$ do
        maps[i,j] ← $GaborFilter$ (map,i,j) ; maps[i,j]
        ← $IFFT$ (map,i,j) ; maps[i,j]
        ← $Normalizations$ (map,i,j) ;
    end
end
saliency ← $Summation$ (maps) ;
```

Algorithm 1: Chrominance pathway for each of the opponent-color images, Cr1 and Cr2.

The chrominance pathway includes the Retina filter with low-pass filters using 2D convolutions and recursive Gaussian filters, normalizations with reduction operations, some simple matrix operations and Fourier transforms. The NVIDIA CUDA fast Fourier transform library (cuFFT) was used to perform the complex Fourier transformations. The reductions use Thrust library, an interface to many GPU algorithms and data structures. Such as the luminance-based static and dynamic pathways implementations, chrominance pathway was tested on a 2.67 GHz quad-core system with 10 GB of main memory, and Windows 7 running on it. CUDA v3.0 programming environment on NVIDIA Geforce GTX 480 was used. The chrominace pathway was evaluated with image sizes of 640×480.

NSS metric. A common metric to compare experimental data to computational saliency maps is the Normalized Scanpath Saliency (*NSS*) [25]. We used this metric to compare C and GS eye positions to their equivalent saliency maps. To compute this, first the saliency maps were normalized to zero mean and unit standard deviation. The NSS value of frame k corresponds to averaged saliency values at the locations of eye positions on the normalized saliency map M_{ls} as shown in the following equation:

$$NSS(k) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{\sigma_k}(M_{ls}(X_i) - \mu_k)$$

where N is the number of the eye positions, μ_k and σ_k are the mean and standard deviation of the initial saliency map of frame k. A high positive value of NSS indicates that the eye positions are located on the salient regions of the computational saliency map. A NSS value close to zero represents no relation between eye position and the computational saliency map, while a high negative value of NSS means that eye positions were not located on the salient regions of computational saliency map.

2.2. Eye-tracking experiment

We used two datasets of eye movements to evaluate the performance of the saliency model: eye movement data from *CRCNS* public dataset of *MTV experiment* [15], and eye movement data obtained from an eye-tracking experiment that we carried out for evaluating the model, called person-present stimuli experiment (*PPS experiment*). The dataset of *MTV experiment* comprises eye movement recordings from 8 distinct subjects watching 50 different video clips (460×640 pixels, 33.185 ms/movie frame). The subjects were asked to follow main actors and actions, and to try to understand overall what happens in each clip [26]. In this experiment only color stimuli was employed.

In *PPS experiment* we recorded eye movement data of 45 volunteers (25 women and 20 men, aged from 25 to 39 years old) while freely viewing videos in two conditions: color and grayscale. An Eyelink 1000 from SR research was used to record the eye positions in a pupil tracking mode. The stimuli consisted of 65 short video extracts of 3 to 5 seconds, called video snippets. Video snippets were extracted from various open source color videos. The stimuli measured 640×480 pixels, subtending a visual angle of 25×19 degrees at a fixed viewing distance of 57 cm. The temporal resolution of video snippets was 25 frames per second. The video stimuli was also presented in grayscale condition to compare the luminance-based model with the eye positions recorded in this condition. The video dataset was converted to grayscale according to the equation 1.

$$L = 0.5010 \times R + 0.4911 \times G + 0.0079 \times B \qquad (1)$$

In most of studies investigating the influence of color information on eye movements, the grayscale version of stimuli is obtained using NTSC conversion, which is based on the weighted sums of R, G and B channels [11, 12]. However such method is not adapted for specific experimental set up. Here, to ensure the luminance matching between color and grayscale video stimuli, the weights of R, G and B channels in equation 1 were found to fit $V(\lambda)$ of standard observer. The equation was found by directly measuring the luminance reflectance of the three color channels of the experimental set up using a Photo Research PR650 spectrometer. Figure 3 depicts example frames of color and grayscale video snippets.

Fig. 3: Example frames in color and grayscale.

Table 1: *NSS* results for *Marat et al.* saliency model and *Itti and Koch* saliency model and without color features for *PPS* experiment.

	stimulus condition	**luminance**	**luminance + chrominance**
Marat et al.	Color	0.59	1.18
	Grayscale	0.60	1.17
Itti and Koch	Color	0.91	0.95
	Grayscale	0.93	0.97

Table 2: NSS results for Marat et al. luminace-based and luminance-chrominance models on *MTV experiment* dataset for saliency.

	luminance	**luminance +chrominance**
Marat et al.	0.65	0.90

3. RESULTS

3.1. Saliency model

First, we evaluated the performance of *luminance-based saliency model* [1] in predicting the eye positions for both stimulus conditions using *NSS* metric. Then we compared the performance of *luminance-based* and *luminance-chrominance* saliency models. As shown in table 1 chrominance pathway improves significantly the performance of the model for color and grayscale stimuli $(Grayscale : t(63) = 4.5, p < 0.01, Color : t(63) = 4.86, p < 0.01)$, while including color features improves slightly the performance of the model of Itti and Koch. [5]. We also compared the performance of luminance-based model with the performance of proposed luminance-chrominace based model on public dataset for saliency from *MTV experiment*. As shown in table 2, for this dataset also the performance of chrominance-luminance model is significantly higher than luminance-based model ($t(49) = 6, p < 0.001$).

3.2. GPU real-time solution

Table 3 shows the execution times of luminance-based pathways for parallel solution proposed by Rahman et al. [2] and chrominance pathway proposed in this article.

Table 3: Timings of sequential (C and Matlab) and parallel (GPU) implementations in ms.

	M_{sl}	M_{cl}	M_{dl}
MATLAB	34.01	22.67	237.03
C	10.73	7.15	31.24
CUDA	0.04	0.03	0.12

The GPU implementation for the three pathways results to significant speed-up over matlab and C that provides a real-time solution.

3.3. Conclusion

In the present manuscript, we have integrated color information into our bio-inspired saliency model. We added the parallel GPU implementation of chrominace pathway to the existing luminance-based pathways. The real-time solution enables the model to be used in many applications such as robotics vision, image analysis and compression. Results show that including a chrominance pathway to the model, improves its predictive power significantly on both *PPS* and *MTV* datasets. Yet, the performance of model on videos stimuli of *PPS* experiment is higher than video stimuli of *MTV* experiment. This might be due to a principal difference between contents of *MTV* stimuli and *PPS*. In *PPS* experiment, most of the video clips include scenes with presence of close-up faces (51 person-present scenes versus 14 person-absent scenes), which are one of the high level stimulus properties that have a significant impact on the gaze location and increases the congruency between observers [27].

REFERENCES

[1] S. Marat, T. Ho Phuoc, L. Granjon, N. Guyader, D. Pellerin, and A. Gurin-Dugu, "Modelling spatio-temporal saliency to predict gaze direction for short videos," *International Journal of Computer Vision*, vol. 82(3), pp. 231–243, 2009.

[2] A. Rahman, D. Houzet, D. Pellerin, S. Marat, and N. Guyader, "Parallel implementation of a spatio-temporal visual saliency model," *Real-Time Image Processing*, vol. 6(1), pp. 3–14, 2010. http://www.gipsa-lab.fr/projet/perception/.

[3] C. E. Connor, H. E. Egeth, and S. Yantis, "Visual attention: bottom-up versus top-down," *Current Biology*, vol. 14, pp. 850–852, 2004.

[4] L. Itti, "Quantifying the contribution of low-level saliency to human eye movements in dynamic scenes.," *Visual Cognition*, vol. 12, pp. 1093 –1123, 2005.

[5] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis.," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, pp. 1254–1259, 1998.

[6] S. Frintrop, *VOCUS: A Visual Attention System for Object Detection and Goal-Directed Search*, Ph.D. thesis, Rheinische Friedrich-Wilhelms-Universität Für Informatik and Fraunhofer Institut Für Autonome Intelliegente Systeme, 2006.

[7] O. Le Meur, P. Le Callet, and D. Barba, "Predicting visual fixations on video based on low-level visual features.," *Vision Research*, vol. 47(19), pp. 2483–2498, 2007.

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

[9] J. M. Wolfe, K. R Cave, and S. L. Franzel, "Guided search: An alternative to the feature integration model for visual search.," *Journal of Experimental Psychology: Human Perception & Performance*, vol. 15, pp. 419–433, 1989.

[10] R. J. Baddeley and B. W. Tatler, "High frequency edges (but not contrast) predict where we fixate: A bayesian system identification analysis," *Vision Research*, vol. 46, no. 18, pp. 2824–2833, 2006.

[11] T. Ho-Phuoc, N. Guyader, and A. Guérin-Dugué, "When viewing natural scenes, do abnormal colors impact on spatial or temporal parameters of eye movements?," *Journal of Vision*, vol. 12(2), pp. 1–13, 2012.

[12] H. P. Frey, C. Honey, and P. Knig, "Whats color got to do with it? the influence of color on visual attention in different categories," *Journal of Vision*, vol. 11(3), pp. 1–15, 2008.

[13] M. Dorr, T. Martinetz, K. Gegenfurtner, and E. Barth, "Variability of eye movements when viewing dynamic natural scenes," *Journal of Vision*, vol. 10, no. 10, pp. 1–17, 2010.

[14] "Nvidia cuda compute unified device architecture - programming guide," 2007.

[15] L. Itti and R. Carmi, "Eye-tracking data from human volunteers watching complex video stimuli. crcns.org. 2009," http://dx.doi.org/10.6080/K0TD9V7F.

[16] S. Hamel, N. Guyader, D. Pellerin, and D. Houzet, "Contribution of color in visual saliency model for videos," in *Sixth International conference on Image and Signal Processing (ICISP 2014)*, LNCS Springer, Ed., 2014, vol. 8509, pp. 213–221.

[17] A. Trémeau, Ch. Fernandez-Maloigne, and P. Bonton, *Image numérique couleur, de l'acquisition au traitement*, Dunod, 2004.

[18] J. Krauskopf, D. R. Williams, and D. W. Heeley, "Cardinal direction of color space," *Vision Research*, vol. 22, pp. 1123–1131, 1982.

[19] L. Bedat, *Aspects psychovisuels de la perception des couleurs. Application au codage d'images couleurs fixes avec compression de l'information.*, Ph.D. thesis, Université de Nantes, 1998.

[20] D. J. Field, "Relations between the statistics of natural images and the response properties of cortical cells," *J. Opt. Soc. Am. A*, vol. 4, pp. 2379–2394, 1987.

[21] K. R. Gegenfurtner, "Cortical mechanisms of colour vision.," *Nature Reviews Neuroscience*, vol. 4(7), pp. 563–72, 2003.

[22] W. H. A. Beaudot and Mullen. K. T., "Orientation selectivity in luminance and color vision assessed using 2-d bandpass filtered spatial noise.," *Vision Research*, vol. 45(6), pp. 687–696, 2005.

[23] T. Ho-Phuoc, N. Guyader, and A. Guérin-Dugué, "A functional and statistical bottom-up saliency model to reveal the relative contributions of low-level visual guiding factors," *Cognitive Computation*, vol. 2(4), pp. 344–359, 2010.

[24] J. Harel, "A saliency implementation in matlab:," http://www.klab.caltech.edu/ harel/share/gbvs.php.

[25] R.J. Peters, A. Iyer, L. Itti, and C. Koch, "Components of bottom-up gaze allocation in natural images.," *Vision Research*, vol. 45, no. 18, pp. 2397–2416, 2005.

[26] R. Carmi and L. Itti, "Visual causes versus correlates of attentional selection in dynamic scenes.," *Vision Research*, vol. 46, pp. 4333–4345, 2006.

[27] A. Rahman, D. Pellerin, and D. Houzet, "Influence of number, location and size of faces on gaze in video," *Journal of Eye Movement Research*, vol. 7(2), pp. 1–11, 2014.