Scalable Visual Sensitivity Profile Estimation

Guangtao Zhai*, Qian Chen*, Xiaokang Yang*†, and Wenjun Zhang*

*Institute of Image Communication and Information Processing
Shanghai Jiao Tong University, Shanghai, 200240, China
†Institute for Computer Science, University of Freiburg, Freiburg, Germany

***Abstract*—We propose a computational model for estimating scalable visual sensitivity profile (SVSP) of video, which is a hierarchy of saliency maps that simulates the bottom-up and top-down attention of the human visual system (HVS). The bottom-up process considers low level stimulus-driven visual features such as intensity, color, orientation and motion. The top-down process simulates the high level task-driven cognitive features such as finding human faces and captions in the video. The nonlinear addition model has been used for integrating low level visual features. A full center-surrounded receptive field profile is introduced to provide spatial scalability of the model. Due to the hierarchical nature, the proposed SVSP can be directly used to augment the visual quality of codings with spatial scalability. To justify the effectiveness of the proposed SVSP, extended experiments of its application in visual quality assessment are conducted.**

***Index Terms*—Human visual system, perceptual quality assessment, scalable video coding,**

I. Introduction

Visual attention is one of the most important mechanisms of the human visual system (HVS). Most of the successful computational visual attention models are based upon Treisman's pioneering work on visual attention [1], which divides the process into the pre-attentive and attentive stages. The pre-attentive stage, also referred as bottom-up attention stage, extracts low-level visual features such as intensity, color, orientation and movement, and integrates them into a saliency map. The attentive stage, or the top-down stage, involves much more complex psychological process, and directs attention into certain objects within the scene. The pre-attentive and attentive stages are also known as stimulus-driven and knowledge-driven attentions, respectively. Koch and Ullman's model [12] firstly proposed the biologically-plausible computational steps in bottom-up attention simulating such central representation and winner-take-all networks. Their work largely motivated Itti's model [2] [16] and the most recent Oliver *et al.*'s model [13]. Compared with bottom-up attention, the investigation on top-down attention lays its emphasis on the top-down modulation processing on bottom-up features of visual attention [2] [14]. Commonly these models lack explicit expression. One exception is Lu *et al.*'s PQSM [15], which takes human face detection as an evidence of top-down attention.

This work was supported by National Natural Science Foundation of China (60332030, 60502034, 60625103), Shanghai Rising-Star Program (05QMX1435), Hi-Tech Research and Development Program of China 863 (2006AA01Z124), NCET-06-0409, the 111 Project and the Alexander von Humboldt Foundation.

With a mathematical visual attention model, it is straightforward to generate a topographic saliency map of visual attention, which indicates the sensitivity level of every location in the input image. However, scalable video coding (SVC) is recently being developed to enable decoding from partial streams with respect to the specific rate and resolution required by a certain application. Therefore, it requires the computational visual attention model to offer more flexibilities, because it may need saliency maps under various spatial/temporal resolutions to facilitate the coding process instead of one with fixed resolution. In this paper, we propose a computational model for scalable visual sensitivity profile (SVSP), i.e., a hierarchy of saliency maps that simulate both the bottom-up and top-down attention of HVS. Due to the hierarchical structure, the proposed SVSP can be directly used to augment the visual quality of SVC applications. In addition, it can be applied to assess the visual quality of image and video, with improved accuracy benefiting from the systematic simulations of both pre-attentive and attentive features.

For the rest part of the paper, the framework of the proposed SVSP is introduced in Section II. The bottom-up and top-down attention models are detailed in Section III and IV, respectively. The results are integrated into SVSP in Section V, followed by the verification of SVSP in image quality assessment in Section VI. And finally Section VII concludes the paper.

II. The Computational Framework for SVSP

The diagram of the computation model of SVSP is shown in Fig.1, and the outputs of each step, using the example of *present debate*, are also illustrated. We take Itti's bottom-up attention model [2] as a reference and make some modifications towards more accurate prediction for specific applications. 1) We extend the receptive field profile computation into a full center-surrounded structure, so as to provide hierarchical saliency maps to be used in SVC. 2) Nothdurft's nonlinear addition model [3] is used to integrate low-level stimulus features instead of Itti's direct summation [2] to account for possible overlap between features map. For the top-down attention part, it is widely known that human face and captions in the picture often indicate useful recognition clues and attract knowledge-driven human attention. As a consequence, we apply face and caption as top-down attention directors. It is noted that other high level attentive features can also be easily added into the proposed framework. These top-down feature maps also take a hierarchical shape to be seamlessly

combined with the afore-computed bottom-up features. With the hierarchical bottom-up and top-down saliency maps, we can integrate them into a final SVSP.

III. Bottom-up Attention Model

A. Low-level Feature Detection

Itti's model operates in the RGB color space for expression and computation simplicity [2]. However, in digital image/video coding systems, the YCbCr color space is often more preferable, due to three reasons: 1) Y, Cb and Cr are uncorrelated; 2) Cb, Cr can be represented using lower bandwidth; and 3) Y can be extracted and used as luminance directly. So in this work, we adapt Itti's model into the YCbCr space. Let $f_1 \cdots f_n$ be n consecutive frames in a video sequence. In YCbCr system, for a frame f_i, Y is taken as the intensity channel, i.e.

$$ci_i = Y_i \tag{1}$$

This luminance component is more accurate than that of Itti's model, where an approximation $ci_i = (R_i + G_i + B_i)/3$ was used. For color componnet, Itti created four broadly tuned color channels of red, green, blue and yellow in RGB space [2], denoted as cr_i, cg_i, cb_i and cy_i, respectively. Considering the conversion matrix between RGB and YCbCr, the color channels in YCbCr space would be

$$cr_i = -0.813(Cb_i - 128) + 2.003(Cr_i - 128) \tag{2}$$

$$cg_i = -1.401(Cb_i - 128) - 1.661(Cr_i - 128) \tag{3}$$

$$cb_i = 2.213(Cb_i - 128) - 0.392(Cr_i - 128) \tag{4}$$

$$cy_i = 3.642(Cb_i - 128) + 0.783(Cr_i - 128) + 0.392(Cb_i - 128) + 2.409(Cr_i - 128) \tag{5}$$

The orientation channel $co_i$is obtained by filtering the intensity channel ci_i in four directions with Gabor filters ($GF(\theta)$)

$$co_i(\theta) = ci_i * GF(\theta), \theta \in \{0^o, 45^o, 90^o, 135^o\} \tag{6}$$

Motion is another important factor that modulates visual attention. For video sequences, we use Ogale's optical flow algorithm [4] to estimate absolute motion of image objects between consecutive frames. For three consecutive frames f_{i-1}, f_i, f_{i+1}, the horizontal and vertical motion channels of f_i are determined as the averaged directional optical flows between (f_{i-1}, f_i) and (f_i, f_{i+1})

$$cm_i^{\Theta} = [of^{\Theta}(f_i, f_{i+1}) + of^{\Theta}(f_{i-1}, F_I)]/2, \Theta \in (h, v) \tag{7}$$

The final motion channel is a combination of the horizontal and vertical motion

$$cm_i = [(cm_i^h)^2 + (cm_i^v)^2]^{1/2} \tag{8}$$

By iteratively dyadic down-sampling for L times of these channels, we can create pyramids for each of these channels of the frame f_i

$$\{ci_i(l), cr_i(l), cg_i(l), cb_i(l), cy_i(l), co_i(l), cm_i(l) | l \in (0, L)\} \tag{9}$$

where $l = 0$ indicates no down-sampling, i.e. the original image size. We use MPEG-4 down-sampling filter [8] to generate the pyramids instead of Itti's Gaussian low-pass filter [2] to be compliable with existing video coding standards. Hereinafter, we omit the subscription "i" in the expressions for brevity.

TABLE I
Qualitative Performance Analysis of Image Quality Assessment Models VC: Variance weighed correlation after nonlinear regression SC: Spearman rank order correlation

	VC			SC		
model	JPEG2000	JPEG	All	JPEG2000	JPEG	All
PSNR	0.8962	0.8596	0.8728	0.8898	0.8409	0.8646
SSIM	0.9367	0.9283	0.9295	0.9317	0.9028	0.9174
$PSNR^{VSP}$	0.9168	0.8688	0.8826	0.9161	0.8566	0.8776
$SSIM^{VSP}$	0.9478	0.9365	0.9401	0.9420	0.9136	0.9267

B. Center-surround Receptive Field Simulation

A full center-surround structure is implemented to simulate the receptive fields in the HVS. The center level c and surround level s are defined as: $c \in [0, 8]$, $s = c + \delta$, $\delta \in [-3, -2, -1, 1, 2, 3]$ and s is thrown away if $s \notin [0, 8]$. The full center-surround receptive field profiles are computed for each pair of the five feature components: intensity, red-green channel, blue-yellow channel, orientation, and motion, based on the computed pyramids.

$$I(c, s) = |ci(c) \circ ci(s)| \tag{10}$$

$$RG(c, s) = |(rc(c) - gc(c)) \circ (gc(s) - rc(s))| \tag{11}$$

$$BY(c, s) = |(bc(c) - yc(c)) \circ (yc(s) - bc(s))| \tag{12}$$

$$O(c, s, \theta) = |co(c, \theta) \circ co(s, \theta)| \tag{13}$$

$$M(c, s) = |cm(c) \circ cm(s)| \tag{14}$$

where the operator $|\circ|$ denotes to convert the size of the surround level s to the center level c through up-sampling or sub-sampling, and then calculate the difference. For compliance with coding standards, AVC 6-tap up-sampling filter [9] and the MPEG-4 down-sampling filter [8] are used. Accordingly, the sizes of these profiles are the same as the corresponding center levels in the pyramids.

C. Non-linear Feature Combination

To combine the channel information to generate one single saliency profile on certain pyramid level, the contents-based global non-linear amplification is firstly used to normalize the profiles [2]. The processed profiles are

$$\bar{I}(c) = \sum_s Nor(I(c, s)) \tag{15}$$

$$\bar{C}(c) = \frac{\sum_s Nor(RG(c, s)) + \sum_s Nor(BY(c, s))}{2} \tag{16}$$

$$\bar{O}(c) = \sum_s Nor\left(O(c, s)\right) \tag{17}$$

$$\bar{M}(c) = \sum_s Nor\left(M(c, s)\right) \tag{18}$$

Note that the RG and BY channels are combined to generate one color channel $\bar{C}$. In order to integrate these profiles, Nothdurft's nonlinear addition model [3] is employed to count for the possible overlaps between stimuli features. The bottom-

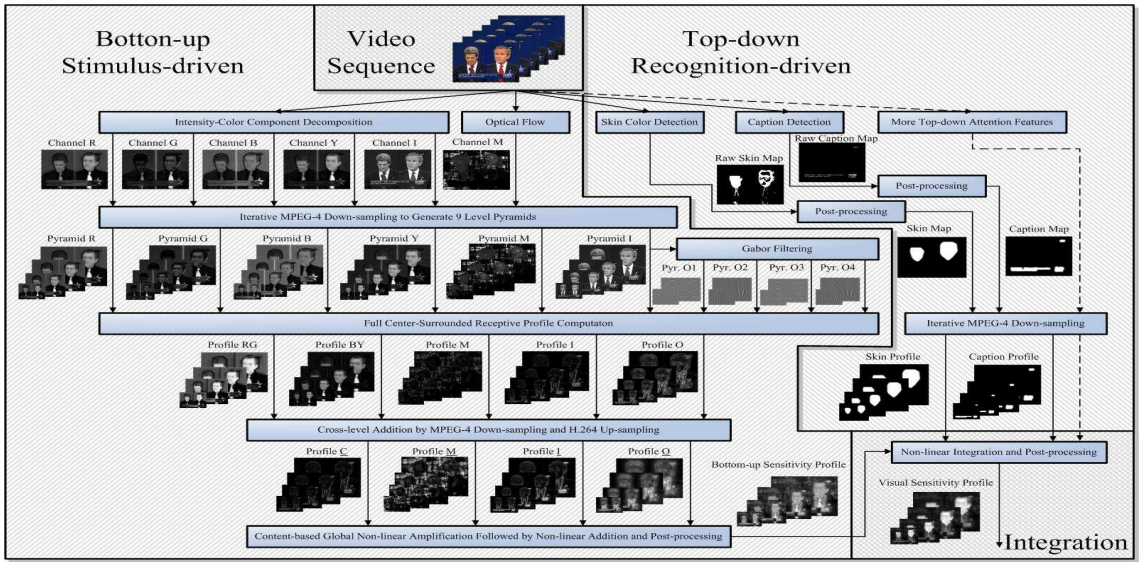


Fig. 1. The framework of computational model for the SVSP.

up attention profile (BAP) is computed as

$$\begin{aligned} BAP^{(c)} &= \bar{I}(c) + \bar{C}(c) + \bar{O}(c) + \bar{M}(c) \\ &-\lambda_{IC} \cdot MIN\left(\bar{I}(c), \bar{C}(c)\right) - \lambda_{IO} \cdot MIN\left(\bar{I}(c), \bar{O}(c)\right) \\ &-\lambda_{IM} \cdot MIN\left(\bar{I}(c), \bar{M}(c)\right) - \lambda_{CO} \cdot MIN\left(\bar{C}(c), \bar{O}(c)\right) \\ &-\lambda_{CM} \cdot MIN\left(\bar{C}(c), \bar{M}(c)\right) - \lambda_{OM} \cdot MIN\left(\bar{O}(c), \bar{M}(c)\right) \end{aligned} \quad (19)$$

The control coefficients λ are set according to psycho-visual experiment findings [3]: $\lambda_{IC} = 0, \lambda_{IO} = 0.2, \lambda_{IM} = 0.25, \lambda_{CO} = 0.8, \lambda_{CM} = 0.2, \lambda_{OM} = 0.5$.

IV. Top-down Attention Simulation

We implement two top-down cognitive-related features of visual attention in the current work, namely skin tone color and caption detections. As has been pointed out by many researchers [15], the skin color area indicates the appearance of people and often attracts human attention. More often than not, caption in video sequence contains much useful information and has already been used as a key element in content-based video indexing and retrieval.

A. Skin Color Detection

Since YCbCr color space has inherent natures in separation between luminance and chrominance and compaction of skin-color clusters, some successful skin detection methods based on YCbCr space have been developed. Among them, we adopt Hsu *et.al*'s elliptical skin model on nonlinear transformed chrominance components for detecting skin areas in color images [5]. For a frame f_i, the detected skin area in a color map is defined as

$$SM_i(x,y) = \begin{cases} 1 & : \quad (x,y) \in skinarea \\ 0 & : \quad otherwise \end{cases} \quad (20)$$

where (x, y) is the pixel index. When false alarm occurs (in our algorithm, we ensure a little over-detection through adjusting the thresholds), a morphological "open" operation is applied, denoted as $SM' = SM \odot se$, where se is a 5x5 disk structure element. After excluding false skin regions through morphological processing, we compute a convex hull to cover the survived skin color regions, so as to combat the possible "holes" resulted from the detection.

B. Caption Detection

Caption in video sequence contains lots of high-level semantic information, which certainly attracts human attention. To fully utilize the temporal information, we incorporate Luo *at. al*'s TFV (Temporal Feature Vector) based caption detection method in our system [6]. The detected caption map is denoted as

$$CM_i(x,y) = \begin{cases} 1 & : \quad (x,y) \in captionarea \\ 0 & : \quad otherwise \end{cases} \quad (21)$$

V. SVSP Integration

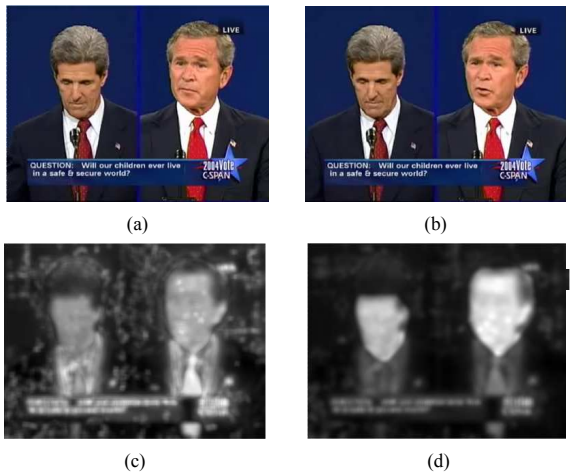
To incorporate with the pyramidal bottom-up attention profile $BAP^{(l)}, l \in [0, L]$, the top-down attention features skin map and caption map are firstly smoothed with a 5x5 Gaussian window and then iteratively filtered with MPEG-4 down-sampling filters [8]. Let the generated pyramids corresponding to skin map and caption map be $SP^{(l)}$ and $CP^{(l)}$ respectively, with $l \in [0, L]$. The visual sensitivity profile is integrated as

$$VSP^{(l)} = BAP^{(l)} \cdot \alpha^{SP^{(l)}} \cdot \beta^{CP^{(l)}} \quad l \in [0, L] \quad (22)$$

where α,$\beta \geq 1$ are weighting coefficients. Note that when $\alpha = \beta = 1$, the top-down features take no effect in modulating bottom-up attention map. Considering the fact that human face by its nature attracts more low-level human attention, we emphasize skin map more and $\alpha = 1.5, \beta = 1.2$ are set in this paper. Fig.2 presents the generated level 0 VSP of the 50th frame in *president debate*.

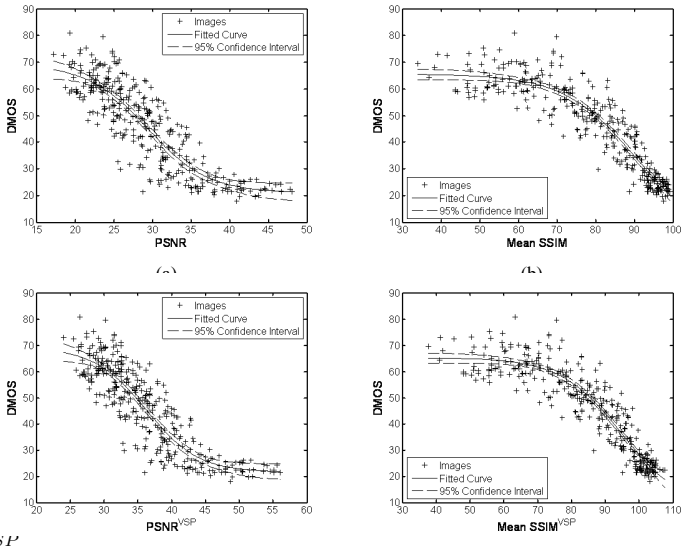
VI. SVSP Verification with Quality Assessment

The proposed SVSP can be used to augment visual quality in either scalable ($0 < l < L$) or non-scalable ($l = 0$) scenarios. Limited by the space, we only justify its validity in quality assessment in non scalable mode. For more applications in scalable modes, one may refer to [17]. We design two simple image quality assessment algorithms by modifying the simple PSNR (Peak Signal to Noise Ratio) and well known



(a) (b) (c) (d)

Fig. 2. (a) Frame 50 of video clip "president debate". (b) Frame 51 of vi clip "president debate". (c) BAP level 0. (d) VSP level 0.



(a) (b) (c) (d)

Fig. 3. Scatter plots of DMOS vs. predictions of image quality metrics

mean SSIM (Structural SIMilarity) [7], namely $PSNR^{VSP}$ and $SSIM^{VSP}$, as

$$PSNR^{VSP} = 10 \cdot log\left(\sum_{x=1}^{M}\sum_{y=1}^{N}\frac{255^2 \cdot [VSP^{(0)}(x,y)]^{-\kappa}}{[f_0(x,y) - f_d(x,y)]^2}\right) \quad (23)$$

$$SSIM^{VSP} = \frac{1}{M \cdot N}\sum_{x=1}^{M}\sum_{y=1}^{N} SSIM(x,y) \cdot [VSP^{(0)}(x,y)]^{\kappa} \quad (24)$$

where $f_0(x,y)$ and $f_d(x,y)$ are original and distorted images, $VSP^{(0)}(x,y)$ is computed from $f_0(x,y)$, $SSIM(x,y)$ is the structural similarity map computed using Wang's algorithm [7], M and N are the image dimensions, and $\kappa = 1$ is simply used in this paper. The proposed methods are tested on LIVE image database release 2 [10] with JPEG and JPEG2000 coded images of various bit rates and collected DMOS (Different Mean Opinion Score) from subjective tests. We evaluate the quantitative performance of the two proposed metrics using methods introduced by Video Quality Experts Group (VQEG) [11]. 1) The correlation between objective and subjective scores after variance-weighted regression, which evaluates the prediction accuracy. 2) The spearman rank order correlation between objective and subjective scores, which evaluates the prediction monotonicity. As can be found in Tab.I, as well as the scatter plots in Fig. 3 the proposed VSP can effectively enhance the performances of image quality metrics by differentiating important and trivial image contents and assign dissimilar weights to different regions.

VII. Conclusion

We propose the scalable visual sensitivity profile (SVSP): a hierarchy of saliency maps that simulate the bottom-up and top-down attention of the HVS. The bottom-up module is based on some earlier work with modifications and adaptations towards a hierarchical representation and more accurate simulation. The top-down module utilizes human skin-color detection and caption extraction as cognition clues. The SVSP's application in image quality assessment has been analyzed to demonstrate the validity of SVSP.

References

[1] A.M.Treisman, G.Gelade, "A feature-integration theory of attention," Cognitive Psychology, pp. 97-136, 1980.
[2] L. Itti, "Models of bottom-up and top-down visual attention." Ph.D. Dissertation California Institute of Technology, Pasadena, California, 2000
[3] H.C.Nothdurft, "Saliency from feature contrast: additivity across dimensions," Vision Research, vol. 44, no. 10, pp. 1183-1201, 2000.
[4] A.S.Ogale and Y.Aloimonos, "A roadmap to the integration of early visual modules," International Journal on Computer Vision:Special issue on early cognitive vision, 2006.
[5] R. L. Hsu, M. bdel-Mottaleb, A. K. Jain, "Face detection in color images," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 24, no. 5, pp. 696-706, 2002.
[6] B. Luo, X. Tang, J. Liu, H. Zhang, "Video caption detection and extraction using temporal information," in IEEE International Conference on Image Processing, 1 ed Barcelona, Spain: Institute of Electrical and Electronics Engineers Computer Society, 2003, pp. 297-300.
[7] Z. Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," IEEE Transactions on Image Processing, vol.13, no.4, pp.600-612, April 2004.
[8] "Mpeg4 video verification model version 18.0," JTC1/SC29/WG11 N3908, Pisa, 2002.
[9] "ITU-T Recommendation H.264 — ISO/IEC 14496-10 AVC," 2004.
[10] "http://live.ece.utexas.edu/research/quality/subjective.htm," 2005.
[11] VQEG, "Final Report from the Video Quality Experts Group on the Validation of Objective Models of Video Quality Assessment," http://www.vqeg. org, Mar.2000.
[12] C.Koch and S.Ullman, "Shifts in selective visual attention:towards the underlying neural circuitry," Human Neurobiology, vol. 4, no. 4, pp. 219-227, 1985.
[13] O. Le Meur, P. Le Callet, D. Barba, and D. Thoreau, "A coherent computational approach to model bottom-up visual attention," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 28, no. 5, pp. 802-817, 2006.
[14] A. Oliva, A. Torralba, M. S. Castelhano, and J. M. Henderson, "Top-down control of visual attention in object detection," in IEEE International Conference on Image Processing, 2003, pp. 253-256.
[15] Z. Lu, W. Lin, X. Yang, E. Ong, and S. Yao, "Modeling visual attention's modulatory aftereffects on visual sensitivity and quality evaluation," IEEE Transactions on Image Processing, vol. 14, no. 11, pp. 1928-1942, 2005.
[16] L.Itti and C.Koch, "Computational Modeling of Visual Attention," Nature Reviews Neuroscience, vol. 2, no. 3, pp. 194-203, 2001.
[17] G. T. Zhai, Q. Chen, X. K. Yang, W. J. Zhang, "Application and Performance of Scalable Visual Sensitivity Profile," submitted to IEEE International Symposium on Circuits and Systems, May, 2008, Washington, USA