Visual Masking Model Implementation for Images & Video.

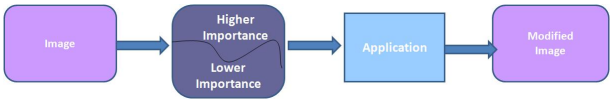
Yuhong Wang, Chi Zhang, Sukesh Kaithaapuzha

***Abstract*—Visual masking refers to the masking effect of the human visual system because of the underlying physiological & psychological mechanisms. Here we attempt to implement a masking model from a combination of various research papers that will give a relevance map of the image in terms of 8x8 pixel blocks.**

***Index Terms*—HVS, CSF, Visual Masking, JND, DCT, eye tracking, motion estimation, visual attention, foveation**

I. Introduction

Visual masking model estimates the masking effect of the HVS. There are various applications where a visual masking model could be used to do efficient image/video processing, some examples are Image/Video filtering for display, Video compression, Watermarking, Encryption/Steganography etc. These applications would utilize the masking model in determining the importance level of each of the pixels, and this information is then in application specific processing.



Our implementation of the visual masking mainly follows the functional block of [1] , the block diagram is displayed as below.

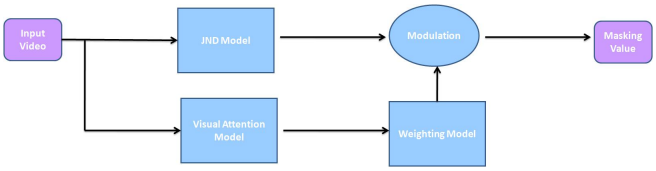


Figure 1. Diagram of JND model sub-blocks

The model consists of three main components JND(Just Noticeable Difference Model): Just Noticeable Difference is defined as the maximum distortion the human visual system cannot perceive. Our implementation of JND is based on [2]. This implementation also considers temporal properties (eye tracking) in addition to the spatial properties. Visual Attention Model : The visual attention model we have considered in this project estimates the attention point of the eye in an image/video based on bottom up space based contrast stimuli (texture, luminance & motion) & top down object based features. This is based on work down in paper [3]. Weighing Model : HVS has the highest spatial resolution & sensitivity at the point of fixation, the estimation of the fixation is incorporated into the overall model by modulating a weighing map generated using foveation method. owing the example. Some components, such as multi-leveled equations, graphics, and tables are not prescribed, although the various table text styles are provided. The formatter will need to create these components, incorporating the applicable criteria that follow.

II. Basic JND Model

JND Model Overview

JND model considered here takes into account both temporal and spatial properties of the human visual system, the model incorporates a pixel domain edge detection using canny edge detection and the utilises the results to do a block type classification. In the next stage the model operates in the frequency domain by performing a discrete cosine transform (DCT) on the input image and incorporates spatial-temporal contrast sensitivity function, the influence of eye movement, luminance adaptation & contrast masking .

$$JND(n,i,j,t) = T(n,i,j,t)a_{Lum}(n,t)$$
$$a_{intra}(n,i,j,t)a_{inter}(n,t)$$

where $a_{Lum}(n,t), a_{intra}(n,i,j,t), and\ a_{inter}(n,t)$account for the effects of luminance adaptation, intra-band masking, and inter-band masking, respectively.

The JND model sub-blocks are explained in detail below as per how we have implemented.

A. Edge Detection & Block Type Classification:

The edge detection we have employed is canny edge detection , the out put of the edge detection is a binary image with the edge pixel identified. We divide this image into blocks(8x8) which can be and calculate the edge density in each of the blocks the blocks are further classifies as Plain blocks, Edge blocks, Texture blocks. The edge density is calculated as edge density = (number of edge pixels in block / number of pixels in the block) The classification thresholds are based on the equation (3) in [1]. The block type classification results are further filtered using a Majority filter in immediate neighborhood blocks for Edge & Texture blocks. Plan blocks does not go through this filter since it is not possible to eliminate a lone Texture/Edge block from any arbitrary image without any contextual information.

B. Baseline Spatio-Temporal CSF Model: The HVS is sensitive to contrast and can only sense a signal whose contrast

is above a certain threshold with respect to a signal frequency . The reciprocal of this is the contrast sensitivity. The baseline spatio-temporal contrast sensitivity that we have implemented is based on equation (1) in the paper [2].This equation operates in the DCT sub band domain and also incorporates the eye movement effect in the form of retinal image velocity which is explained below.

Eye Movement Effect (Eye Tracking) The eye movement is classified into three types : i) Smooth-Pursuit Eye Movement : Tracks moving object and reduces retinal velocity ii)Natural Drift Eye Movement : Refers to very slow eye movement and is a used as measure for viewing static images iii)Saccadic Eye Movement : refers to rapidly moving objects to which HVS has low sensitivity.

Because of the different eye movements the perceived Retinal velocity is different from the Image plane velocity. Retinal Image Velocity is defined V= VI -VE. where VI is the image plane object velocity and VE is the eye movement velocity. In our implementation predefine equation (6) in paper [2] for calculating VE. The Image plane velocity is implement in using two different methods for motion estimation

C. Motion Estimation Motion estimation is the process of determining motion vectors that describe the transformation from one 2D image to another; usually from adjacent frames in a video sequence. It is an ill-posed problem as the motion is in three dimensions but the images are a projection of the 3D scene onto a 2D plane. In our JND model, motion estimation is used to get the image velocity representing eye movement effect.

Motion estimation using Optical flow calculation: Optical flow is the distribution of apparent velocities of movement of brightness patterns in an image. Optical flow calculation is a very popular gradient-based image matching method. It can give important information about the spatial arrangement of objects viewed and the rate of change of this arrangement. We estimate the direction and speed of object motion from one image to another or from one video frame to another using the Horn-Schunck method. By assuming that the optical flow is smooth over the entire image, the Horn-Schunck method computes an estimate of the velocity field that minimizes this equation:

$$E = \iint (I_x u + I_y v + I_t)^2 \, dxdy + \alpha \iint \left\{ \left(\frac{\partial u}{\partial x} \right)^2 + \left(\frac{\partial u}{\partial y} \right)^2 + \left(\frac{\partial v}{\partial x} \right)^2 + \left(\frac{\partial v}{\partial y} \right)^2 \right\} dxdy$$

Motion estimation using block matching method Block Matching Algorithm is a way of locating matching blocks in a sequence of digital video frames for the purposes of motion estimation. The purpose of a block matching algorithm is to find a matching block from a frame i in some other frame j, which may appear before or after i.

D. Luminance Adaptaion:

The JND model we have implemented also incorporates the luminance adaptation ,which is a property of the HVS where the eye has higher visibility threshold for dark and light regions and is more sensitive to noise in medium gray regions. The average local intensity of a block is determined by the dc component of a DCT block we use an equation (13) from paper [2] which utilizes this property

E. Contrast Masking:

The extent of contrast masking depends on the local intensity activity of the image. We perform a DCT domain Intra & Inter-band contrast masking. In this method a DCT block is divided into DC, low frequency (LF), medium frequency (MF) and High Frequency and calculate a Texture Energy and utilize the block type classification done earlier to implement equation (15) for paper [2].

III. Visual attention model

One major factor in masking is how the human eye directs attention to various parts of an image or image sequence. There are two kinds of attention features: bottom-up features, processed by the brain from neutral detail into areas of interest, and top-down features, automatically recognized as specialized qualities and transformed into evaluation of its details. In the project we implement bottom-up features of color and texture variation to find a weighted map of visual attention. In some literature, top-down processing of faces, skin tones, common objects, and patterns and motifs are incorporated into a visual attention model. However, considering that top-down recognition code is still in a highly developmental stage and in the interest of computational ability, we focus our project on just the bottom-up components. As shown below, the input video sequence is evaluated for color and texture contrast, using a k-means clustering algorithm, combined into one conglomerate map based on correlation between stimuli, and truncated according to limitations of human attention.

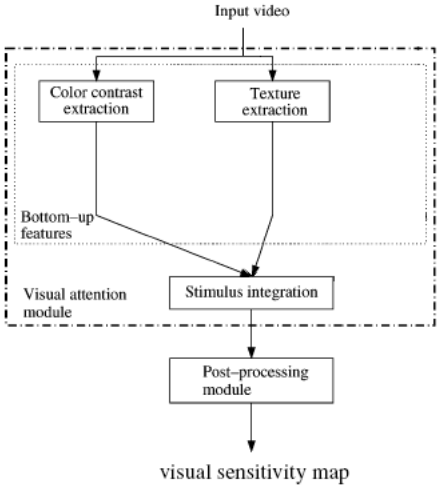


Figure 2. Diagram of visual attention process

For color, we take each block and find the average RGB values inside the block and then apply the k-means method, which seeks to find $arg\,min \sum_{i=1}^{k} \sum_{\mathbf{x}_j \in S_i} \|\mathbf{x}_j - \boldsymbol{\mu}_i\|^2$, for $\mathbf{x}_j = (R_j, G_j, B_j)$, where there are k sets S_i and

each has a mean value $\boldsymbol{\mu_i}$. Then if there is a cluster larger than some fraction of pixels, we designate that cluster as the background and compute relative distances, $d = \sqrt{\left(R-\overline{R}\right)^2+\left(G-\overline{G}\right)^2+\left(B-\overline{B}\right)^2}$, and these values are discretized into ranges to create proper scaling. Otherwise, if the image is relatively uniform, default to using the center of the image as the focal point. Texture variation is done similarly to color contrast, except we count edge pixels in each block and use $d = \mid n_e - \overline{n}_e \mid$.

How the two stimuli combine is dependent on their correlation. According to literature, color and texture are slightly correlated features, so a value of around 0.25 is a good choice, thus we have $S_{combined} = s_c + s_t - 0.25\min(s_c, s_t)$. Post-processing is based on a thresholded decaying exponential $k = e^{-\left(\frac{\rho - S_{combined}{}^{-1}}{S_{combined}{}^{+1}}\right)^2}$ if the radius is within a standard deviation , otherwise we take $k = 1$. The maximal attention values after convolution with the kernel are taken. Finally, the map is scaled and limited by the maximal attention capacity, predefined to equal the block area.

IV. WEIGHTED JND

Foveation is the tendency of the human eye to have highest resolution at points of highest attention and exponentially decreasing sensitiviy with increasing eccentricity away from the focal points. Our model incorporates the $k = 10$ most attention-weighted *fixation points* to use as the focal points. We then calculate $W = e^{\frac{\alpha f \xi}{e_0}\min\limits_k\left\{\arctan\frac{\sqrt{(x-xf_k)^2+(y-yf_k)^2}}{V}\right\}}$, where f is the local frequency, and xf_k and yf_k are the coordinates of the kth fixation points and V is the distance from the observer to the screen.

The final interpretation of the model is in the frequency domain of the DCT, with weighting from foveation based on frequency eccentricity and block type classification combined with frequency analysis, all dependent on block location and DCT coefficient values. Each block is a region where we could introduce a certain amount of noise or quantization error, the frequency of which correspond to the values of the weights in the block.

V. IMPLEMENATION RESULT

The Implementation was done in Matlab. Here we present the results of individual stages of the algorithm

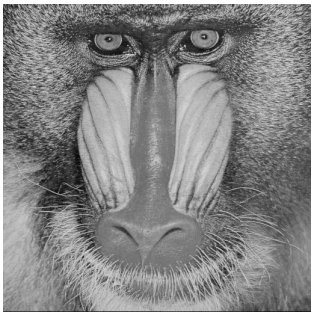


Figure 3. Input image

Figure 4. Edge Detection Output

Figure 5. BLOCK TYPE CLASSIFICATION

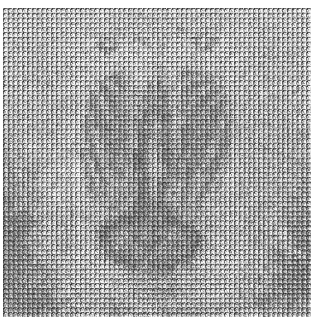


Figure 6. JND Model Output

VI. ACKNOWLEDGMENTS

We would like to thank Professor Bernd Girod and TA David Chen for their feedback and support for this project.

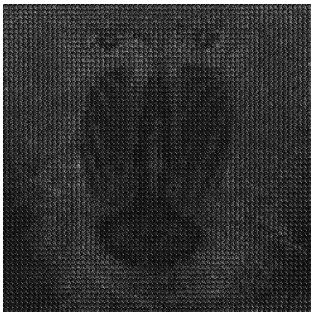


Figure 7. Visual Mask with Visual Attention in the Centre

VII. APPENDIX

Yuhong Wang programmed the weighted JND model with visual attention and foveation, without motion estimation. Sukesh Kaithakapuzha & Chi Zhang did a JND model for video with motion estimation and Majority Filter based block type classification for video. Report & Poster was done by Sukesh, Yuhong & Chi together. Andriod implementation issues discovered by all three.

REFERENCES

[1] Zhongkang Lu, Weisi Lin, Xiaokang Yang, EePing Ong, Susu Yao, Modeling Visual Attention's Modulatory Aftereffects, *IEEE Transactions on Image Processing*, Vol. 14, No. 11, November 2005, pp. 1928-1942.

[2] Anmin Liu, Maansi Verma and Weisi Lin,;"Modeling the Masking Effect of the Human Visual System with Visual Attention Model", Nanyang Technological University, Singapore.

[3] Yuting Jia, Weisi Lin, and Ashraf A. Kassim "Estimating Just-Noticeable Distortion for Video".

[4] Zhongkang Lu, Weisi Lin, Xiaokang Yang, EePing Ong, and Susu Yao, "Modeling Visual AttentionâĂŹs Modulatory Aftereffects on Visual Sensitivity and Quality Evaluation" .

[5] B.K.P. Horn and B.G. Schunck, "Determining optical flow." *Artificial Intelligence, vol 17, pp 185-203, 1981K. Elissa*.

[6] Andrew Burton and John Radford âĂIJThinking in Perspective: Critical Essays in the Study of Thought Processes." (1978)

[7] David H. Warren and Edward R. Strelow (1985). âĂIJElectronic Spatial Sensing for the Blind: Contributions from Perception. SpringerâĂİ

[8] Philip H.S. Torr and Andrew Zisserman," Feature Based Methods for Structure and Motion Estimation", *ICCV Workshop on Vision Algorithms, pages 278-294, 1999*