# AN EFFECTIVE PERCEPTUAL WEIGHTING MODEL FOR VIDEOPHONE CODING

*X.K. Yang, W.S. Lin, Z.K. Lu, E.P. Ong, S.S. Yao*

Institute for Infocomm Research, Agency for Science, Technology and Research, Singapore 119613
Email: xkyang@ieee.org, {xkyang,wslin,zklu,epong,ssyao}@i2r.a-star.edu.sg

**ABSTRACT**

In this paper, a perceptual weighting model is proposed for effective rate control so as to enhance perceptual coding quality of videophone, by exploiting two categories of factors affecting the perception of the human visual system: stimulus-driven factors and cognition-driven factors. In order to achieve a simple but effective perceptual weighting model, we use luminance adaptation and texture masking as the stimulus-driven factors, while skin color serves as the cognition-driven factor in the videophone application. Both objective and subjective quality evaluations of videophone-like sequences in H.263 platform validate the effectiveness of our perceptual weighting model.

## 1. INTRODUCTION

Videophone is expected to be a killer-application of next-generation wireless. A typical videophone scene is basically composed of a human body (usually a speaker) and a background. The conveyed visual information of primary interests in videophone communication is usually the human bodies in the scene.

Block-based motion compensation has been widely adopted by the prevalent video coding standards, such as H.261/263 and MPEG 1/2/4. It is based on the translational rigid motion model of blocks, and its popularity is due to the low computational complexity and the low overhead requirements to represent the motion field. However, the translational rigid motion model fails for zooming, rotational motion, and deformations of nonrigid objects (such as face and hand motion [1]). As a nonrigid object, the human body involves complex motion, rotation (of hands and the head) and local deformations. If the translational rigid motion model of blocks is applied to videophone coding, it results in poorer coding quality for the foreground region than the background region, as will be illustrated in Section 2. Therefore it is desirable to devise a perceptual model to differentiate the foreground and background regions for rate control so that the available bits can be assigned to maximize the perceptual quality of coded video.

Several schemes [2, 3] have been proposed recently to improve the performance of coding systems in terms of perceptual quality of foreground objects, by exploiting certain prior knowledge in videophone applications. These schemes firstly segment facial regions from the scene, and then apply finer quantization in facial region and coarser quantization in non-facial region. However, the existing schemes lack a unified perceptual importance model for rate control. Consequently the bit-rates between facial region and non-facial region are heuristically controlled, and the quantization schemes cannot be adaptive to local perceptual significance within the facial region and the non-facial region.

In this paper, a perceptual weighting model is proposed for effective rate control so as to enhance perceptual coding quality of videophone, by exploiting two categories of factors affecting the perception of the human visual system: stimulus-driven factors and cognition-driven factors. We firstly evaluate the typical block-based coding scheme for the videophony application and justify the necessity of allocating bits effectively throughout the frame according to the Human Visual System (HVS) sensitivity for quality perception. In order to achieve a simple but effective perceptual weighting model, we then use luminance adaptation and texture masking as the stimulus-driven factors, while skin color serves as the cognition-driven factor in the videophone application. Both objective and subjective quality evaluations of videophone-like sequences in H.263 platform validate the effectiveness of our perceptual weighting model.

## 2. ANALYSIS OF BLOCK-BASED CODING FOR VIDEOPHONY

Due to its simplicity, block-based coding has been widely adopted by the prevalent video compression standards, so is the primary choice for videophony. We now examine the peak signal-to-noise ratio (PSNR) characteristics of block-based coding with videophony-like sequences. Let $PSNR_F$, $PSNR_B$ and $PSNR_W$ denote the PSNR of the foreground region,the background region and the whole image, respectively.

We use four typical videophony-like scenes from standard test sequences: *Carphone*, *Foreman*, *Silent*, and *Suzie*. These four sequences are compressed using an H.263 encoder with the TMN8 rate control scheme [4]. The foreground region is detected using the skin color detection algorithm (to be described in Section 3.2), and the detection results are illustrated in Figure 1. The average $PSNR_F$, $PSNR_B$ and $PSNR_W$ of the four sequences, coded by TMN8 as well as the perceptual rate control (PRC) scheme with our proposed perceptual weighting model, are presented in Table 1, for a wide range of bit rates (Note: PSNRs of PRC are listed in this table in order to save paper space and they will be discussed in Section 4). When we investigate the relationship of PSNRs for TMN8, it can be seen that $PSNR_F$ is consistently lower than the $PSNR_B$ for all sequences. This implies that foreground objects need more bits with block-based coding due to nonrigid deformations and the underlying motion of more complex nature (i.e., rotation). The conventional video coding scheme without considering unequal perceptual importance in rate control results in lower

PSNR for the foreground. Such results are undesirable because of the inconsistence with the HVS perception. Therefore, it is necessary to develop a new rate control scheme that takes the unequal perceptual importance among different objects into account.

(a) Carphone (b) Foreman (c) Silent (d) Suzie

Figure 1: Skin maps for the 120-th frame.

Table 1: Comparison of PSNR (dB)

| Bit Rate (kpbs) | | | 32 | 64 | 96 | 128 |
|---|---|---|---|---|---|---|
| Carphone | $PSNR_F$ | TMN8 | 32.88 | 35.67 | 37.31 | 38.29 |
| | | PRC | 33.47 | 36.97 | 38.88 | 39.77 |
| | $PSNR_B$ | TMN8 | 34.45 | 37.47 | 39.39 | 40.57 |
| | | PRC | 33.47 | 36.97 | 38.88 | 39.77 |
| | $PSNR_W$ | TMN8 | 34.23 | 37.22 | 39.08 | 40.23 |
| | | PRC | 33.66 | 36.66 | 38.62 | 39.72 |
| Foreman | $PSNR_F$ | TMN8 | 29.52 | 32.66 | 34.31 | 35.36 |
| | | PRC | 29.66 | 33.65 | 35.82 | 37.21 |
| | $PSNR_B$ | TMN8 | 30.09 | 32.77 | 34.37 | 35.52 |
| | | PRC | 29.25 | 31.58 | 32.95 | 33.99 |
| | $PSNR_W$ | TMN8 | 29.93 | 32.73 | 34.34 | 35.45 |
| | | PRC | 29.31 | 31.96 | 33.46 | 34.53 |
| Silent | $PSNR_F$ | TMN8 | 30.84 | 34.15 | 36.11 | 36.94 |
| | | PRC | 31.57 | 35.17 | 37.40 | 38.31 |
| | $PSNR_B$ | TMN8 | 32.57 | 36.13 | 38.89 | 41.10 |
| | | PRC | 32.15 | 35.52 | 38.24 | 40.47 |
| | $PSNR_W$ | TMN8 | 29.93 | 32.73 | 34.34 | 35.45 |
| | | PRC | 29.31 | 31.96 | 33.46 | 34.53 |
| Suzie | $PSNR_F$ | TMN8 | 33.12 | 35.97 | 37.57 | 38.66 |
| | | PRC | 33.56 | 36.92 | 38.89 | 40.05 |
| | $PSNR_B$ | TMN8 | 35.68 | 38.43 | 40.15 | 41.32 |
| | | PRC | 34.86 | 37.26 | 38.86 | 39.96 |
| | $PSNR_W$ | TMN8 | 34.98 | 37.76 | 39.45 | 40.58 |
| | | PRC | 34.53 | 37.18 | 38.85 | 39.97 |

## 3. PERCEPTUAL WEIGHTING MODEL FOR VIDEOPHONE CODING

Figure 2 is a diagram of the proposed perceptual weighting model for videophone coding. There are four major modules in Figure 2: stimulus-driven sensitivity detection, skin color detection, perceptual integration, and postprocessing.

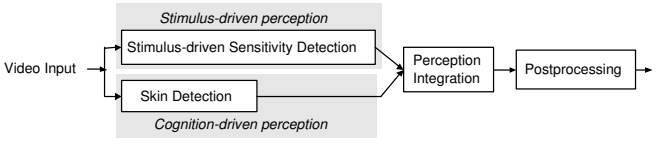



Figure 2: The perceptual weighting model for videophone coding

### 3.1. Stimulus-driven Sensitivity

Stimulus-driven sensitivity factors refer to the factors relating to the optical property of eyes and the retina structure, as well as the signal transmission property via the vision path. There are various such factors that could be modelled into video quality assessment, e.g., color, spatial masking, temporal masking, motion. Considering the low delay, low power and low resolution constraints of the videophone application, we only use two spatial masking related factors in this study: luminance adaptation and texture masking. Luminance adaptation refers to the fact that the HVS is sensitive to luminance contrast rather than absolute luminance value, while texture masking denotes that textured regions can hide more error than smooth areas. In our previous work [5], an nonlinear additivity model for masking (NAMM) has been proposed to combine luminance adaptation and texture masking effects into a just noticeable distortion (JND) profile of image.

In this paper, the stimulus-driven sensitivity for a pixel in an image is defined by combining the luminance adaptation and texture masking effects (as the reciprocal of the just noticeable distortion (JND) [5]). The stimulus-driven sensitivity can be given by:

$$S(x,y) = (T_l(x,y)+T_t(x,y)-C_{l,t}\cdot\min\{T_l(x,y),T_t(x,y)\})^{-1} \quad (1)$$

where $T_l(x,y)$ and $T_t(x,y)$ are the visibility thresholds for background luminance adaptation and texture masking for the pixel located at $(x,y)$, and they can be determined by the methods proposed in [6]; and $C_{l,t}(\in(0,1))$ is the gain reduction factor due to overlapping between two masking stimuli.

### 3.2. Skin Color as Cognition-driven Perception Factor

Cognition-driven perception factors reflects the human's cognitive processing, such as object/pattern recognition based on the knowledge and experience. In the videophone application, the presence of the human body no doubt is a cognitive-driven factor attracting visual attention. In general, the faces (sometimes hands as well) are the most significant part of visual attention in the conversational (usually head-and-shoulder) application. Skin color can be therefore used as a cognition-driven perceptual clue, since both faces and hands are normally present in the regions with skin color. Face detection is currently still a challenging task because of the variability in scale, location, orientation, and pose [7]. If false face detection occurs from time to time in a video sequence and if such face maps were used for rate control, the perceptual quality would be greatly reduced due to the possible variation in human face detection. The detection of skin color is much easier and more robust.

Several color spaces have been utilized to label pixels as skin [7]. $YC_bC_r$ color space is adopted here since it has been used in the prevalent image/video compression standards, and leads to good performance of skin color detection in terms of the separation between luminance and chrominance, and the compactness of the skin clusters. The difficulty of detecting the low-luma and high-luma skin tones can be efficiently overcome by applying nonlinear transform in $YC_bC_r$ color space. Hsu's nonlinear transform of Chroma and the elliptical skin model in $YC_bC_r$ color space [8] is used in this paper.

Figure 1 shows the detected skin maps for four standard testing sequences at the 120-th frame. The rough segmentation for skin regions provides effective indication on the whereabouts of foreground objects.

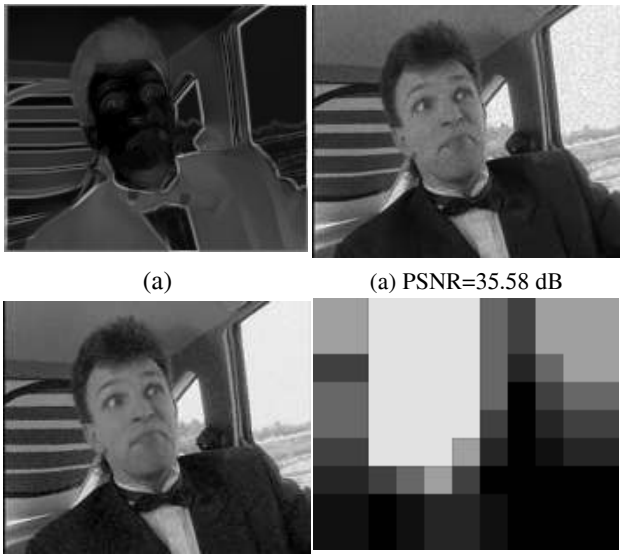
### 3.3. Perceptual Integration

Since the distortion in the skin region of videophone is the most intolerable from the viewpoint of cognition-driven perception, we compute the cognition-integrated perceptual sensitivity by simply scaling $S(x,y)$ values for the pixels within the skin map $\mathcal{F}$ as follows:

$$S_c(x,y) = \begin{cases} \frac{\max\limits_{1\le x\le X,\ 1\le y\le Y}(S(x,y))}{\max\limits_{(x,y)\in\mathcal{F}}(S(x,y))} \cdot S(x,y) & \text{if } (x,y)\in\mathcal{F} \\ S(x,y) & \text{otherwise} \end{cases} \quad (2)$$

where $X$ and $Y$ denote the image dimensions.

Figure 3 illustrates how the $S_c(x,y)$ profile reflects the perceptual sensitivity in an image. The $S_c(x,y)$ map of *Carphone* is given in Figure 3(a), where the brightness represents the sensitivity strength (i.e., the brighter, the larger stimulus-driven sensitivity). The original image of the 120-th frame of *Carphone* sequence is corrupted by random noise injection as in Figure 3(b). As can be seen, the perceptual quality of the randomly noised image is very poor. If we shape the same amount of noise (therefore with similar PSNR) according to the $S_c(x,y)$ profile (i.e., more noise in less sensitive areas), the noise is almost invisible (as in Figure 3(c)). $S_c(x,y)$ reflects the local perceptual sensitivity for coding error, and therefore is the basis of sensible bit allocation for both foreground and background objects.



(a) (a) PSNR=35.58 dB

(b) PSNR=35.53 dB (d)

Figure 3: The 120-th frame of *Carphone*: (a) The $S_c(x,y)$ map; (b) Image corrupted by random noise injection; (c) Image corrupted by noise injection with the cognition-integrated sensitivity map; (d) The $w(n)$ map.

### 3.4. Postprocessing: Avoidance of Over-adjustment

In typical block-based video coders, the number of the bits used and the resultant distortion for a given macroblock depend on the macroblock's quantization parameter (QP), which determines the step size Q for quantizing the transformed coefficients. We can determine a perceptual sensitivity weight for the $n$-th macroblock ($1 \le n \le N$, and $N = XY/256$), denoted as $w(n)$, as follows:

$$w(n) = \frac{\sum_{i=1}^{16}\sum_{j=1}^{16} S_c(n,i,j)}{A} \quad (3)$$

where $S_c(n,i,j)$ represents the cognition-integrated sensitivity at the $(i,j)$-th pixel of the $n$-th macroblock, and $A$ is the number of pixels in a macroblock (i.e., $A = 16 \times 16$).

To avoid too frequent changes in QP, which is differentially encoded in a raster-scan order, $w(n)$ is filtered with a $p \times p$ morphological close operator. The postprocessing also reduces the influence of the inaccuracy of skin region detection towards the final sensitivity weights.

The resultant $w(n)$ map of *Carphone* is given in Figure 3(d), where the brightness represents the strength of perceptual sensitivity. It can be seen that $w(n)$ is in line with the HVS perception.

### 3.5. Perceptual Rate Control (PRC)

In video phone situations, the distortion for the $n$-th macroblock is mainly introduced by quantizing its DCT coefficients. With the rate-distortion model and Lagrangian optimization [9], the optimized quantization step size $Q^*(n)$ can be determined:

$$Q^*(n) = \sqrt{\frac{A\cdot K\cdot\sigma(n)}{(B - A\cdot N\cdot C)\cdot w(n)}\sum_{n=1}^{N} w(n)\sigma(n)} \quad (4)$$

where $K$ and $C$ are constants, $B$ is the number of available bits for the current frame, and $\sigma(n)$ is the standard deviation of the luminance and chrominance values in the macroblock.

## 4. PERFORMANCE EVALUATION

Table 1 compares the average PSNRs between the TMN8 rate control scheme and the proposed PRC scheme with the above perceptual weighting model for foreground, background and whole image in the four sequences with a range of bit rates. Similar to Section 2, the foreground is indicated by the skin region. It can be seen that $PSNR_F$ is significantly improved by the PRC scheme in comparison with the TMN8 rate control scheme. For *Carphone* and *Foreman* sequences, $PSNR_F$ becomes higher than $PSNR_B$ after the proposed PRC scheme is incorporated. For the other four sequences, $PSNR_F$ is still lower than $PSNR_B$ after the proposed PRC scheme is incorporated, but the gap between $PSNR_F$ and $PSNR_B$ is greatly reduced. We can see that the increase of $PSNR_F$ is at the cost of the slight decrease of $PSNR_B$ and $PSNR_W$.

Next, we performed subjective quality evaluation, in order to further confirm the overall coding quality improvement by the PRC scheme, in spite of the slight decrease of $PSNR_W$. Double Stimulus Continuous Quality Scale (DSCQS) method, proposed by Rec.

ITU-R BT.500 [10], was used to evaluate the subjective quality of a decoded sequence (obtained with the proposed PRC or the H.263 TMN8 rate control scheme) relative to its associated original sequence. Each display session for an original sequence and an associated decoded sequence is: *Video Sequence 1*, *two seconds of grey screen*, *Video Sequence 2*, *two seconds of grey screen*. The display repeats twice before the viewers are requested to vote for the quality of each sequence. Both the display order of the sequences in a session and the order of the four test sequences were randomized for viewers. The Mean Opinion Score (MOS) scales for viewers to vote for the quality after viewing are: Excellent (100-80), Good (80-60), Fair (60-40), Poor (40-20) and Bad (20-0). Ten observers (five of them are with average image processing knowledge and the rest are naive) were involved in the experiments. The subjective visual quality assessment was performed in a typical laboratory environment with normal lighting, using a 21" EIZO T965 professional color monitor with resolution of $1600 \times 1200$. The viewing distance is approximately six times of the image height.

Difference Mean Opinion Scores (DMOS) are calculated as the difference of MOSs between the original video and the decoded video. The smaller the DMOS is, the higher perceptual quality of the decoded video has when compared with the original video. Table 2 compares the averaged DMOSs over the all 10 observers for the four sequences. From the table, we can see that the subjective rating is consistently better for the decoded sequences with with the PRC, and an average subjective quality gain of 12.1 measured in DMOS is achieved by the proposed scheme.

Table 2: Comparison of subjective quality

| Sequences | Bit-rate (kbps) | DMOS | | |
|---|---|---|---|---|
| | | TMN8 | PRC | gain by PRC |
| *Carphone* | 64 | 64 | 47 | 17 |
| | 96 | 48 | 33 | 15 |
| *Foreman* | 64 | 70 | 45 | 15 |
| | 96 | 51 | 32.6 | 18.4 |
| *Silent* | 48 | 52 | 46 | 6 |
| | 64 | 40 | 31 | 9 |
| *Suzie* | 48 | 68 | 61 | 7 |
| | 64 | 52 | 43 | 9 |
| Average DMOS gain achieved by PRC | | | | 12.1 |

## 5. CONCLUSIONS

In videophone/videoconferencing, the block-based video coding schemes tend to result in lower coding quality with foreground objects (e.g., the head and hands of the talking person(s)) than background objects. This is because of the poorer motion estimation as the result of non-translational motion and the nonrigid deformations of human bodies in the scene. Hence, it is desirable to allocate bits effectively throughout the frame according to the HVS sensitivity for quality perception, since the available bandwidth is very limited in such an application.

In this paper, a perceptual weighting model has been proposed to achieve effective rate control for enhancing perceptual coding quality of videophone compression. For efficiency, luminance adaptation and texture masking are extracted as stimulus-driven factors, while skin color is used as the cognition-driven factor. The resultant perceptual sensitivity weight for each macroblock is formed to guide the determination of quantization step with the rate-distortion model, and allows fully adaptive bit-allocation for both foreground and background objects. It has been demonstrated that in comparison with the existing H.263 TMN8 rate control scheme the proposed PRC scheme achieves higher PSNR for foreground objects at a fixed bit rate. The subjective viewing tests further confirm the overall perceptual quality improvement by the PRC scheme.

## 6. REFERENCES


[1] A. M. Tekalp, *Digital Video Processing*, Prentice Hall, 1995.

[2] D. Chai and K.N. Ngan, "Face segmentation using skin-color map in videophone application," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 9, no. 4, pp. 551–564, June 1999.

[3] J. Hartung, A. Jacquin, J. Pawlyk, J. Rosenberg, H. Okada, and P. E. Crouch, "Object-oriented h.263 compatible video coding platform for conferencing applications," *IEEE J. Select. Areas Comm.*, vol. 1, no. 3, pp. 264–277, 1999.

[4] ITU-T/SG15, "Video codec test model, near-term, version 8 (tmn8)," Portland, June 1997.

[5] X.K. Yang, W.S. Lin, Z.K. Lu, E.P. Ong, and S.S. Yao, "Just-noticeable-distortion profile with nonlinear additivity model for perceptual masking in color images," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Hong Kong, April 2003, vol. 3, pp. 609–612.

[6] C.-H. Chou and C.-W. Chen, "A perceptually optimized 3-d subband image codec for video communication over wireless channels," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 6, no. 2, pp. 143–156, 1996.

[7] M.-H. Yang, D. J. Kriegman, and N. Ahuja, "Detecting faces in images: A survey," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 24, no. 1, pp. 34–58, Jan. 2002.

[8] R.-L. Hsu, A.-M. Moharmed, and A.K. Jain, "Face detection in color images," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 24, no. 5, pp. 696–706, May 2002.

[9] J. Ribas-Corbera and S. Lei, "Rate control in dct video coding for low-delay communications," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 9, no. 1, pp. 172–185, Feb. 1999.

[10] ITU-R, "Methodology for the subjective assessment of the quality of television pictures, itu-r rec. bt. 500-9," 1999.