Perceptually Scalable Extension of H.264

Hojin Ha, Jincheol Park, Sanghoon Lee, *Member, IEEE,*
Alan Conrad Bovik, *Fellow, IEEE*

***Abstract*—We propose a novel visual scalable video coding (VSVC) framework, named VSVC H.264/AVC. In this approach, the non-uniform sampling characteristic of the human eye is used to modify scalable video coding (SVC) H.264/AVC. We exploit the visibility of video content and the scalability of the video codec to achieve optimal subjective visual quality given limited system resources. To achieve the largest coding gain with controlled perceptual quality degradation, a perceptual weighting scheme is deployed wherein the compressed video is weighted as a function of visual saliency and of the non-uniform distribution of retinal photoreceptors. We develop a resource allocation algorithm emphasizing both efficiency and fairness by controlling the size of the salient region in each quality layer. Efficiency is emphasized on the low quality layer of the SVC. The bits saved by eliminating perceptual redundancy in regions of low interest are allocated to lower block-level distortions in salient regions. Fairness is enforced on the higher quality layers by enlarging the size of the salient regions. The simulation results show that the proposed VSVC framework significantly improves the subjective visual quality of compressed videos.**

***Index Terms*—Frequency weighting, H.264/AVC, human visual system, perceptual coding, scalable video coding, visual attention.**

I. Introduction

The EXPLOSIVE growth of multimedia applications such as network video broadcasting, video-on-demand, and video conference has energized networked visual communication as an active research area. To transmit voluminous video data over the available bandwidth of networks, considerable efforts have been applied to the development of video compression techniques such as H.261, H.263, H.264, MPEG-1, 2, and 4 [1], [2]. To improve the performance of visual communication systems, it is necessary to consider both the perceptual quality of the reconstituted and displayed videos [14]–[24], as well as the seamless delivery and assurance of quality using scalable video coding (SVC) techniques to mediate perceptual video quality as a function of bitrate [7]–[12].

Recently, a new SVC H.264/AVC has been developed as an amendment of the H.264 and MPEG-4 part 10 video standards [4]–[6]. The notion of perceptual weighting is enabled by varying the quantization parameter (QP) from 1 to 51 in the standard [13], [14]. If the QP is constant over all macroblocks (MBs) in each picture, then the perceptual importance of each MB is regarded to be likewise equal regardless of content. Of course, this may result in MBs located in regions of high interest or visual attention being quantized to an annoying degree, leading to loss of subjective quality in the reconstituted video.

In our proposed visual scalable video coding (VSVC) framework, we incorporate perceptual weighting into the resource allocation algorithm. This idea is not entirely new. In [33], an improvement of visual quality was achieved by allocating more encoding bits for regions having high perceptual salience. In [15]–[21], a non-uniform spatial filtering law, called *foveation*, was employed to define spatial perceptual weights on the MBs. Larger weights were applied near presumed visual fixation points, which were represented at high resolution, while lower weights were assigned to peripheral points. This process of foveation weighting attempts to match the non-uniform density of photoreceptors over the retina. The local spatial bandwidth (LSB) rapidly decreases with distance from the presumed fixation point(s).

Fig. 1 is an example of controlling perceptual quality. In one image, the QP is fixed at 40, while in the other, local perceptual quality at the MB level is controlled by varying the QP in the range 34 to 46. In this example, the 16th frame of the *Soccer* sequence was used for visual quality inspection, where the presumed visual fixation point is on the soccer player near the right center. It is visually observable that higher perceptual quality is obtained by controlling the QP as a function of each MB. This suggests that adaptive perceptual weighting by spatially localized control of the QP is promising for generally improving the perceptual video quality.

Approaches that incorporate elements of visual perception into video coding have been studied in [14]–[24]. In particular, perceptual quality can be effectively improved by utilizing foveation in MPEG/H.263 video coding, if the fixations points can be acquired or accurately guessed. Resource allocation is accomplished as a function of the spatial distance from foveation point(s) [16]–[21]. Fixation point selection can be directly determined through the use of eye-tracking, or if that is

Manuscript received October 27, 2008; revised April 12, 2010 and December 28, 2010; accepted February 13, 2011. Date of publication March 28, 2011; date of current version November 2, 2011. This research was supported in part by the Ministry of Knowledge Economy, Korea, under the National HRD support program for convergence information technology supervised by the National IT Industry Promotion Agency (NIPA), under Grant NIPA-2010-C6150-1001-0013, and by the Basic Science Research Program through the National Research Foundation of Korea funded by the Ministry of Education, Science and Technology, under Grant 2010-0011995. This paper was recommended by Associate Editor H. Gharavi.

H. Ha is with Digital Media and Communications Research and Development Center, Samsung Electronics, Suwon 443-373, Korea (e-mail: hojini@samsung.com).

J. Park and S. Lee are with the Department of Electrical Engineering, Yonsei University, Seoul 120-749, Korea (e-mail: dewofdawn@yonsei.ac.kr; slee@yonsei.ac.kr).

A. C. Bovik is with the Laboratory for Image and Video Engineering, Department of Electrical and Computer Engineering, University of Texas, Austin, TX 78712 USA (e-mail:bovik@ece.utexas.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCSVT.2011.2133370

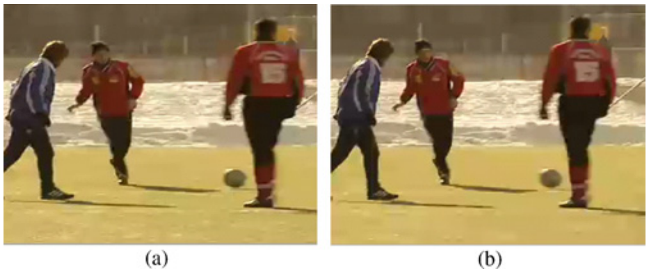

Fig. 1. Comparison of perceptual quality. (a) Fixed QP allocation for each MB. (b) Dynamic QP allocation for each MB in a picture.

inconvenient, by the use of visual salience measures based on color, skin detection, luminance edges, motion, or other visual attractors [28]–[31]. Motion and other spatiotemporal features have been used to identify regions of visual importance [22]–[28]. In [24], motion information is used to discover perceptually significant regions at low complexity. In our application, the sophisticated motion prediction tools in the H.264/AVC coder suggests that spatio-temporal features are readily available for deciding assessing visual saliency [26]–[28].

The approach taken in this paper is to allocate limited resources to the MBs in each frame in order to mediate the perceptual quality of each layer. When the number of target encoding bits for a layer is given, an optimization procedure is formulated to enable block-level resource allocation based on the video content.

First, we propose a perceptual weighting scheme based on foveation to enable the capture of visible frequencies in video. The scheme consists of two parts: visual salience is determined using motion information [23], [24]; then, the LSB is calculated for each MB based on a foveation model [15]–[21].

Second, we develop a resource allocation algorithm for optimizing perceptual quality by utilizing the computed motion-based salience to control the scalable quality layers in the SVC. The allocation algorithm enhances perceptual quality by allocating more bits to highly localized salient regions at the lower quality layers. If degradation occurs in less salient regions, higher quality will still be attained in more salient region(s). At the higher quality layers, the salient region(s) is expanded. Fairness among MBs is fulfilled by allocating more bits to region(s) of lower saliency at the highest quality layer. Thus, a tradeoff between efficiency and fairness is mediated based on the number of available bits and the associated quality layer.

We provide simulations where we measure the performance gain in terms of efficiency and fairness relative to conventional algorithms. To measure efficiency, we utilize the foveated peak signal-to-noise ratio (FPSNR). To measure fairness, we plotted the distribution of mean square error (MSE) of a frame.

II. System Description

A. Motivation

Perceptual weighting of MBs is an approach that has been effectively used for single-layer video coding using rate control [16], [20], [21], [25], [28]. For SVC H.264/AVC coding, we are unaware of any developed approach that, e.g., adaptively selects a QP and allocates bits to each layer based on perceptual saliency or foveation. Such an approach is feasible and promising; by using a rate control algorithm, the number of target bits could be decided, and a constant QP determined for each layer, on a frame-by-frame basis. This would require flexibly assigning perceptual weights to each MB, as well as an appropriate bit allocation scheme.

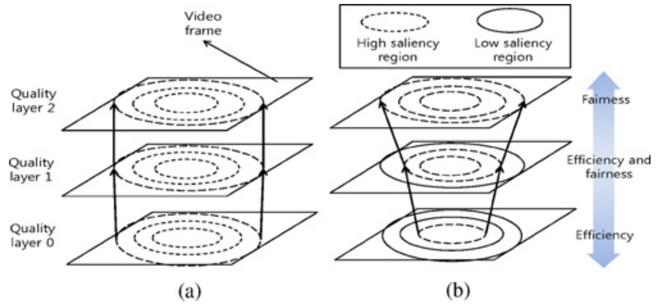

Fig. 2. Proposed VSVC framework compared to a conventional SVC scheme. (a) Conventional SVC scheme. (b) Proposed VSVC scheme.

To improve the coding performance of each quality layer, our proposed VSVC approach exploits the non-uniform sampling of the eye's sensory apparatus by controlling the sizes of identified salient regions across layers, and by controlling the number of bits for each MB via a rate control algorithm. Fig. 2 shows the mechanics of the proposed VSVC framework compared to a conventional approach. At each quality layer, the region indicated by dotted lines indicates regions of presumed higher salience, while the solid lines indicate regions of lower salience.

The conventional approach shown in Fig. 2(a) applies an equal perceptual weighting across the layers. However, the human visual response is higher for lower frequency information, which can be taken into account when selecting the quantization levels. In the conventional approach, no account of visual attention or spatial assignment of visual importance or salience is used in defining the three quality layers.

By comparison, Fig. 2(b) shows the various features of our proposed VSVC framework. The three quality layers are configured to optimize perceptual quality, i.e., an MB-level rate control mechanism computes the perceptual weights as a function the identified perceptually most salient regions in each quality layer, which are dynamically selected. For the lowest quality layer (layer 0), a small region of visual importance is maintained over which perceptual quality will be maintained. The region(s) of presumed perceptual interest is larger in layer 2, and larger still in layer 3. Significant savings in bit allocation in all three layers can be obtained in this way, with high efficiency in layers 0 and 1. Fair resource allocation can be achieved in layer 2 to maintain the perceptual quality of all MBs by expanding the sizes of the salient region(s).

B. Overview of the Proposed VSVC Algorithm

Fig. 3 is a block diagram for the proposed VSVC scheme. Fig. 3(b) and (c) shows the result of applying foveation-based

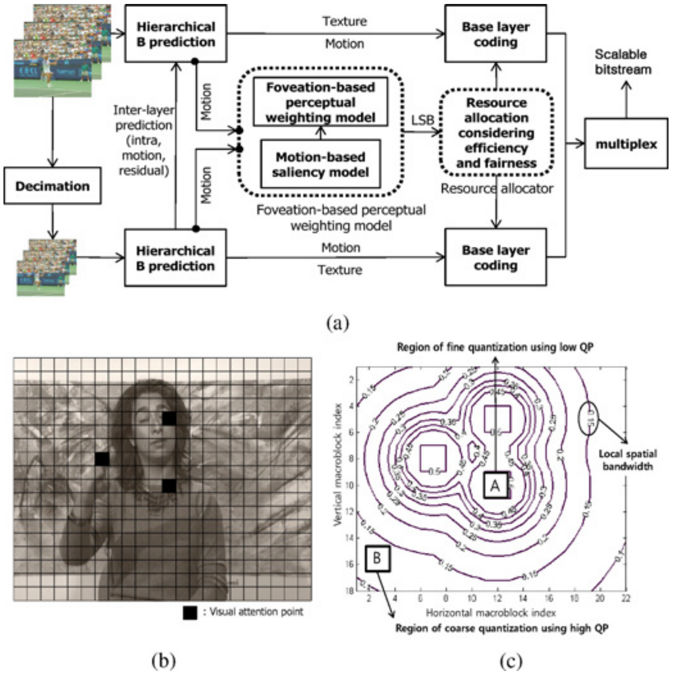
Hierarchical B prediction
Texture
Motion
Base layer coding
Scalable bitstream
Motion
Foveation-based perceptual weighting model
Motion-based saliency model
LSB
Resource allocation considering efficiency and fairness
multiplex
Inter-layer prediction (intra, motion, residual)
Decimation
Motion
Resource allocator
Foveation-based perceptual weighting model
Hierarchical B prediction
Motion
Texture
Base layer coding
(a)
: Visual attention point
(b)
Region of fine quantization using low QP
Local spatial bandwidth
A
B
Vertical macroblock index
Horizontal macroblock index
Region of coarse quantization using high QP
(c)

Fig. 3. Overview of the proposed VSVC scheme. (a) Block diagram of the proposed VSVC scheme. (b) Result of applying the motion-based salience model. (c) Depiction of the foveation-based perceptual weighting scheme. Foveation-induced bandwidths are plotted as iso-contours within which the perceptual weights, quantization, and bit allocations are determined.

perceptual weighting on the 35th frame of the *Silent* video test clip.

The proposed VSVC structure is based on the basic SVC H.264 design, which is classified as a layered video coder. Fig. 3(a) shows the typical coder structure with two spatial layers represented using a solid line. Since the proposed VSVC algorithm extends the coarse-grain Signal-to-noise ratio (SNR) scalability, the inter-layer prediction mechanism using the upsampling operation is omitted [34], [35]. The dotted processing blocks are added for the proposed SVC algorithm.

Next, the LSB for each MB is found by the perceptual weight allocator, which determines motion-based salience, determines which MBs fall within the salient region, and applies foveation-based perceptual weights.

Fig. 3(b) illustrates the outcome of the motion-based salience model. The face and left hand, both of which are in motion, are selected as region of heightened visual interest.

The LSB is decreased exponentially from the center of each salience region, which we will call the *foveation point*. The intention is that when a visual fixation falls on the region, the projection of the distribution of LSBs onto the retina will approximately match the non-uniform distribution of the photoreceptors. A map illustrating the foveation-based perceptual weighting model is shown in Fig. 3(c). Resources are allocated according to spatial placement within the indicated isocontours of the foveation-induced LSBs. In this example, MBs in region A are located in a high salient region and are thus finely quantized. MBs in region B are located in regions of lower saliency resulting in coarser quantization.

By considering visual importance together with bit allocation at each quality layer, the resource allocator seeks to optimize efficiency and fairness to achieve improved perceptual video quality. Encoding each quality layer depends on the QP, on motion information, and on identifying salient regions. Finally, the encoded bits are multiplexed to produce a scalable video bitstream.

III. Foveation-Based Perceptual Weighting Allocator

We introduce two perceptual models: one is motion-based and the other is foveation-based. Using these models, it is possible to calculate the LSB for each MB in each quality layer. The perceptual weights are defined over the spatial and temporal domains using the following observations.

1) Moving objects in video are strong attractors of visual attention and tend to draw visual fixations [23], [24], [26], [29], [37], [38].
2) The eye is sensitive to the temporal correlation of moving objects [22].
3) Fixated regions in a video are perceived at high resolution via foveal sampling [15], [20], [21], [29].

The motion-based saliency model exploits the first and second assumptions. The third assumption is employed in the foveation-based perceptual weighting model.

Fig. 4 diagrams the foveation-based perceptual weight allocator, and shows results from its operation. The weight allocation algorithm consists of four stages. Stages A–C accomplish the motion-based saliency determination, while Stage D obtains the foveation-based perceptual weighting. This means that we regard the salient MBs as an extended form of foveation points in the unit of MB rather than in the unit of pixel for obtaining the LSB of MBs in each frame. Simulation results are shown for each stage, which are explained below, using the 23rd frame of the *Soccer* video clip.

A. Motion-Based Saliency Model

1) *Stage A: Partition Selection:* Fig. 4(a) diagrams the motion-based saliency model. The model utilizes motion intensity (speed) and the MB partition, both of which can be obtained from the motion estimation (ME) module in the hierarchical B prediction. In H.264/AVC, there exist nine different prediction modes for intra MBs and seven different prediction modes for inter MBs. The block partition of the kth MB in the nth frame, which is denoted as $P_{n,k}$, is calculated by performing rate-distortion optimization (RDO)-based coding mode selection algorithm [25], [34]–[36]. From Fig. 4(b), it may be observed that blocks having small partition size are associated with inhomogeneities, such as edges. Blocks having large partition size are associated with homogeneous regions [24], [27].

To attempt to distinguish moving objects of interest from the background, speed is employed. For the kth MB in the nth frame, the speed is given by

$$I_{n,k} = \sqrt{(MV^x_{n,k})^2 + (MV^y_{n,k})^2} \tag{1}$$

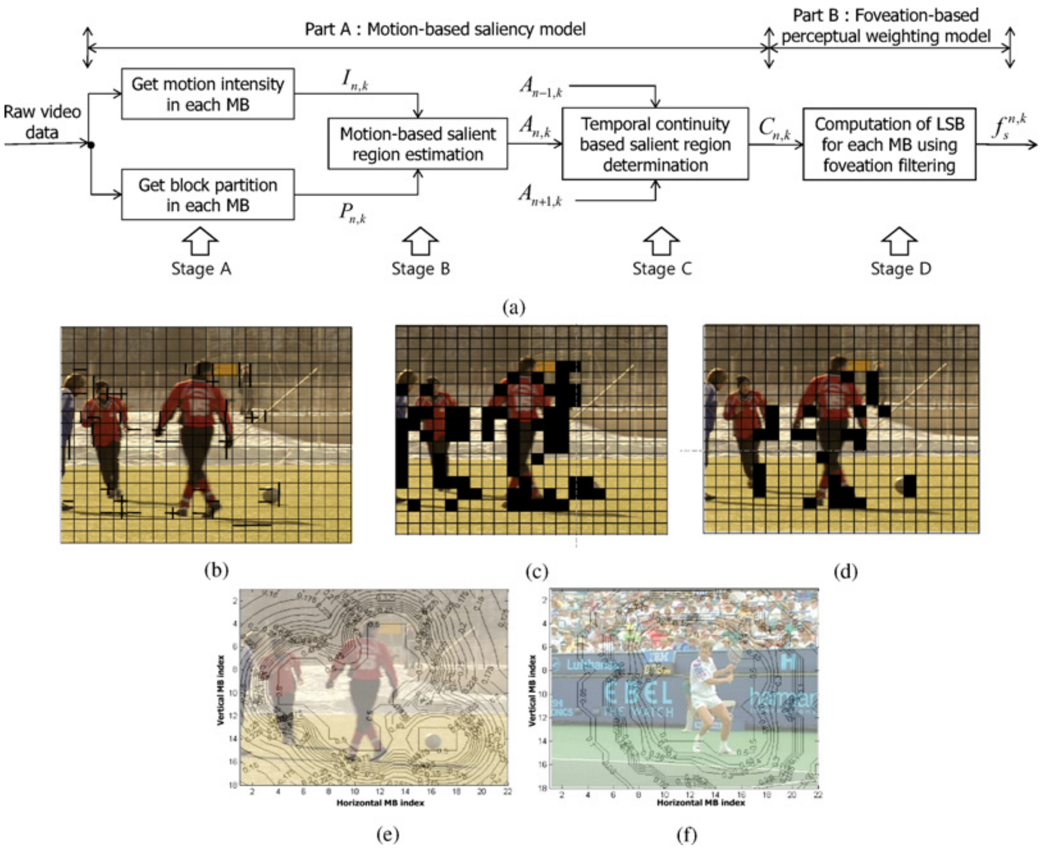
Part A : Motion-based saliency model
Part B : Foveation-based perceptual weighting model
Raw video data
Get motion intensity in each MB
Get block partition in each MB
$I_{n,k}$
$P_{n,k}$
Motion-based salient region estimation
$A_{n-1,k}$
$A_{n,k}$
$A_{n+1,k}$
Temporal continuity based salient region determination
$C_{n,k}$
Computation of LSB for each MB using foveation filtering
$f_s^{n,k}$
Stage A
Stage B
Stage C
Stage D
(a)
(b)
(c)
(d)
Vertical MB index
Horizontal MB index
(e)
(f)

Fig. 4. Block diagram and simulation results for the foveation-based perceptual weight allocator. (a) Block diagram. (b) Result of Stage A. (c) Result of Stage B. (d) Result of Stage C. (e) Result of Stage D in *Soccer* video clip. (f) Result of Stage D in *Stefan* video clip.

where $MV^x_{n,k}$ and $MV^y_{n,k}$ are the horizontal and vertical components of motion. Using $I_{n,k}$ and $P_{n,k}$, salient regions are found in each frame.

2) *Stage B: Determination of Salient Regions:* If $P_{n,k}$ indicates a small partition size, the corresponding block may occur at or near an object boundary. However, the block may also belong to the background. Using only $P_{n,k}$, it is not possible to determine whether the MB should be considered as a salient MB. The MB may be part of a salient region, provided that the speed of the MB is nonzero. However, if the speed is too large, the perceptibility of the moving MB may significantly decrease. Hence, an upper bound on the speed of the MB is required, i.e., if the MB belongs to a salient region, then its speed must fall below a threshold [23], [29], [32], [37] $0 < I_{n,k} < \Delta_k$ where Δ_k is the sum of the global mean, μ_k, and standard deviation, δ_k, of motion intensities in the kth frame as follows:

$$\Delta_k = \mu_k + \delta_k \tag{2}$$

where the motion intensity $(I_{l,k})$ is calculated from (1). In the case of an intra-coded MB, the motion intensity of the MB is set to zero since there is no motion vector. However, for *forward*, *backward*, and *bidirectional*-coding types, motion vectors for each MB are utilized for obtaining the motion intensity regardless of the direction of prediction.

If the speed of the MB falls outside of this range, then the MB is regarded as a part of a non-salient region. A binary function $A_{n,k}$ is used to indicate whether or not the kth MB belongs to a salient region; $A_{n,k}$=1 indicates saliency while $A_{n,k}$=0 indicates otherwise. The result of Stage B is exemplified in Fig. 4(c). It may be observed that edges of the moving soccer players and the moving ball are selected as candidate salient regions.

3) *Stage C: Motion Consistency:* Assuming that salient regions typically survive across consecutive frames [22], we define a binary function $C_{n,k}$ that indicates the temporal continuity of motion flow of salient objects, as follows:

$$C_{n,k} = \begin{cases} 1, & \text{if } W_{n,k} \geq S \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

where $W_{n,k} = A_{n-1,k} + A_{n,k} + A_{n+1,k}$ and $S = 2$. Based on this criterion, moving blocks having transient motion ($A_{n,k}$=1, $C_{n,k} = 0$) are considered as non-salient. This motion consistency criterion can vary according to the frame rate, e.g., by increasing S at a high frame rate and by diminishing S at a low frame rate.

The processing of Stage C is exemplified in Fig. 4(d). It may be observed that many blocks with $A_{n,k}$=1 are removed in the process, owing to their low motion consistency. Yet there remain a number of MBs that are promising salient objects.

Using $A_{n,k}$ and $C_{n,k}$, the salient region selection algorithm is summarized.

1) Calculate $I_{n,k}$ and $P_{n,k}$ for the kth MB.
2) Determine whether $P_{n,k}$ < MODE_16 × 16. If so, proceed to the next step. Otherwise, indicate that the MB is not salient by setting $A_{n,k} = 0$.

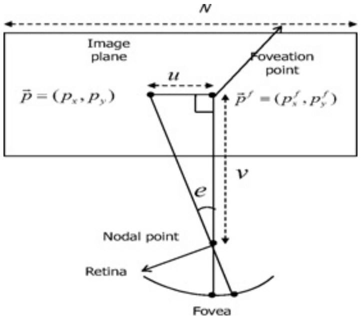

Fig. 5. Viewing geometry for the foveation model.

3) Determine whether $0 < I_{n,k} < \Delta_k$. If so, indicate that kth MB is a candidate salient MB: $A_{n,k} = 1$. Otherwise, set $A_{n,k} = 0$.
4) Determine whether $W_{n,k} > S$. If so, then kth MB is salient: $C_{n,k} = 1$. Otherwise, it is not $C_{n,k} = 0$.

The block size for the MB partition is selected by performing the RDO-based coding mode selection algorithm. Thus, the block partition is, in fact, dependent on the target bitrate. In most cases, the higher the target bitrate, the greater the number of blocks with smaller partition mode, so that the number of salient MBs increases, and vice versa. However, this is not critical to the algorithm from the point of view of efficiency and fairness. If the target bitrate is high, it is beneficial that the salient region is larger, so that fairness is guaranteed, and vice versa. Thus, the variation of block size as a function of the target bitrate is not critical to the following efficiency-fairness (EF) resource allocation algorithm. In the worst case, no MB in a frame could be selected as a salient MB at a low bitrate. In this situation, the center MBs of a frame could be candidate salient MBs based on the tendency of human fixations to seek the center of the frame side [15], [20], [21].

B. Foveation-Based Perceptual Weighting Model

Light reflected from objects in the environment passes through the optics of the eye and onto the retinal photoreceptors (cones and rods). We make the very reasonable assumption that the video is bright enough to elicit photopic vision, hence the visual input is dominated by the responses of the cones. The density of the cones and associated receptive field neurons is non-uniformly distributed across the retinal topography, peaking in density at the fovea, which also is centered on the optical axis of the eye. The point on an object surface that projects light onto the center of the fovea is correctly termed a point of visual fixation. The term *foveation point*, which we shall use, is related, but different. Since our application is presenting video images to human observers, it is presumed that the points of visual fixation fall on the displayed video being viewed. A foveation point in a video is a coordinate in space-time where it is expected that humans fixations are likely to be placed. Therefore, foveation points in video are represented with a high spatial resolution. The resolution is made to fall off systematically away from a foveation point, unless another foveation point is approached. This scheme is applicable for multiple foveation points as well. It can be seen that there are multiple foveation points in Fig. 4(d). If the LSBs for each foveation point overlap, then the larger LSB is chosen to determine the perceptual weight. In this way, the video presentation is made to have high resolution where the observers visual fixations are known or predicted to be placed, thus seeking to match the sampling capability of the retina. There has been useful work done on determining the visual resolution response (contrast sensitivity) as a function of the placement of the stimulus on the retinal relative to the fovea, which is known as the retinal eccentricity [39]–[41].

Fig. 5 diagrams a viewing geometry where $\vec{p}^f = (p_x^f, p_y^f)$ (pixels) indicates a foveation point that is also a fixation point, v is the viewing distance from the eye to the display image plane, and N is the number of pixels along the horizontal axis. The distance u from the point $\vec{p} = (p_x, p_y)$ (pixels) to the foveation point $\vec{p}^f$ is $u = d(\vec{p})/N$, where $d(\vec{p}) = \|\vec{p}^f - \vec{p}\|_2$. The vertical distance v is defined similarly. The eccentricity is then defined as the visual angle [20] as follows:

$$e(v, \vec{p}') = \tan^{-1}(\frac{u}{v}) = \tan^{-1}(\frac{d(\vec{p}')}{Nv}). \tag{4}$$

For a given eccentricity, $e(v, \vec{p})$, the local spatial cut-off frequency (cycle/degree) f_c is defined by setting the contrast sensitivity to 1.0 (the maximum possible contrast) and is calculated as follows:

$$f_c(e(v, \vec{p})) = \frac{e_2 \ln(\frac{1}{CT_0})}{\alpha(e(v, \vec{p}) + e_2)} \tag{5}$$

where CT_0 is a minimum contrast threshold, e_2 is a half-resolution eccentricity constant, and α is a spatial frequency decay constant. In this model, higher spatial frequencies than f_c are less visible or invisible.

In a displayed digital image, the effective display visual resolution r (pixels/degree) is expressed in terms of the viewing distance v (cm) and display resolution N (pixel/cm) as follows:

$$r = pv \tan(\frac{\pi}{180}) \approx Nv\frac{\pi}{180}. \tag{6}$$

The highest displayable spatial frequency is half of r as follows:

$$f_d(v) = \frac{r}{2} \approx Nv\frac{\pi}{360}. \tag{7}$$

Using (5) and (7), the local foveated bandwidth of a given MB $\vec{p}_k$ and viewing distance v is

$$\hat{f}_s^{n,k} = \min(f_c(e(v, \vec{p}_k)), f_d(v)). \tag{8}$$

The LSB is explicitly the local foveated bandwidth within which the maximum possible contrast can be recognized as a function of the distance from a foveation point. Thus, an MB will have a lower LSB when it is further from the foveation point. Likewise, high frequency coefficients tend to be more severely quantized as the QP is increased. Therefore, in Section IV, we propose a scheme to adjust local QPs as a function of the LSB.

1) *Stage D: Computation of LSB Using Foveation Model:* As shown in Fig. 4(d), there can be multiple foveation points. Thus, multiple LSBs overlap. In this situation, the largest LSB is chosen to determine the perceptual weight as follows:

$$f_s^{n,k} = \max_f \left\{ \acute{f}_s^{n,k}\left(d^{f,k}\right)\right\} \tag{9}$$

where the local foveated bandwidth is expressed as a function of the distance, $d^{f,k}$, between the fth foveation point and the kth MB. Using $f_s^{n,k}$, the LSB is obtained for each MB as shown in Fig. 4(e). In Fig. 4(f), the 6th *Stefan* video test clip is additionally shown as a result. This process makes it possible to eliminate visual redundancy from non-salient regions to improve coding efficiency.

2) *Computational Complexity:* The foveation-based perceptual weighting model consists of Stages A–D. In Stage A, the root mean square (RMS) value of directional motion estimates is computed on each MB. This value depends on the motion partition. The 4 × 4 sub-block is the smallest for a 16 × 16 MB, in which case the MB consists of 16 sub-blocks. Suppose that all the MBs of a video frame have 4 × 4 sub-blocks. In this case, the maximum number of RMS values to be computed is $16K$ where K is the total number of MBs in the frame. After Stage A is finished, Stages B and C perform simple threshold comparisons to determine the saliency regions of all MBs. Based on the selected foveation points, the LSB is then computed. Since the LSB can be obtained using table look-up there is negligible computational overhead.

IV. Optimal Resource Allocation: Efficiency and Fairness

The LSB of each MB can be obtained using the foveation principles outlined in the preceding. In order to adjust video quality as a function of the LSB, the QP is adjusted according to the saliency of each MB. Here, resource allocation cross the layers is developed considering efficiency.

Now let $r_k(q_k)$, $d_k(q_k)$, and q_k be the rate, distortion, and QP of the kth MB. Let M be the number of MBs in a frame, and R_T^n be the number of target bits for a frame. Given QP q_k for the kth MB, then $r_k(q_k)$ and $d_k(q_k)$ are calculated for each MB using the RDO algorithm in the reference software SVC H.264/AVC [12]. The QPs for coding the M MBs then compose a quantization state vector $\vec{Q} = (q_1, q_2, \cdots, q_M)$. Given the above, we propose an optimal resource allocation algorithm for each quality layer in the following. For brevity, we drop the dependence on v in $f_s^{n,k}(v)$ in (7) and express it as $f_s^{n,k}$.

We assume a viewing distance, $v = 40\,\text{cm}$. If the viewing distance is increased, then the LSB decreases more slowly from the foveation point, and vice versa. At a greater distance, a larger region is covered by the angle subtended by the fovea, causing the sensitivity to be more uniform over a larger part of the image. In this case, fairness increases in importance (and vice versa). This is reflected in the design of the LSB model (5).

A. *EF* Resource Allocation Algorithm

Here we introduce the EF algorithm that seeks to achieve a desirable tradeoff between efficiency and fairness in each quality layer. If efficiency is only considered at the expense of fairness, then MBs having large LSB will obtain the most encoding bits. Other MBs having low LSB will likely have very poor perceptual quality. The more efficient the scheme becomes, the more unfair the enhancement quality layers will become. While this is the idea of our scheme, it is still important to carefully design the resource allocator to mediate efficiency while maintaining an appropriate modicum of fairness at each quality layer.

The quantity $f_s^{n,k}$ in (7) is used to determine the size of the salient regions in each layer. For a given efficiency level l, let $EF(l)$ denote the threshold above which $f_s^{n,k}$ must fall for an MB to fall within a salient region, and vice versa. Fig. 6(a) shows the layered mechanism of the proposed VSVC framework.

Using efficiency level 9 in Fig. 6(a), the highest efficiency is achieved by setting $EF(9) = EF_{max}$, which is the largest possible LSB. This makes it possible to maintain optimal perceptual quality over the smallest salient region. Using efficiency level 4 in Fig. 6(a), both efficiency and fairness are moderately realized by assigning an intermediate value of the LSB to $EF(4)$, which leads to an enlarged salient region. Finally, using efficiency level 0 in Fig. 6(a), increased fairness can be assured by assigning a lower value $EF(0)$. This makes the salient region as large as possible. The LSB of the MBs within the salient region is set to the highest value. We term this scheme the EF algorithm, representing the balanced tradeoff between efficiency and fairness at each layer.

To decide the required number of salient region sizes, $f_s^{n,k}$ is mapped onto discrete levels of frequency sensitivity, which vary with the frequency indices of the transform coefficients, the coefficient magnitudes, and the block luminances [14], [23], [24], [26]. Let α and β be the indexes of 2-D transform coefficients in a block. The normalized local spatial frequency (cycle/degree) can be expressed as follows:

$$N_s(\alpha, \beta) = \frac{1}{2N}\sqrt{\alpha^2 + \beta^2} \tag{10}$$

where $N_s(\alpha, \beta)$ is normalized by 0.5 [23]. Fig. 7(a) shows the contours of $N_s(\alpha, \beta)$ as a function of the transform coefficient index. In our implementation, ten values of $N_s(\alpha, \beta)$ are used to define the size of the salient region.

To modify the perceptual weighting of each MB as a function of quality layer, $f_s^{n,k}$ is quantized using the values of $N_s(\alpha, \beta)$. Denote the quantized versions of $f_s^{n,k}$ by $\hat{f}_s^{n,k}$. The quantization process is as follows:

$$\hat{f}_s^{n,k} = \max\left\{z \in \bar{N}_s | z \le f_s^{n,k}\right\} \tag{11}$$

where $\bar{N}_s = \left\{N_s(1,1), N_s(1,2), \cdots, N_s\left(\bar{\alpha}, \bar{\beta}\right)\right\}$. $\bar{\alpha}$ and $\bar{\beta}$ are the maximum transform coefficient indices. In SVC H.264/AVC, $\bar{\alpha} = 4$ and $\bar{\beta} = 4$. For example, if $f_s^{n,k} = 0.19$, then $\hat{f}_s^{n,k} = 0.17$. This follows since for the next level $N_s(\alpha, \beta) = 0.25$, the nearest discrete value of $f_s^{n,k}$ is 0.17. Fig. 7(b) shows the contours of $\hat{f}_s^{n,k}$ derived from $f_s^{n,k}$ by using (11).

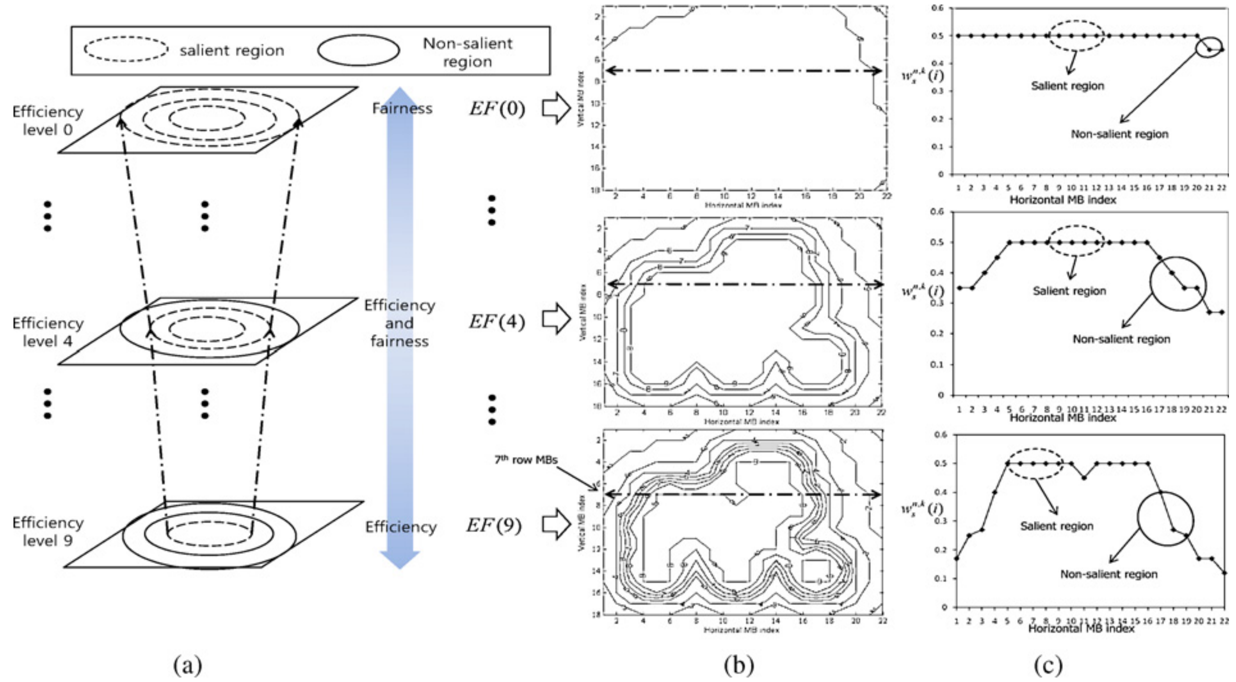

Fig. 6. Proposed VSVC framework. (a) Distribution of resources to achieve efficiency and fairness across layers in the proposed VSVC framework. (b) Illustration of variation of salient region size and shape across the quality layers. (c) Distribution of $w_s^{n,k}(i)$ in the MBs in an arbitrary image row (seventh row).

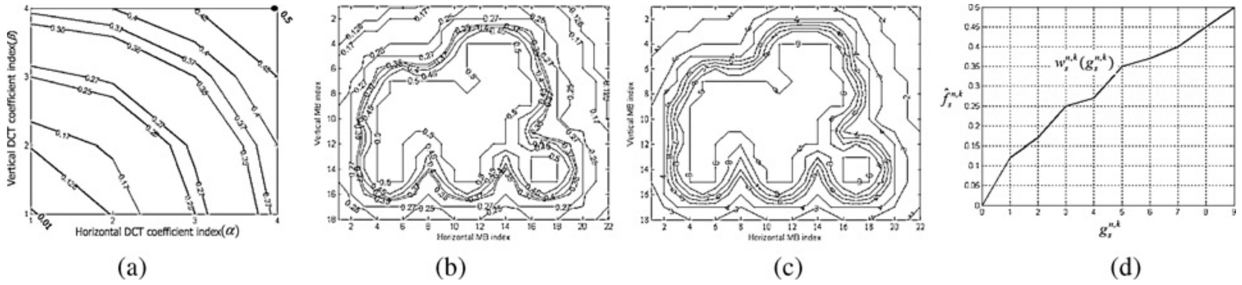

Fig. 7. Plots of coefficient indices. (a) Contours of $N_s(\alpha, \beta)$ as a function of coefficient index. (b) Contours of $\hat{f}_s^{n,k}$ from $f_s^{n,k}$. (c) Contours of $g_s^{n,k}$ from $\hat{f}_s^{n,k}$. (d) Relation between $g_s^{n,k}$ and $\hat{f}_s^{n,k}$.

The values of $\hat{f}_s^{n,k}$ are mapped onto integer values denoted $g_s^{n,k}$ in the range [0, 9] as depicted in Fig. 7(c) which is denoted as $g_s^{n,k}$. The relation between $g_s^{n,k}$ and $\hat{f}_s^{n,k}$ is defined by a weighting function $w_s^{n,k}$ as follows:

$$\hat{f}_s^{n,k} = w_s^{n,k}(g_s^{n,k}). \tag{12}$$

Fig. 7(d) shows the weighting function of $w_s^{n,k}$ in (12). If $g_s^{n,k}$ = 9, then the associated MB obtains the highest discrete value $\hat{f}_s^{n,k}$ = 0.5. At the other extreme, if $g_s^{n,k}$ = 0, then lowest value $\hat{f}_s^{n,k}$ = 0.01 is assigned to the MB.

Suppose that there are L quality layers. Then, for a given quality layer l, the perceptual weighting of each MB in the EF algorithm is fixed by defining $\hat{w}_s^{n,k}$ as follows:

$$\hat{w}_s^{n,k}(g_s^{n,k}) = \begin{cases} w_s^{n,k}(L), & \text{if } g_s^{n,k} \geq l \\ w_s^{n,k}\left(g_s^{n,k} + L - 1\right), & \text{otherwise} \end{cases} \tag{13}$$

where L = 9 in our implementation. Thus, the region with $g_s^{n,k} \geq l$ becomes salient and $\hat{w}_s^{n,k}$ is decreased in proportion to the distance from the salient region.

Fig. 6(b) and (c) depicts the operation of the EF algorithm. The *Soccer* test clip is utilized. The algorithm is illustrated for three efficiency levels $l = 0, 4, 9$ with the corresponding QPs of 42, 36, and 28, while showing the variation in the size of the salient region.

B. *EF Resource Allocation*

The resource allocation algorithm is designed to minimize the overall distortion subject to the target rate constraint using a rate control algorithm as follows:

$$\min_{\vec{Q}} \left[D(\vec{Q})\right] = \min_{\vec{Q}} \sum_{k=1}^{M} \left[\hat{w}_s^{n,k}(g_s^{n,k}) \cdot d_k(q_k)\right] \tag{14}$$

subject to (13) and

$$\text{subject to} \begin{cases} \sum_{k=1}^{M} r_k(q_k) \leq R_T \\ q_{\min} \leq q_k \leq q_{\max} \end{cases} \tag{15}$$

where $l = \{0, 1, 2, \ldots, 9\}$ and L = 9. The first constraint is the number of target bits. The second constraint represents the feasible range of variation of the QP. A large difference in the QP between adjacent MBs may result in an abrupt change in perceptual quality, which may be noticeable and visually annoying.

For a given R_T, the goal of (14) is to allocate an optimal QP for each MB to minimize the overall perceptual distortion. This problem can be solved by using a greedy algorithm. At each iteration of the greedy algorithm, each MB is evaluated to determine which can achieve the greatest (weighted) distortion reduction using the least bits, if there are encoding bits available. To conduct this evaluation, define

$$\phi_k = \frac{\hat{w}_s^{n,k}(g_s^{n,k}) \cdot (d_k(q_k) - d_k(q_k - 1))}{r_k(q_k - 1) - r_k(q_k)}. \tag{16}$$

This quantity evaluates the gradient of the reduced weighted distortion that is achieved when q_k is decremented by one. Which MB most effectively enhances the expected perceptual quality via rate control is determined as

$$\hat{k} = \arg\max_k [\phi_k]. \tag{17}$$

If $q_{max} \le q_{\hat{k}} - 1 \le q_{min}$, then the QP for the $\hat{k}$th MB is decremented by 1. Otherwise, another MB is selected that satisfies (17). The current rate allocated to MB $\hat{k}$ is updated, by setting $R_{\text{sum}} = R_{\text{sum}} + r_{\hat{k}}(q_{\hat{k}})$. This procedure is repeated until $R_{\text{sum}} \le R_T$. The iteration algorithm is summarized as follows.

1) Initialize R_T, $R_{\text{sum}} = 0$, $q_k = q_{\max}$ for all MBs.
2) For each layer l, $w_s^{n,k}(i)$ is assigned to MBs using (15).
3) Calculate ϕ_k using (16) for all MBs.
4) Find $\hat{k}$ (17) that satisfies $q_{\min} \le q_{\hat{k}} \le q_{\max}$ using (15).
5) If $R_{\text{sum}} + r_{\hat{k}}(q_{\hat{k}}) > R_T$, this procedure is terminated. Otherwise, go to Step 6.
6) If $q_{\hat{k}} < q_{\min}$, set $\phi_{\hat{k}} = 0$ and go to Step 4.
7) Update: $q_{\hat{k}} = q_{\hat{k}} - 1$ and $R_{\text{sum}} = R_{\text{sum}} + r_{\hat{k}}(q_{\hat{k}})$.
8) Calculate $\phi_{\hat{k}}$ using (16) for the $\hat{k}$th MB and go to Step 4.

This iterative procedure is continued until Step 5 is satisfied. After the procedure is terminated, $\vec{q}_k = q_k^*$ which is the optimal QP of the kth MB. The vector $\vec{Q}$ consists of the optimal QP values q_k^* for $1 \le k \le M$.

V. Simulation Results

We use the reference software SVC H.264/AVC [12] to demonstrate the effectiveness of the proposed VSVC algorithm. The proposed VSVC algorithm and the conventional SVC algorithm using frame level framework (the previous version of SVC H.264/AVC) are used for performance comparison.

Encoding configurations are as follows. The RDO mode and the loop filter are enabled. The content-based adaptive binary arithmetic coding option is turned on and variable block sizes with a search range of 32 are utilized for block ME.

For scalable video, the maximum temporal level is 5 based on groups of pictures with 16 frames. All frames except I frames are coded as B frames, which use forward and backward referencing. Let $QP_{T_{\max}}$ be the QP at the highest temporal level, which is known prior to coding. The remaining QP_Ts for the temporal levels $0 \le T < T_{\max}$ are determined by the reference software [12], [34], [35]. We consider SNR scalability to improve perceptual quality without including temporal and spatial scalability.

TABLE I
Target Bitrate and Used Bitrate Comparisons for Different Quality Layers with Conventional SVC Algorithm, EF(9), and EF(2)

	Quality	Target	Used Bitrate		
	Layer	Bitrate	Conventional	EF(9)	EF(2)
Stefan	BL	280.0	279.98	280.86	279.70
	EL	550.0	551.18	547.62	549.55
City	BL	120.0	119.77	121.05	120.70
	EL	260.0	261.13	260.70	258.50
Soccer	BL	160.0	166.80	162.24	166.96
	EL	340.0	339.08	340.12	338.21
Silent	BL	100.0	101.54	99.17	100.76
	EL	220.0	219.53	217.54	219.35

BL: base layer; EL: enhancement layer.

Thus, one spatial resolution is encoded for quality layer scalability. The parameters $q_{\max}$ and $q_{\min}$ for each temporal level T are set to $QP_T + 6$ and $QP_T - 6$, respectively.

The foveation-based perceptual weighting model is configured assuming a block size of 4 × 4 for evaluating $\hat{f}_s^{n,k}$. The following test video clips with a CIF resolution of 30 f/s were used for the performance comparison: *Soccer*, *Silent*, *Stefan*, and *City* where 300 frames are used for each test video clip. The simulation results are studied with respect to two aspects: objective perceptual quality and subjective perceptual quality.

A. Objective Visual Quality Evaluation

In this section, we evaluate the proposed VSVC algorithm at different quality layers. Two different quality layers of the EF algorithm are considered: $l = 2$ and 9, which represent pure fairness and efficiency, respectively. Two types of quality layers are considered: the base and the enhancement quality layers, where the values of initial $QP_{T_{\max}}$ are set to 42 and 36, respectively. Using the initial $QP_{T_{\max}}$, the total used bitrate is controlled to the target bitrate. Table I compares the target and used bitrates for each quality layer.

To evaluate efficiency in the base quality layer, we utilize the FPSNR first developed in [16]. Specifically, the mean square error weighted by quantized LSB $\hat{f}_s^{n,k}$ is defined here as follows:

$$\text{FMSE}_{\text{LSB}} = \frac{1}{\sum_{k=1}^{M}\sum_{j=1}^{J}(\hat{f}_s^{n,k})^2} \sum_{k=1}^{M}\sum_{j=1}^{J}(o_k(j) - r_k(j))^2 \cdot (\hat{f}_s^{n,k})^2 \tag{18}$$

where $o_k(j)$ is the jth pixel of the kth MB in the original video frame, and $r_k(j)$ is the jth pixel of the kth MB in the reconstructed image. M and J represent the total number of MBs in a frame and the total number of pixels in an MB, respectively. Then, the overall FPSNR is

$$\text{FPSNR} = 10 \cdot \log_{10} \frac{255^2}{\text{FMSE}_{\text{LSB}}}. \tag{19}$$

TABLE II

AVERAGE FPSNR AND PSNR COMPARISON FOR DIFFERENT QUALITY LAYERS WITH THE CONVENTIONAL SVC ALGORITHM, EF(9), AND EF(2)

	Scheme	*Stefan*	*City*	*Soccer*	*Silent*
Average	Conventional	30.29	31.63	31.44	31.51
FPSNR	$EF(9)$	30.97	31.93	32.06	32.33
(BL)	$EF(2)$	30.39	31.55	31.57	31.96
Average	Conventional	32.26	33.43	32.90	33.24
FPSNR	$EF(9)$	32.15	33.19	32.84	33.17
(EL)	$EF(2)$	32.56	33.71	33.05	33.65
Average	Conventional	28.01	29.64	29.82	30.53
PSNR	$EF(9)$	27.72	29.35	29.57	30.12
(BL)	$EF(2)$	27.85	29.72	29.79	30.41
Average	Conventional	30.85	32.48	31.95	33.42
PSNR	$EF(9)$	30.47	31.84	31.75	32.61
(EL)	$EF(2)$	30.61	31.95	31.81	32.81

BL: base layer; EL: enhancement layer.

Table II tabulates the average FPSNR and PSNR comparison for each quality layer. In the base layer, the $EF(9)$ scheme delivers a higher FPSNR than $EF(2)$ and the conventional SVC scheme (in the range 0.3–0.68 dB), since $EF(9)$ allocates more resources to the identified perceptually important region since there is a lack of encoding bits. As shown in Fig. 8, $EF(9)$ consistently delivers higher FPSNR values across frames. Conversely, the average traditional PSNR of $EF(9)$ is lower than that of conventional SVC for all of the test clips in Table II. This is expected, since conventional SVC utilizes a resource allocation scheme in order to lower the highest distortion for each MB in a frame. Yet this is not an effective measure of the perceptual quality.

As shown in Table II, the $EF(9)$ scheme yields the lowest average FPSNR, since it makes an effort to reduce the distortion of the smallest salient region. As compared to conventional SVC, $EF(2)$ obtains a higher FPSNR (in the range of 0.15–0.38) over all of the test video clips, indicating that $EF(2)$ provides more efficient perceptual quality improvement than the conventional algorithm. Although the average PSNR of $EF(2)$ is lower than that of conventional SVC, SVC reduces the overall distortion over the MBs without regard to the perceptual relevance of the salient region.

B. Subjective Quality Comparison

In order to conduct subjective quality comparison, we use a video test clip reconstructed by both the proposed VSVC algorithm with the EF algorithm and by conventional SVC. We present a few exemplary reconstructed pictures from different quality layers. In addition, we have made several demo sequences available for public download at [44], to enable visual comparison between the result of conventional SVC and the proposed VSVC schemes.

Fig. 8 shows a frame from the video clips reconstructed using the base (166.25 kb/s) and enhancement (198.62 kb/s) layers in the conventional SVC algorithm, where $EF(8)$ (165.37 kb/s) and $EF(4)$ (197.14 kb/s) are used for the base and enhancement layers. The 98th frame of the *City* test video clip is used.

Fig. 8(a) and (b) compares the subjective quality for the base layer. When conventional SVC was used, noticeable perceptual degradation occurs in the middle of the frame owing to the lack of adequate encoding bits. On the contrary, $EF(8)$ preserves details in the video frame better than does the conventional SVC scheme. This is a good demonstration that efficiency is more important than fairness toward improving subjective video quality, when there is a lack of encoding bits.

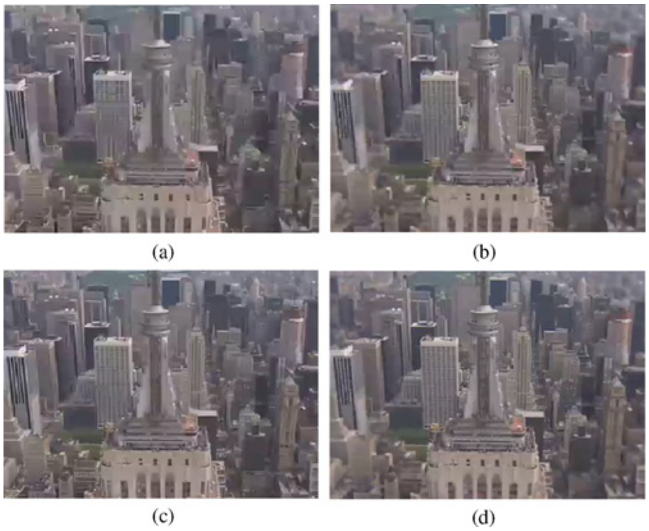

Fig. 8. Subjective quality comparisons for different quality layers on the 98th frame of the *City* test video clip. (a) Base layer in conventional SVC. (b) Base layer in $EF(8)$. (c) Enhancement layer in conventional SVC. (d) Enhancement layer in $EF(4)$.

Fig. 8(c) and (d) compares the subjective quality for the enhancement quality layer. Since $EF(8)$ allocates more encoding bits to MBs having a large perceptual weighting, a noticeable degradation of the subjective quality occurs in the remaining MBs. However, the proposed fairness scheme improves the subjective video quality more effectively than does the efficiency scheme on the enhancement quality layer. Conversely, when conventional SVC is used, noticeable perceptual degradation occurs in the middle of the frame owing to a lack of adequate encoding bits in the base layer.

Fig. 9 shows the degree of reduced distortion over the frame using conventional SVC, $EF(8)$, and $EF(4)$. The distortion is defined as the MSE between the original and reconstructed images divided by 10 000. The 98th frame of the *City* test video clip is used.

Fig. 9(a) and (b) represents the distribution of resources in conventional SVC for different quality layers. Since resource allocation in conventional SVC is based on the MSE, resources are allocated equally into all MBs in the base and enhancement quality layers.

The distribution of reduced distortions in the proposed SVC is shown in Fig. 9(c) and (d) for the different quality layers. It may be observed that $EF(8)$ allocates resources in the perceptually important region with $l = 8$ in the base quality layer. In the enhancement quality layer, $EF(4)$ enlarges the range of allocated resources in the same region with $l = 4$.

Fig. 10(a) and (b) shows the video clips reconstructed using the base layer in conventional SVC (164.43 kb/s) and in $EF(8)$ (162.25 kb/s) with an efficiency level of $l = 8$, respectively. The 16th frame of the *Stefan* test video clip is used for comparison.

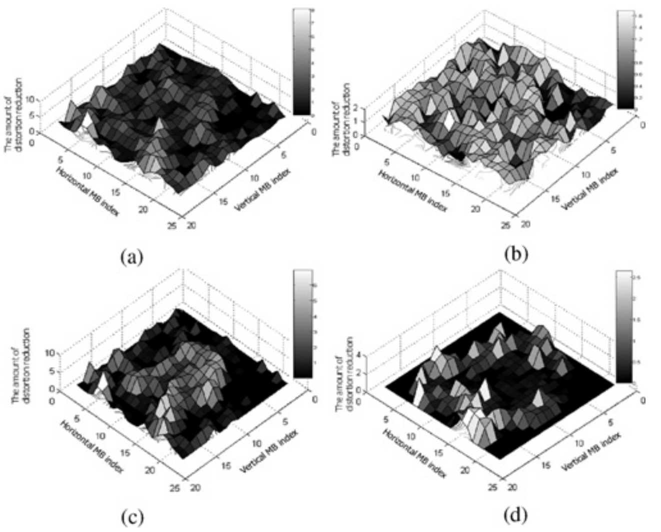

Fig. 9. Distortion and QP distribution comparisons for different quality layers on the 98th frame of the *City* test video clip. (a) Amount of reduced distortion in base layer of conventional SVC. (b) Amount of reduced distortion in the enhancement layer of conventional SVC. (c) Amount of reduced distortion in the base layer of $EF(8)$. (d) Amount of reduced distortion in the enhancement layer of $EF(4)$.

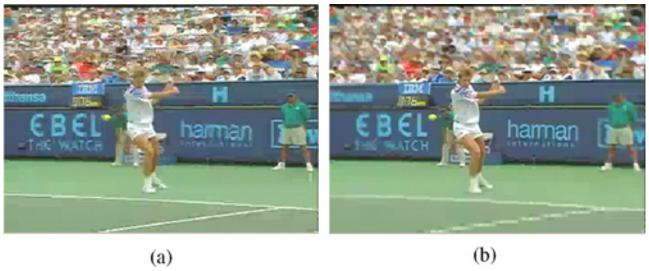

Fig. 10. Subjective quality comparisons on the 14th frame of the *Stefan* test video clip. (a) Base layer in conventional SVC. (b) Base layer in $EF(8)$.

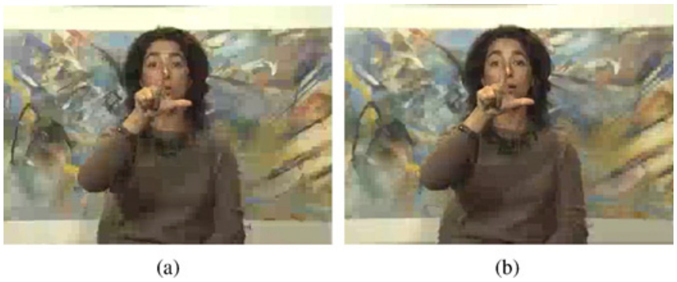

Fig. 11. Subjective quality comparisons on the 103th frame of the *Silent* test video clip. (a) Enhancement layer in conventional SVC. (b) Enhancement layer in $EF(4)$.

Using conventional SVC, a noticeable degradation of quality occurs owing to the lack of encoding bits. However, the proposed efficiency algorithm allocates more resources into the salient region to improve the visual quality. While the spectators in the background are more distorted, the tennis player, which has been assigned high perceptual importance, is maintained with higher quality.

Fig. 11(a) and (b) shows the video clips reconstructed using the enhancement layer in the conventional SVC (131.40 kb/s) and in $EF(4)$ (130.11 kb/s) with an efficiency level of $l = 4$, respectively. The 103rd frame of the *Silent* test video clip is used.

Based on the proposed motion-based saliency model, the region where hand movement for sign language is selected as the salient region. While the enhancement layer in $EF(4)$ focuses on the quality improvement of the selected salient region, conventional SVC still assigns more bits into the background. Consequently, the visual quality in $EF(4)$ is better than that of conventional SVC owing to the salient region-based resource allocation. The proposed fairness scheme in VSVC effectively inhibits perceptual degradations in areas of identified perceptual importance by allocating increased resources to those MBs.

From these visual quality comparisons and the FPSNR analysis, it is apparent that the perceptual weighting scheme of the proposed SVC can efficiently improve the subjective quality of H.264 compressed videos in identified perceptually important regions.

VI. Conclusion

We have proposed an MB-level perceptual weighting framework as an extension to SVC in H.264/AVC. To enable adaptation to the non-uniform resolution of the visual photoreceptors, we developed a foveation-based perceptual weighting allocator. The proposed VSVC scheme was developed based on two essential objectives, namely, enforcing efficiency and fairness in the quality layer to improve coding performance. To find an optimal tradeoff between efficiency and fairness, we proposed the EF resource allocation algorithm, which preferentially allocates coding resources to salient regions. The size of the salient regions is increased as the efficiency level is raised.

The simulations showed that the VSVC algorithm achieves higher perceptual quality using FPSNR than does conventional SVC by 0.3–0.8 dB in the different quality layers. In summary, it can be concluded that the proposed VSVC algorithm is a promising solution for achieving high-quality and reliable video communications over variable rate channels with good control of QoS. Such algorithms that emphasize perceptual quality as a function of visual importance, saliency, computed or measured fixations, or other similar criteria are likely to continue growing in importance, owing to the ongoing increases in display sizes and video bandwidths, expectations for higher video quality, and generally, increased ubiquity of video in our daily environment.

References

[1] ITU-T and ISO/IEC JTC 1, *Generic Coding of Moving Pictures and Associated Audion Information, Part 2: Video*," document ITU-T Rec. H.262 and ISO/IEC 13818-2 (MPEG-2 Video), Nov. 1994.

[2] ITU-T and ISO/IEC JTC 1, *Advanced Video Coding for Generic Audio-Visual Services*," document ITU-T Rec. H.264 and ISO/IEC 14496-10 (MPEG-4 AVC), May 2003.

[3] H. Schulzrinne, S. Casner, R. Frederick, and V. Jacobson, "RTP: A transport protocol for real-time application," document RFC 1889, Internet Official Protocol Standards (STD 1), Jan. 1996.

[4] *Text of ISO/IEC 14496-10:2005/FDAM 3 Scalable Video Coding*, document N9197, Joint Video Team (JVT) of ISO-IEC MPEG and ITU-T VCEG, Lausanne, Switzerland, Sep. 2007.

[5] ITU-T and ISO/IEC JTC 1, *Advanced Video Coding for Generic Audio Visual Services*, ITU-T Recommendation H.264 and ISO/IEC 14496-10 (MPEG-4 AVC), Version 1: May 2003, Version 2: May 2004, Version 3: Mar. 2005, Version 4: Sep. 2005, Version 5 and Version 6: Jun. 2006, Version 7: Apr. 2007, Version 8 (including SVC extension): consented in Jul. 2007.

[6] *Text of ISO/IEC 14496-4:2001/PDAM 19 Reference Software for SVC*, document N9195, Joint Video Team (JVT) of ISO-IEC MPEG and ITU-T VCEG, Sep. 2007.

[7] S. J. Choi and J. W. Woods, "Motion-compensated 3-D subband coding of video," *IEEE Trans. Image Process.*, vol. 8, no. 2, pp. 155–167, Feb. 1999.

[8] H. M. Radha, M. V. D. Schaar, and Y. Chen, "The MPEG-4 find-grained scalable video coding method for multimedia streaming over IP," *IEEE Trans. Multimedia*, vol. 3, no. 1, pp. 53–68, Mar. 2001.

[9] Y. W. Chen and W. A. Pearlman, "Three-dimensional subband coidng of video using the zerotree method," *Proc. SPIE*, vol. 2727, pp. 1302–1312, Mar. 1996.

[10] J. Park, H. Lee, S. Lee, and A. C. Bovik, "Optimal channel adaptation of scalable video over a multi-carrier based multi-cell environment," *IEEE Trans. Multimedia Special Issue*, vol. 11, no. 6, pp. 1062–1071, Oct. 2009.

[11] U. Jang, H. Lee, and S. Lee, "Optimal carrier loading control for the enhancement of visual quality over OFDMA cellular networks," *IEEE Trans. Multimedia*, vol. 10, no. 6, pp. 1181–1196, Oct. 2008.

[12] J. Reichel, H. Schwarz, M. Wien, and J. Vieron, *Joint Scalable Video Model 9 of ISO/IEC 14496-10:2005/AMC3 Scalable Video Coding*, document JVT-X202, Joint Video Team (JVT) of ISO-IEC MPEG and ITU-T VCEG, Geneva, Switzerland, Jul. 2007.

[13] J. Reichel, M. Wien, and H. Schwarz, "Scalable video model 3.0," document N6716, ISO/IEC JTC 1/SC 29/WG 11, Palma de Mallorca, Spain, Oct. 2004.

[14] B. Ciptpraset and K. R. Rao, "Human visual weighting progressive image transmission," in *Proc. ICCS*, Nov. 1987, pp. 1040–1044.

[15] S. Lee, M. S. Pattichis, and A. C. Bovik, "Foveated video quality assessment," *IEEE Trans. Multimedia*, vol. 4, no. 1, pp. 129–132, Mar. 2002.

[16] S. Lee, M. S. Pattichis, and A. C. Bovik, "Foveated video compression with optimal rate control," *IEEE Trans. Image Process.*, vol. 10, no. 7, pp. 977–992, Jul. 2001.

[17] S. Lee and A. C. Bovik, "Fast algorithms for foveated video processing," *IEEE Trans. Circuit Syst. Video Technol.*, vol. 13, no. 2, pp. 149–162, Feb. 2003.

[18] S. Lee, A. C. Bovik, and Y. Y. Kim, "High quality, low delay foveated visual communications over mobile channels," *J. Visual Commun. Image Representat.*, vol. 16, no. 2, pp. 180–211, Apr. 2005.

[19] H. Lee and S. Lee, "Compression gain measurements by using ROI-based data reduction," *IEICE Trans. Fundamentals*, vol. E89-A, no. 11, pp. 2985–2989, Nov. 2006.

[20] Z. Wang and A. C. Bovik, "Embedded foveation image coding," *IEEE Trans. Image Process.*, vol. 10, no. 10, pp. 1397–1410, Oct. 2001.

[21] Z. Wang, L. Lu, and A. C. Bovik, "Foveation scalable video coding with automatic fixation selection," *IEEE Trans. Image Process.*, vol. 12, no. 2, pp. 243–254, Feb. 2003.

[22] K. Yang, C. C. Guest, K. E. Maleh, and P. K. Das, "Perceptual temporal quality metric for compressed video," *IEEE Trans. Multimedia*, vol. 9, no. 7, pp. 1528–1535, Nov. 2007.

[23] C. Tang, "Spatiotemporal visual considerationfor video coding," *IEEE Trans. Multimedia*, vol. 9, no. 2, pp. 231–238, Feb. 2007.

[24] J. Chen, J. Zheng, and Y. He, "Macroblock-level adaptive frequency weighting for perceptual video coding," *IEEE Trans. Consumer Electron.*, vol. 53, no. 2, pp. 775–781, May 2007.

[25] J. Lee, "Rate-distortion optimization of parameterized quantization matrix for MPEG-2 encoding," in *Proc. ICIP*, Oct. 1998, pp. 383–386.

[26] Y. F. Ma and H. J. Zhang, "A model of motion attention for video skimming," in *Proc. ICIP*, vol. 1. 2002, pp. I-129–132.

[27] C. W. Ngo, T. C. Pong, and H. J. Zhang, "Motion analysis and segmentation through spatio-temporal slices processing," *IEEE Trans. Image Process.*, vol. 12, no. 3, pp. 341–355, Mar. 2003.

[28] H. B. Yin, "Adaptive quantization in perceptual MPEG video encoders," in *Proc. PCS*, Apr. 2006, pp. 3–14.

[29] R. A. Rensink, "A model of saliency-based visual attention for rapid scene analysis," in *Proc. ACM 2nd Int. Symp. Smart Graphics*, 2002, pp. 63–70.

[30] A. B. Watson, G. Y. Yang, J. A. Solomon, and J. Villasenor, "Visibility of wavelet quantization noise," *IEEE Trans. Image Process.*, vol. 6, no. 8, pp. 20–29, Aug. 1997.

[31] A. B. Watson, J. Hu, and J. F. McGowan, III, "DVQ: A digital video quality metric based on human vision," *J. Electron. Imag.*, vol. 10, no. 1, pp. 1164–1175, Aug. 1997.

[32] R. A. Rensink, J. K. O. Regan, and J. J. Clark, "To see or not to see: The need for attention to perceive changes in scenes," *Psychol. Sci.*, vol. 8, pp. 368–373, Sep. 1997.

[33] B. Girod, "What's wrong with mean-square error?" in *Digital Images Human Vision*, A. B. Watson, Ed. Cambridge, MA: MIT Press, 1993.

[34] H. Schwarz, D. Marpe, and T. Wiegand, "Overview of the scalable video coding extension of the H.264/AVC standard," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 17, no. 9, pp. 1103–1120, Sep. 2007.

[35] H. Schwarz, D. Marpe, and T. Wiegand, "Analysis of hierarchical B pictures and MCTF," in *Proc. ICME*, Jul. 2006, pp. 1929–1932.

[36] I. Amonou, N. Cammas, S. Kervadec, and S. Pateux, "Optimized rate-distortion extraction with quality layers in the scalable extension of H.264/AVC," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 17, no. 9, pp. 1186–1193, Sep. 2007.

[37] L. Itti, "Quantifying the contribution of low-level saliency human eye movement in dynamic scenes," *Vis. Cognit.*, vol. 12, no. 6, pp. 1093–1123, Aug. 2005.

[38] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Patt. Anal. Mach. Intell.*, vol. 20, no. 11, pp. 1254–1259, Nov. 1998.

[39] W. S. Geisler and J. S. Perry, "A real-time foveated multiresolution system for low bandwidth video communication," *Proc. SPIE*, vol. 3299, pp. 294–305, Jul. 1998.

[40] J. G. Robson and N. Graham, "Probability summation and regional variation in contrast sensitivity across the visual field," *Vis. Res.*, vol. 21, no. 3, pp. 409–418, 1981.

[41] M. S. Banks, A. B. Sekuler, and S. J. Anderson, "Peripheral spatial vision: Limited imposed by optics, photoreceptors and receptor pooling," *J. Opt. Soc. Am.*, vol. 8, pp. 1775–1787, Nov. 1991.

[42] *Final Report from the Video Quality Experts Group on the Validation of Objective Quality Metrics for Video Quality Assessment*. (2008) [Online]. Available: http://www.its.bldrdoc.gov/vqeg/projects/multimedia

[43] ITU-T, *Subjective Video Quality Assessment Methods for Multimedia Applications*, document Rec. ITU-T P.910, 2002.

[44] Demo Sequences. (2010). *Perceptually Scalable Extension of H.264* [Online]. Available: ftp://wireless.yonsei.ac.kr

Hojin Ha received the B.S. degree from Myongji University, Yongin, Korea, in 1998, the M.S. degree from Hanyang University, Ansan, Korea, in 2000, both in control and instrumentation engineering, and the Ph.D. degree in electrical and electronic engineering from Yonsei University, Seoul, Korea, in 2009.

Since 2000, he has been a Research Engineer with Digital Media and Communications (DMC) Research and Development Center, Samsung Electronics, Suwon, Korea. His current research interests include multimedia communications, multimedia signal processing, peer-to-peer networking, and hypertext transfer protocol adaptive streaming.

Jincheol Park was born in Korea in 1982. He received the B.S. degree in information and electronic engineering from Soongsil University, Seoul, Korea, in 2006, and the M.S. degree in electrical and electronic engineering from Yonsei University, Seoul, in 2008. Since 2008, he is pursuing the Ph.D. degree from the Wireless Network Laboratory, Yonsei University.

From June 2010 to June 2011, he was a Visiting Researcher in the Laboratory for Image and Video Engineering with Prof. A. C. Bovik at the University of Texas at Austin, Austin. His current research interests include wireless multimedia communications and video quality assessment.

Sanghoon Lee (M'05) was born in Korea in 1966. He received the B.S. degree in electrical engineering from Yonsei University, Seoul, Korea, in 1989, the M.S. degree in electrical engineering from the Korea Advanced Institute of Science and Technology, Daejeon, Korea, in 1991, and the Ph.D. degree in electrical engineering from the University of Texas, Austin, in 2000.

From 1991 to 1996, he was with Korea Telecom, Seongnam, Korea. From June 1999 to August 1999, he was with Bell Labs, Lucent Technologies, Seoul, working on wireless multimedia communications. From February 2000 to December 2002, he worked on developing real-time embedded software and communication protocols for 3G wireless networks, Lucent Technologies. In March 2003, he joined the faculty of the Department of Electrical and Electronics Engineering, Yonsei University, where he is currently an Associate Professor. His current research interests include image/video quality assessments, wireless multimedia communications, multihop sensor networks, and 4G wireless networks.

Dr. Lee is an Associate Editor of the IEEE Transactions on Image Processing and an Editor of the *Journal of Communications and Networks*. Currently, he is the Chair of the IEEE Standard Working Group for 3-D quality assessment.

Alan Conrad Bovik (S'80–M'81–SM'89–F'96) is the Curry/Cullen Trust Endowed Chair Professor with the University of Texas at Austin, Austin, where he is also the Director of the Laboratory for Image and Video Engineering. He is a Faculty Member with the Department of Electrical and Computer Engineering and the Center for Perceptual Systems in the Institute for Neuroscience. His current research interests include image and video processing, computational vision, and visual perception. He has published over 500 technical articles in these areas and holds two U.S. patents. His several books include the most recent companion volumes, *The Essential Guides to Image and Video Processing* (New York: Academic, 2009). He is a registered Professional Engineer in the State of Texas and is a frequent consultant to legal, industrial, and academic institutions.

He was named the SPIE/IS&T Imaging Scientist of the Year for 2011. He has also received a number of major awards from the IEEE Signal Processing Society, including the Best Paper Award in 2009, the Education Award in 2007, the Technical Achievement Award in 2005, and the Meritorious Service Award in 1998. He received the Hocott Award for Distinguished Engineering Research at the University of Texas at Austin, the Distinguished Alumni Award from the University of Illinois at Urbana-Champaign, Urbana, in 2008, the IEEE Third Millennium Medal in 2000, and two journal paper awards from the International Pattern Recognition Society in 1988 and 1993. He is a Fellow of the Optical Society of America, a Fellow of the Society of Photo-Optical and Instrumentation Engineers, and a Fellow of the American Institute of Medical and Biomedical Engineering. He has been involved in numerous professional society activities, including the Board of Governors of the IEEE Signal Processing Society from 1996 to 1998, the Co-Founder and Editor-in-Chief of the IEEE Transactions on Image Processing from 1996 to 2002, an Editorial Board Member of the *Proceedings of IEEE* from 1998 to 2004, the Series Editor for *Image, Video, and Multimedia Processing* (San Mateo, CA: Morgan and Claypool) from 2003 to present, and the Founding General Chairman, 1st IEEE International Conference on Image Processing, held in Austin, TX, in November 1994.