ICICS-PCM 2003
15-18 December 2003
Singapore

User-Oriented Approach in Spatial and Temporal Domain Video Coding

Chia-Chiang Ho, Wei-Ta Chu, Chen-Hsiu Huang and Ja-Ling Wu

Communication and Multimedia Laboratory
Department of Computer Science and Information Engineering
National Taiwan University

Abstract

Present scalable coding techniques focus on enhancing the quality of the whole frame when more bandwidth is available. However, not all regions in a frame are equally important to user perception. This paper proposed a user-oriented approach to facilitate scalable coding in spatial and temporal domains. By combining user attention and foveation models, the regions of user's focus in a frame (spatial domain) or the segment of a video clip with steady saliency values (temporal domain) are prioritized as far as perceived quality is concerned.

1. Introduction

One of the challenges to video encoding is in reducing storage or transmission bandwidth required by the compressed bitstream, while preserving mostly the perceived quality. Various human visual system (HVS) properties had been adopted to develop specific coding techniques to face such a challenge. Typical video encoding schemes, including normal video coding and scalable coding, treat different parts of the source video as being equally important, and try to provide a constant quality across all these parts. However, we know that when viewing one video sequence, people often pay more attention, consciously or unconsciously, to some specific visual object than others. Such an attracting object, for example, may be the leading actor or actress of a play, a violent car in a car flow, or a colorful parrot in a dim rainforest. This "focusing effect" is what we called "user attention". On the other hand, rigid psychological experiments have confirmed that the HVS has a property of nonuniform spatial resolution, with respect to the attentional focus in sight. This property is called "foveation". By combining user attention and foveation techniques, we develop a scalable coding scheme that preserves perceived quality as far as possible when the limit of transmission bandwidth or storage becomes stricter.

Based on MPEG-4 FGS coding scheme [1], foveation model is exploited to perform spatially selective enhancement, and user attention model is used to facilitate temporal scalability. This approach envisions scalable coding by two ways:

- **Spatial scalability:** With the concept of base layer and enhancement layer in MPEG-4 FGS scheme [1], when encoding the base layer, we use pre-filtered foveated frames that suppress quality of those regions users pay less attention to. The differences between original video frames and foveated frames are encoded as enhancement layer. When the available bandwidth is fixed, we could allocate more bits to the attention regions and thus provide better perceived quality.
- **Temporal scalability:** Different frames may have different attraction to human beings. With the help of temporal domain user attention model, we calculate importance values of different frames and skip comparably unimportant frames when the transmission bandwidth is not enough. An adaptive threshold mechanism is applied to meet different situations.

This paper is organized as follows: In section 2, the user attention model and the foveation model are introduced. Then, a novel spatial and temporal scalable coding scheme is proposed in section 3. Section 4 describes some subjective and objective experimental results. Finally, concluding remarks are given in section 5.

2. The User Attention Model and The Foveation Model

2.1 User Attention Model

Attention refers to the ability of one human to focus and concentrate upon some visual or auditory 'object', by careful observing or listening. Assuming limited processing resources of one human, attention also refers to the allocation of these resources. Here, the resource can refer to either neurological or cognitive resource. The former is often referred as *bottom-up attention* and the later *top-down attention*. Bottom-up attention [5] models what people are attracted to see. Saliency of early visual features is computed to form a set of feature maps. On the other hand, top-down attention was usually modeled by detecting some meaningful (semantic) objects or video features.

As far as video coding (or video streaming) is concerned, modeling user attention may help for reducing required storage resources (or network bandwidth), or improving user perception. We propose a bottom-up approach [2],

0-7803-8185-8/03/$17.00 © 2003 IEEE

which integrates color and intensity (low level), motion (medium level), and face (high level) saliency maps, to model user attention. For features in different levels, various saliency maps are generated to capture users' focus. Then a priority-based or linear combination rule is applied to fuse them into one integrated saliency map. This approach serves as the fundamental basis to find the foveation point.

2.2 Foveation Model

Early vision systems lean on the anatomy of the human retina. The light that reflects objects is imaged onto the retina by the lens. The retina, which consists of three layers of neurons (photoreceptor, bipolar and ganglion), is responsible for detecting the light from these images. The brain decodes these images into information that we know as vision. There are two kinds of photoreceptors: rods and cones. Cone cells are responsible for daylight vision. The density of con cells is higher at the fovea and drops significantly with increasing eccentricity (i.e., the viewing angle). Photoreceptors deliver data to bipolar cells and then to ganglion cells. The distribution of ganglion cells is also highly non-uniform. Figure 1 shows the relationship between cell densities and eccentricity (cited from [9]).

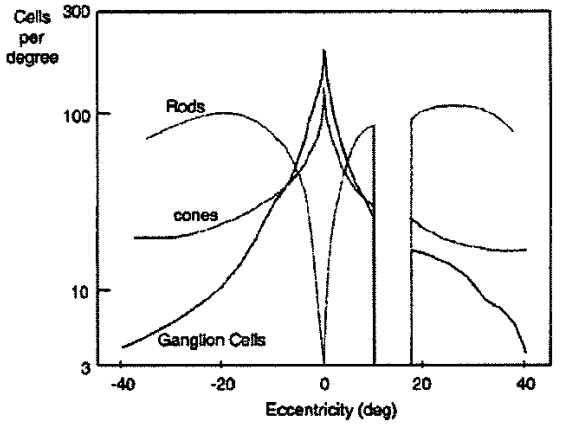

Figure 1: Photoreceptors and ganglion cell density as a function of eccentricity.

These non-uniform distributions correspond to the non-uniform spatial resolution ability of human eyes, confirmed by rigid psychological experiments (e.g., [6][7][8]). Specifically, it is found that the sampling density and contrast sensibility over an image are found decreasing dramatically with increasing viewing angle, with respect to the point human eyes gaze at. Foveation, one kind of HVS model, is thus developed to model this property.

In [4], an experimental-proven foveation model was presented. The contrast threshold of human eyes was considered as:

$$CT(f,e) = CT_0 \exp(\alpha f \frac{e_2 + e}{e}),$$

where f is the spatial frequency (cycle/degree), e is the eccentricity (degrees), CT_0 is the minimal contrast threshold, α is the spatial frequency decay constant, and e_2 is the half-resolution eccentricity constant (degrees).

We considered the visibility of the DCT-domain foveation model in [3] based on existed experimental results and proposed the *critical amplitude,* which is found by multiplying the maximal value of the *(m,n)*-th coefficient $A^*_{m,n}$:

$$A_c(f_{m,n}, e) = \frac{A^*_{m,n} CT_0}{r + (1-r)\cos^2\theta_{m,n}} \exp\left(\frac{\alpha(e_2 + e)\sqrt{m^2 + n^2}}{2e_2 Nw}\right)$$

It means that if the *(m,n)*-th DCT coefficient of one block with eccentricity e smaller than $A_c(f_{m,n}, e)$, for human eyes, it is indistinguishable from zero. In another point of view, we deduce the *critical eccentricity* $e_c(m,n)$ as:

$$e_c(m,n) = \frac{2e_2 Nw}{\alpha\sqrt{m^2 + n^2}} \ln\left(\frac{r + (1-r)\cos^2\theta_{m,n}}{CT_0}\right) - e_2$$

For one block with center point *(x,y)*, we can thus have the critical condition for each *(m,n)*-th DCT coefficient: $e(x,y) > e_c(m,n)$; and if it is true, no matter how large the *(m,n)*-th DCT coefficient is, for human eyes, it is indistinguishable from zero.

To combine the deduced foveation model with standardized image or video coding schemes, we develop two foveation filters that act on DCT blocks: *amplitude threshold based* and *breakpoint based foveation filters.* Generally, the block refers to an 8x8 pixel area. The foveation point is restricted to be the center point of one particular block, which is denoted as the 'foveation block'. When applying the amplitude threshold based filter, for each DCT block in the image, coefficients (except DC) smaller than their corresponding critical amplitudes are set to zero. On the other hand, breakpoint based filter eliminates a series of DCT coefficients at the end of that block, in zigzag scanning order.

To ease heavy computation incurred, we can calculate critical amplitudes (or critical eccentricities) in advance, and generate *amplitude threshold maps* (or *breakpoint maps*) for different viewing distances. In this way, the required computation at runtime can be reduced largely.

Foveation model can be considered as a form of the region-of-interest concept. Many researches have been devoted to object segmentation and object-based video processing, in the last decade. However, automatic mechanisms are considered hard to achieve for general usage. Foveation model implicitly alleviates the object boundary restriction,

and we think it may be a compromising mechanism for object-based applications.

3. User-Oriented Video Coding

3.1 Spatial Domain Approach

With the help of the foveation and user attention model, we propose an architecture of the so-called user-oriented video encoder (as shown in Figure 2), which is expected to improve the performance of video encoding in spatial domain.

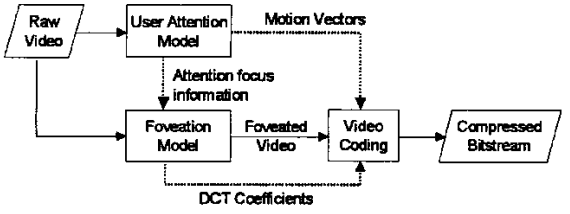

Figure 2: The proposed architecture for the user-oriented video encoding.

First, the input raw video goes through the focus detection module, which is built based on the prescribed attention model, to find out focus point(s) of each video frame. According to the information of focus point(s), the raw video goes through the foveation filter. All 8x8 blocks in an input frame will go through DCT, foveation filter, and finally IDCT modules to get the foveated frames. The foveated frames are then sent to the video encoder for normal video encoding.

Since motion estimation has been performed when calculating motion saliency map (the medium level of user attention model [2]), the related motion vector information can be sent to the encoder to avoid or lessen the burden of motion estimation. Moreover, intermediate foveated DCT blocks after the foveation filter is also sent to the video encoder. When the encoder chooses intra mode for particular frames or macroblocks, no additional DCT operations are needed.

Based on this architecture, we combine foveation model and user attention to facilitate both non-scalable and scalable coding schemes in the spatial domain:

- **Non-scalable Coding:** Under the constraint of the foveation model, the encoder can discard unimportant visual information as much as possible. In this way, the compression gain can be increased without sacrificing perceived quality of the compressed video. For the purpose of reducing storage requirement, this scenario effectively makes benefits.

- **Scalable Coding:** The encoder can selectively preserve higher quality for those focused regions, in trade of worse quality for periphery regions, to maximally match users' expectation. This scenario can be applied to generate the base layer bitstream of a scalable video when the bitrate constraint is very strict. The block diagram of the proposed foveation model based scalable coding is shown in Figure 3. The difference between the original video and the foveated video is then compressed as enhancement bitstream(s). When more bandwidth is available, the streaming server can enhance video quality by selectively adding enhancement layers according to the foveation model. The extra bitstream can be added to the regions that catch the user's attention, instead of uniformly enhancing the whole frame.

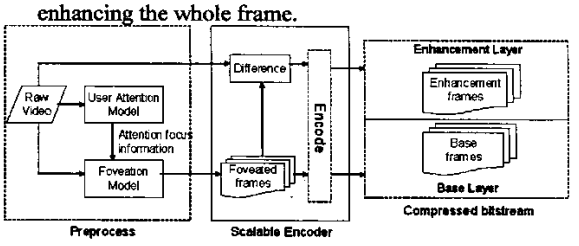

Figure 3: Foveation model based scalable coding on the base layer

In our experiments, we found the choice of foveation parameters has different considerations for non-scalable and scalable encoding schemes. Thus we define a *cut-off eccentricity*, say e_{cutoff}, and perform foveation filtering with different foveation parameters on different regions. We can thus define that the central region is composed of those blocks with eccentricities equal to or smaller than e_{cutoff}, and the periphery region is composed of those blocks with eccentricities larger than e_{cutoff}. After filtering out the region with larger eccentricity, the base layer contains more information for the central region than that of the periphery region, under the very limited bitrate budget.

3.2 Temporal Domain Approach

The proposed user attention model constructs saliency maps for video frames base on various visual features, including color, intensity, motion, and face regions. With a weighted linear combination, the user attention model fuses saliency maps of different characteristics and computes the saliency value with respect to user perception. For example, the video segments with violent motion or face may attract more attention. Therefore, not only the spatial domain coding could be achieved by foveation model, temporal scalability is another approach to reduce the required bit-rate.

In the proposed user attention model [2], the value of each pixel in the combined saliency map is normalized to yield a dynamic range of [0, 255]. Let P_i denote the value of the i-th pixel, and *Score* denote the saliency score of this frame. For each frame, an integrated score is calculated as follows:

For each pixel in the combined saliency map
if $P_i \geqq 128$, $Score = Score + 1$.

After computing the saliency score of all video frames, we could construct a saliency curve to illustrate the trend of this video clip, as shown in Figure 4. To meet the limit of bandwidth for base layer, we design a window-based approach to selectively skip inconspicuous fames by the following steps:

(1) **Quantization:** We first quantize the saliency curve to several stages. The quantization step is defined as $\sigma * f$, where σ is the standard deviation of all frames with a given window and f is a factor to adaptively adjust the quantization step according to the characteristic of the window.

(2) **Variance Calculation:** The given window slides along the time axis gradually to cover all frames. The variance of the frames within the window is calculated to form the basis of saliency.

(3) **Scalable Coding:** If the variance of one video segment is larger than a pre-defined threshold, we say that this segment dazzled users and doesn't possess high semantic meaning. This video segment is then encoded in the enhancement layer due to storage or transmission restriction. On the other hand, the video segments with smooth saliency values are encoded as the base layer.

A good example of the video segments with high saliency variances is the scene with glittering light, such as the scene in a dancing party. This kind of video segments are first encoded as enhancement layer and enhance the whole video if more bandwidth is available. In Figure 4, frames from 481th to 540th and 681th to 840th are encoded as the enhancement layer.

Although the proposed approach generates the bitstream meeting the limited bandwidth, uncontinuous video frames may annoy users due to skipping a few frames of a solid video segment. Therefore, we apply a smooth mechanism to avoid this kind of skipping. In a window, if more than 80% of frames whose variances are smaller than a pre-defined threshold, the frames with high variances are not skipped. An example is the 641th frame in Figure 4. On the contrary, the frames with short sudden decrease in variance are skipped. Such a short-term low-variance frames are not long enough to capture users' focuses.

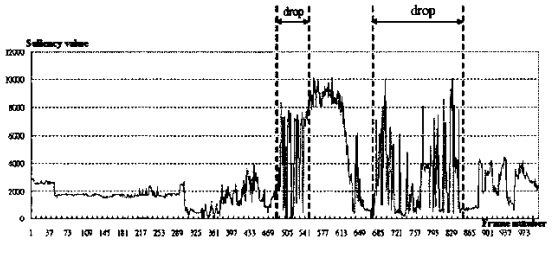

Figure 4. An example of saliency values in different frames

4. Experimental Results

4.1 Non-Scalable Spatial Coding

In this experiment, the foveation model parameters used are: α = 0.106, e_2 = 2.3, CT_0 = 1/64, and the viewing distance D is set as 1 to 6. From the foveation model, the critical eccentricities drop more with a larger viewing distance. The foveation center is set as the detected focus point using our user attention modeling system. Furthermore, the foveation filters can be adjusted to suit specific needs of applications. We can properly increase the minimum contrast threshold by modifying CT_0 as:

$$CT_1(k) = CT_0 + kS,$$

where k = 0, 1, ..., K, and S is the fixed step size.

Figure 5 shows some results of applying foveation filters to the sequence 'foreman', with different viewing distances (D = 1 and D = 6, respectively).

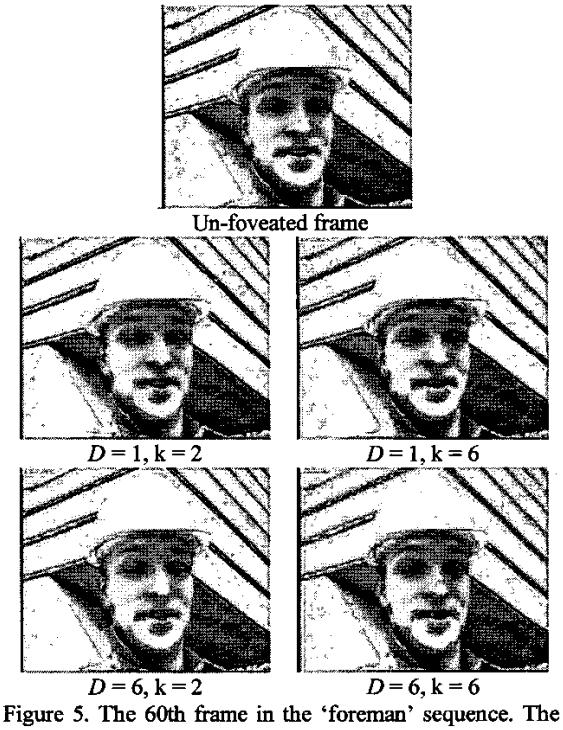

Figure 5. The 60th frame in the 'foreman' sequence. The foveation center is obtained from focus detection using saliency maps.

Applying a foveation filter to a raw video can help for reducing consumed bit-counts of MPEG encoding, while keeping up the quality of focused regions. To show this, we applied the foveation filter (with D = 6 and $CT_1 = CT_0$) to several test sequences, and compared the consumed bit-counts of un-foveated and foveated MPEG-1 videos. All MPEG-1 bitstreams were generated with following settings: The frame rate is 24; the coding pattern is

IBBPBBPBBPBB; a fixed quantization scale (16) is used for all frames; the motion search range is from –15 to +15. Table 1 shows the comparision results. It can be seen that applying foveation filtering does help for bitrate reduction.

Table 1. Bitrate savings of applying foveation filtering to various MPEG-1 encoded sequences

Se-quence	Original bitstream size(bytes)	Original bitstream bitrate (kbps)	Foveated bitstream size (bytes)	Foveated bitstream bitrate (kbps)	Birate Saving Ratio (%)
foreman	1329545	831	1230589	769	7.4
mobile	3913246	2445	3326153	2078	15.0
butterfly	392447	721	374541	688	4.5

In this paper, we propose a novel bitrate allocation approach based on foveation model. Here we adopt MPEG-1 codec in the experiment, but this approach can also be applied to more advance coding techniques such as MPEG-4 advance profile and H.264.

4.2 Scalable Temporal Coding

We propose a feasible approach to perform user-oriented temporal coding based on the prescribed user attention model. However, this work involves semantic-level coding which is highly human perception dependent. Different people may have different subjective assessments. Thus we can hardly provide an objective experiment results to approve this approach.

In our preliminary experiments, we found that this approach provides satisfactory results in some categories of videos. For example, in a news video, the segments with smooth frames, such as the scenes of anchorperson and close-up shot are preserved to be the base layer. Other segments with frequent scene changes are encoded as enhancement layer.

5. Conclusion

In this paper, we proposed a user-oriented approach combining user attention and foveation models to facilitate scalable coding in spatial and temporal domains. In the user attention model, saliency maps in different levels are generated to find user focus, and then, foveation model is applied to suppress the areas that users don't care. With the combination of these two models, scalable coding is achieved in both the spatial and the temporal domains. This framework could be extended to develop a transcoder that selectively transcodes a part of a video frame to meet different requirements in different devices. For example, we could just crop a focus region to fit the display resolution of a PDA, instead of resampling the whole frame. Prioritized protection is another extension by giving more redundant information to protect attentive regions.

References

[1] W. Li, "Overview of fine granularity scalability in MPEG-4 video standard," IEEE Transactions on Circuits and System for Video Technology, Vol. 11, No. 3, pp. 301-317, 2001.

[2] C.-C. Ho, W.-H. Cheng, T.-J. Pan and J.-L. Wu, "A User-Attention Based Focus Detection Framework and Its Applications" accepted by the 4th IEEE Pacific-Rim Conference on Multimedia (PCM'03).

[3] C.-C. Ho and J.-L. Wu, "A foveation-based rate shaping mechanism for MPEG videos," in Proc. 3th IEEE Pacific-Rim Conference on Multimedia (PCM'02), pp. 485-492, 2002.

[4] W. S. Geisler and J. S. Perry, "A real-time foveated multiresolution system for low-bandwidth video communication," SPIE Proceedings: Human Vision and Electronic Imaging, Vol. 3299, pp. 294-305, 1998.

[5] L. Itti and C. Koch, "Computational modeling of visual attention," Nature Reviews Neuroscience, Vol. 2, No. 3, pp. 194-203, 2001.

[6] J. G. Robson and N. Graham, "Probability summation and regional variation in contrast sensitivity across the visual field," Visual Research, Vol. 21, pp. 409-418, 1981.

[7] M. S. Banks and A. B. Sekuler and S. J. Anderson, "Peripheral spatial vision: limits imposed by optics, photoreceptors and receptor pooling," Journal of the Optical Society of America A, Vol. 8, pp. 1775-1787, 1991.

[8] W. S. Geisler, "Visual detection following retinal damage: predictions of an inhomogeneous retino-cortical model," SPIE Proceedings: Human Vision and Electronic Imaging (VCIP'96), pp. 119-130, 1996.

[9] Z. Wang and A. C. Bovik, "Embedded foveation image coding," IEEE Transactions on Image Processing, pp. 1397-1410, 2001.