Human Attention Region-of-Interest in Video Coding

Sylvia O. N'guessan and Nam Ling
Department of Computer Engineering
Santa Clara University
500 El Camino Real, Santa Clara, CA, USA
{sylvia.nguessan, nling}@scu.edu

***Abstract*—We propose a new scheme that exploits characteristics of motion vectors combined with luminance contrast to automatically detect human attention regions of interest (HAROIs) in every I-frame or intra-coded blocks in a group of pictures (GOP). These HAROIs can then be used for adaptive quantization. Our ultimate goal is to obtain a generic HAROI detection scheme of low complexity that can be used to improve compression while maintaining video quality and perception. Experimental results show consistency between actual human attention regions and the most relevant regions identified by our algorithm. Our algorithm also produces better compression of I-frames while improving both peak-signal-to-noise ratio (PSNR) and structural similarity (SSIM).**

Keywords- video coding; visual communications; human attention model; region-of-interest; motion vectors; contrast; quantization

I. INTRODUCTION

Region of Interest (ROI) techniques exploit the concept of psycho-visual redundancies and most importantly the philosophy of multiple human attention models that support the idea that a user only attends to a small portion of a video at a certain point in time. Evidently, the advantages of using the ROI concept have been exposed through numerous researches confirming it to be a good approach to reduce the compression ratio, better the coding efficiency and bit rate [1, 2, 3, 5] while maintaining relatively good image perceptual quality and PSNR [1, 2, 7, 8, 10]. Most ROI detection schemes revolve around a specific object [4, 5]. In the context of video streaming and video conferencing, it would be ideal to have a generic detection scheme that would adapt to the user's eye gaze focus.

Multiple visual attention models have emerged from psychophysical sciences and psychology in the aim of finding a general logic for how the human attention reacts to images or video [9, 17]. Consequently, we have decided to create a human attention based ROI (HAROI) classification scheme in the video encoding phase that relies on motion vectors and luminance contrast.

Section II overviews ROI classification schemes, Section III exposes the human attention models that are reflected in our method, Section IV describes our method, Section V discusses our results and concluding remarks are given in Section VI.

II. ROI CLASSIFICATION SCHEMES

ROI implementations are classically characterized by the selection of one or multiple regions of a video frame and the reduced quantization of that selected area during the encoding process. Some techniques automatically select the ROI or have the users interactively select it as it was done on a real-time video stream in [1, 8]. Other schemes focused on human physical traits like the head and shoulders [2], face [6], or skin tone [3]. Finally, ROI can be identified by tracking a specific object. For example, in [4], the center of the soccer ball corresponds to the center of the ROI.

If multiple ROIs are retrieved [4, 5, 7], they are then classified by loosely simulating the foveation technique described in [7]. With this particular classification scheme, a main region of interest is selected as having the highest priority and the other regions are assigned priorities proportional to the distance from that main ROI. For example, in the work done in [4], ROI priority is assigned relative to the distance of the player around the soccer ball; a player closest to the ball has highest ROI priority.

III. VISUAL ATTENTION MODELS

Visual attention models show where the human eyes are most likely going to be drawn to within an image [10]. They can be classified in two major groups: the top-down models which are object detection specific [2, 3] and the bottom-up models which are color-contrast based or motion based [4, 5, 11]. Our work focuses on the latter group. We use the assumption that a top-down models group is a subset of a

bottom-up models group. In other words, we argue that finding an ROI in a video based on motion and color contrast is a more generic approach than finding and classifying a specific object as an ROI. This explains why our work revolves around motion vectors and block luminance information.

To the best of our knowledge, there has not been any ROI detection scheme combining both motion vectors and luminance contrast information to classify human attention regions in a video. This has thus motivated our efforts based on three important observations of human attention research.

1) *Observation 1*: Human eyes are drawn to regions of high motion [4, 10, 11].
2) *Observation 2*: The work done in [9] concludes that attention moves smoothly when an observer is tracking an object in motion but moves abruptly when attention is directed to another location of interest.
3) *Observation 3*: Humans pay attention to regions with higher contrast. Contrast is a relevant factor in our work because it shows the distinctiveness of an object between itself and its environment. The saliency map visual attention model was originally proposed by *Itti et al* [12]. A saliency map is a gray level image where the regions with high contrast are considered to be HAROIs. Contrast based ROI has fueled a plethora of studies each supporting it to be an effective mechanism to match human attention [11, 12].

Our method is influenced by the Motion Attention Model (MAM) and the Static Attention Model (SAM) both described in [16]. The MAM is characterized on the estimation of a motion vector field that has three inductors: an intensity inductor, a spatial coherence inductor, and a temporal inductor. On the other hand, the SAM proposes a contrast-based saliency map to determine the brightness of a block relative to the entire image. We use this same idea, by obtaining information of the average luminance of a block relative to the entire image. The following section details our method in six steps.

IV. Proposed HAROI Method

In step 1, motivated by *Observation 1,* for each block i, we compute the motion vector magnitude $MvMag_{i\_a\text{-}b}$ between two consecutive frames a and b.

In step 2, inspired by *Observation 2,* for each block i, with horizontal and vertical block positions i_x and i_y (left top block has position (0,0)), and horizontal and vertical sizes (in pixels) of motion vector $Mvx_{i\_a\text{-}b}$ and $Mvy_{i\_a\text{-}b}$ we find the block j with horizontal and vertical positions j_x and j_y such that: $j_x = i_x + \lceil Mvx_{i\_a\text{-}b}/c \rceil$ and $j_y = i_y + \lceil Mvy_{i\_a\text{-}b}/c \rceil$ where c is an integer value indicating block size. For our experiment, we chose c to be 16. The block j (the impacted block) will have its impact count $IC_{j\_a\text{-}b}$ incremented by 1 and its impact force $IF_{j\_a\text{-}b}$ incremented by $MvMag_{i\_a\text{-}b}$. $IC_{j\_a\text{-}b}$ and $IF_{j\_a\text{-}b}$ are both indicators of the likelihood of human attention based on motion. The impacted block represents the likelihood of it becoming the new point of attention.

We then repeat steps 1 and 2 for all consecutive frames of a GOP and compute the average $MvMag_i$, IC_i, and IF_i for block i.

In step 3, we compute for each block $rIEd_i$ which is the ratio of the Euclidean distance between the block i and the center of the frame over half of the diagonal length of the frame. This step is inspired by the foveation technique [7] used to classify ROIs and the concept of spatial coherence inductor [11]. Also, since we know that human attention tends to be drawn to the center of the image, $rIEd_i$ then provides information about the closeness of a block to the center of the frame.

In step 4, based on *Observation 3*, we compute for each block i its average luminance Lu_i. We then compute the average block luminance of the entire I-frame μLu and its luminance standard deviation σLu. We also keep track in the process of the highest luminance value *maxLu* and lowest luminance value *minLu* for a block for that particular frame. The values *maxLu, minLu,* Lu_i, μLu, and σLu provide information for each block i about how contrasted it is in relation to the entire- I-frame. In other words, we loosely model a saliency map. This phase is crucial for the classification algorithm to determine each group of neighboring block based on σLu.

In step 5 (region creation process), we form groups by traversing the frame row by row. Each time we start a new row, a new region is created. Within a row, two adjacent blocks belong to the same region if each of their Lu_i obeys the criteria defined by equation (2) where k is an integer variable that ranges from 0 to k_{max}-1, and R is an integer that determines the size of the luminance range for each region:

$$k_{\max} = \text{round}\left[\frac{maxLu - minLu}{R \times \sigma Lu}\right] \quad (1),$$

$$minLu + kR(\sigma Lu) \le Lu_i \le minLu + (k+1)R(\sigma Lu) \quad (2).$$

If that condition is not satisfied a new region is created. Each time a new region is formed, we check to see that its first block belongs to the region above it. If so, we append the block to the region above it. As we go along this process, for each region, we update *ratioDistance* (the average of all $rIEd_i$'s of each block of the region), AvLu_{region} (the average of all Lu_i's of each block belonging to that region), *ratioLuminance* (AvLu_{region} over *maxLu*), *ratioMVMagnitude* (the average of all $MvMag_i$'s over *maxMvMag* (the largest *MvMag* of the frame)), *ratioImpactForce* (the average of all IF_i's over *maxMvMag*), *ratioImpactCount* (the average of all IC_i's over $maxIC_{\text{region}}$ (*IC* of the region with the highest *IC*)). For each region, the *ratioContrast* and the *regionScore* are computed as follows:

$$ratioContrast = \frac{2\times|(\mu Lu - AvLu_{region})|}{(maxLu - minLu)} \quad (3),$$

$$\begin{aligned} regionScore &= v_d \times (1 - ratioDistance) \\ &+ v_{lu} \times ratioLuminance + v_c \times ratioContrast \\ &+ v_{mvm} \times ratioMVMagnitude + v_{if} \times ratioImpactForce \\ &+ v_{ic} \times ratioImpactCount \end{aligned} \quad (4).$$

TABLE I. EXPERIMENTAL RESULTS JM VERSUS PROPOSED

	I-Frame (Frame 0) Results							Rate Distortion Results Frame Rate = 30 Hz Number of Frames = 100	
Sequences	***Y-PSNR (JM) (dB)***	***Y-PSNR (Proposed) (dB)***	***Y-SSIM (JM)***	***Y-SSIM (Proposed)***	***Bits/Frame reduction***	***Number of HAROIs***	***TRP***	***Y BD bit rate change***	***BD Y-PSNR change***
Akiyo (QCIF) 176×144	40.722	40.814	0.9802	0.9806	5.16%	22	72%	-7.48%	0.425
Soccer (CIF) 352×288	38.586	38.815	0.9532	0.9540	2.50%	67	30%	-7.75%	0.303
Crew (SD) 704×576	43.929	43.934	0.9468	0.9468	3.47%	168	75%	-0.03%	0.227
Sunflower (HD) 1920×1080	43.929	43.934	0.9762	0.9762	8.85%	647	69%	-0.845%	0.127

Figure 1. Akiyo I-Frame 0

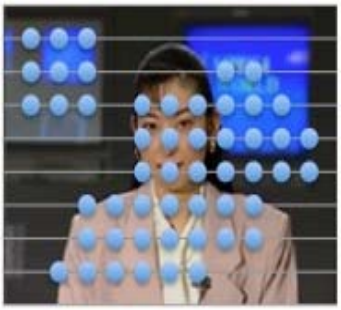

Figure 2. Top 72% HAROIs (Akiyo)

Figure 3. Soccer I-Frame 0

Figure 4. Top 30% HAROIs (Soccer)

Figure 5. Crew I-Frame 0

Figure 6. Top 75% HAROIs (Crew)

Figure 7. Sunflower I-Frame 0

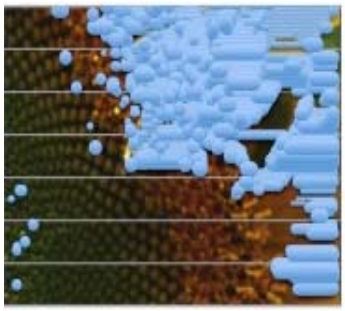

Figure 8. Top 69% HAROIs (Sunflower)

(v_d, v_{lu}, v_c, v_{if}, v_{mvm}, and v_{ic} are real weight values that can be assigned depending on the need). Finally, the *HumanAttentionScore* is the ratio of the region score over the highest *regionScore* of the entire frame.

In step 6, we reduce the quantization parameter (QP) for blocks with *HumanAttentionScore* within the range [0.85, 1], otherwise we increase it. In both cases, the QP is evaluated in proportion to the *HumanAttentionScore*.

The overall additional complexity per I-frame for all six steps is bounded by O(n) where n is the number of blocks of same size in the I-frame.

V. EXPERIMENTAL RESULTS

We implemented our algorithm using the H264/AVC JM Reference (version 17.2) software [13]. The values $k_{max,}$, v_d, v_{lu}, v_c, v_{mvm} , v_{if}, and v_{ic} of equation (4) were respectively set to 4, 1.25, 2, 2, 2.25, 2.5, and 2.5. The choice was made empirically by assigning more weights to luminance, contrast, impact count, and impact force, compared to that of distance to the center. Choosing k_{max} values ranging from 4 to 6 have always achieved the best HAROI detection accuracy. Higher values resulted in more regions of smaller sizes that were scattered across the frame. Lower values resulted in less accuracy in the HAROI detection. Choosing k_{max} to be 4 gave us regions concentrated in the same proximity.

Table I shows comparisons of PSNR, SSIM, and bits per frame (I) between the JM approach and our method for various videos (*Akiyo, Soccer, Crew* and *Sunflower*) of different resolutions. It also lists the number of HAROIs detected by our algorithm. We chose a group of ten individuals and had them select 1 to 5 regions that 'caught' their eyes on the first frame of the video sequence. We used that data to measure the *Top Regions Percentage* (*TRP*). TRP was computed by first sorting in a list all HAROIs in decreasing order of *HumanAttentionScore*. We then started from the beginning of the list and increment the count of HAROIs until we fully matched all regions selected by all ten individuals. We evaluated the percentage of that count of HAROIs over the total number of regions detected within the I-frame; this corresponded to the *TRP*. Figures 1, 3, 5 and 7 represent the first frame of each video sequences while Figures 2, 4, 6, 8 are respectively their equivalents but we added dots to represent the locations of *TRP's*. Figures 2 and 4 (*Akiyo* and *Soccer*) support the argument we made earlier about human attention models: the fact that top-down models (object detection) is a subset of a bottom-up model (contrast, motion). In the *Akiyo* video, the face is included in the 40% of the top HAROIs. In the *Soccer* video, the players and soccer ball are selected without any form of object detection with a *TRP* of 30%. The bee of the *Sunflower* video, detected within 30% of the top HAROIs, is because it is the main element with motion, although its contrast is not high. Generally, from our experimental results, a *TRP* of 75% should guarantee the strongest consistency with actual human attention while 40% of the top HAROIs correspond to the most salient regions.

Finally, the rate distortion results for each sequence were obtained with the first 100 frames at 30 Hz [13]. We used the Bjøntegaard method [14] to measure the average differences between R-D curves of JM versus our proposed algorithm. On average, we obtained a bit rate reduction of more than 7% for QCIF and CIF sequences with an average PSNR gain of 0.36 dB. On the other hand, we obtained a smaller bit rate reduction of less than 1% for SD and HD sequences with a smaller PSNR gain averaging 0.175. The difference between these low resolution (QCIF, CIF) and high resolution (SD, HD) videos is explained as follows: for lower resolution the increased QPs correspond to smaller regions than those where QPs decreased; higher resolution videos have larger denser regions where we increased the QPs. As a whole, visual quality is maintained or improved. Observing from the experiments conducted using our algorithm the bit-rates of I-frames are reduced in most cases from 2% to 9% without perceivable visual quality loss.

VI. CONCLUSION

In this work, we have shown that our low complexity algorithm can be used to detect HAROIs. We have compiled an approach that allows adaptive quantization algorithms to have flexibility. The current results reinforce the fact that our method is a good approximation of the human attention model. We have improved or maintained PSNR and SSIM values while increasing compression. Future work will factor in the use of eye-tracking software and camera motion to better refine our HAROIs detection and classification scheme.

REFERENCES

[1] D. Grois, E. Kaminsky, and O. Hadar, "Adaptive bit-rate control for region-of-interest scalable video coding," IEEE Convention of Electrical and Electronics Engineers in Israel, pp. 761-765, November 2010.

[2] Z. Bojkovic and D. Milovanovic, "Multimedia coding using adaptive regions of interest," Seminar on Neural Network Applications in Electrical Engineering, pp. 67- 71, September 2004.

[3] Y. Liu, Z. G. Li, and Y. C. Soh, "Region-of-interest based resource allocation for conversational video communication of H.264/AVC," IEEE Transactions on Circuits and Systems for Video Technology, vol. 18, no. 1, pp.134-139, January 2008.

[4] J. Y. Kim, C. H. Yi, and T. Y. Kim, "ROI-centered compression by adaptive quantization for sports video," IEEE Transactions on Consumer Electronics, vol. 56, no. 2, pp. 951-956, May 2010.

[5] M. Firoozbakht, J. Dehmeshki, M. Martini, Y. Ebrahimdoost, H. Amin, M. Dehkordi, A. Youannic, and S. D. Qanadli, "Compression of digital medical images based on multiple regions of interest," Proceedings of the International Conference on Digital Society, pp. 260-263, Feb. 2010.

[6] H. Zheng, Y. Lu, and X. Feng, "Improved compression algorithm based on region of interest of face," Proceedings of the International Conference on Artificial Reality and Telexistence Workshops, pp. 345-348, November 2006.

[7] D. Agrafiotis, D. R. Bull, N. Canagarajah, and N. Kamnoonwatana, "Multiple priority region of interest coding with H.264," Proceedings of the IEEE International Conference on Image Processing, pp. 53-56, October 2006.

[8] M. Makar, A. Mavlankar, P. Agrawal, and B. Girod, "Real-time video streaming with interactive region-of-interest," Proceedings of the IEEE International Conference on Image Processing, pp. 4437-4440, September 2010.

[9] S. Shioiri, T. Inoue, K. Matsumura, and H. Taguchi, "Movement of visual attention," Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics, vol. 2, pp. 5-9, 1999.

[10] C. M. Tsai, C. W. Lin, W. Lin, W. H. Peng, "A comparative study on attention-based rate adaptation for scalable video coding," Proceedings of the IEEE International Conference on Image Processing, pp. 969-972, November 2009.

[11] Y.-F. Ma, X.-S. Hua, L. Lu, and H.-J. Zhang, "A generic framework of user attention model and its application in video summarization," IEEE Transactions on Multimedia, vol. 7, no. 5, pp. 907–919, October 2005.

[12] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 20, no. 11, pp. 1254–1259, November 1998.

[13] "H264/AVC JM Reference Software"
Website: http://iphome.hhi.de/suehring/tml/download/ .

[14] G. Bjøntegaard, "Calculation of average PSNR difference between RD-curves", document VCEG-M33.doc, ITU-T VCEG, 13th Meeting, Austin, TX, USA, April 2001.