REGION-OF-INTEREST BASED COMPRESSED DOMAIN VIDEO TRANSCODING SCHEME

Aniruddha Sinha[1], Gaurav Agarwa1[2] and Alwin Anbu[3]

Motorola India Electronics Ltd.
{[1]aniruddha.sinha@ motorola.com, [2]gaaga@rediffmail.com, [3]alwin_pdy@yahoo.co.in}

ABSTRACT

We propose a fast video transcoding technique based on Region-of-Interest (ROI) determination. The ROIs are identified using the properties of the Human Visual System (HVS), applied in the compressed domain. We use the edge, motion and spatial frequency content of the video frame as the parameters in identifying perceptually important regions in compressed domain [9]. The TM5 rate-control of the video transcoder is modified to assign relatively more bits to the ROIs, thereby providing better quality to the areas that are likely to attract the viewer's attention. The ROI determination in compressed domain has been modified to give more robust results. Our algorithm achieves a speed-up of 15-20 times as compared to similar algorithms in pixel domain while giving comparable results. This computation is only around 8% of the total transcoder complexity.

1. INTRODUCTION

Recent trend has shown a tremendous demand for video technology in multimedia applications. The video is transmitted in standard compressed formats such as MPEG, H263 etc. For video services in heterogeneous networks it is necessary to convert the bit-rate of the compressed video to match the bandwidth of the transmission channels of low capacity. Therefore, a great lack of flexibility arises in the transmission of these bit streams due to the wide dissimilarity in the capabilities, of the various applications at the receiving end. These applications have variable encoding format, bitrate and display size requirements.

The solution to this problem lies in video transcoding, which enables seamless video communication between two devices with varying capabilities. Various schemes have been suggested for transcoding architecture [7]. When the bit-rate is reduced, the overall quality of the video is degraded, equally throughout the frame, due to re-quantization.

In this paper we suggest a region-of-interest (ROI) based transcoding in the compressed domain. In this scheme, the rate control assigns more bits to the ROIs. This approach allows for a non-uniform treatment of the scene content. This improves the visual quality of selected regions of the frame, without perceptually degrading the background and unimportant areas. The ROIs are determined in compressed domain based on the Human Visual System (HVS)[9]. Extensive simulation shows that we achieve a better visual quality compared to existing transcoding applications without any significant increase in complexity.

The paper is arranged as follows: Section 2 gives an overview of the HVS and the current research on its applications to image and video coding. Section 3 presents the modified algorithm of finding ROIs in the compressed domain. Section 4 presents the rate control modifications. Section 5 presents the results and Section 6 shows the speed-up achieved by operating in the DCT domain as compared to similar algorithms that operate in the pixel domain. Section 7 concludes the paper.

2. HUMAN VISUAL SYSTEM

Various experiments have been conducted so far, to identify the processes involved and the behavior of the HVS [7,8]. Different schemes have been proposed to develop HVS models based on these results. Itti, et al [1] proposed a saliency based visual attention model, which decomposes the input into a set of topographic feature maps based on intensity, color and orientation contents of the image. Luo and Singhal [3] demonstrated the effectiveness of using an amorphous spatial context model for determining the neighborhood of a region when computing relative saliency features based on color, texture and shape properties. For dynamic visual scenes, such as video, the motion present in a region plays an important part in deciding its perceptual importance [5]. Osberger and Maeder [2] used contrast, size, shape, color and motion in deciding the regions of interest in a video frame.

Thus, the major factors that decide perceptual importance of a region are its motion, spatial frequency, edge and color contents. We use these factors in the compressed domain for identifying the ROIs. Computation of these parameters in the compressed domain is fast; at the same time, the accuracy with which they are able to isolate the ROIs is good.

3. PROPOSED ALGORITHM

There are various schemes already proposed for video transcoding in both spatial and compressed (DCT) domains. In general, a video transcoder can be used to convert any video format to any other video format at any arbitrary bit-rate and frame rate. In this paper, we focus on MPEG4 Simple profile transcoder to generate a low bit-rate video from high bit-rate video, for QCIF size images. (This concept can easily be format to any other video format at any arbitrary bit-rate and frame rate.)

There are mainly two stages in the video transcoder, the first stage is used to partially decode the incoming bit-stream and the second stage is used to re-quantize and encode it at the desired low bit-rate [7]. In Fig.1, Q_1^{-1} and Q_2^{-1} represent the

0-7803-8484-9/04/$20.00 ©2004 IEEE

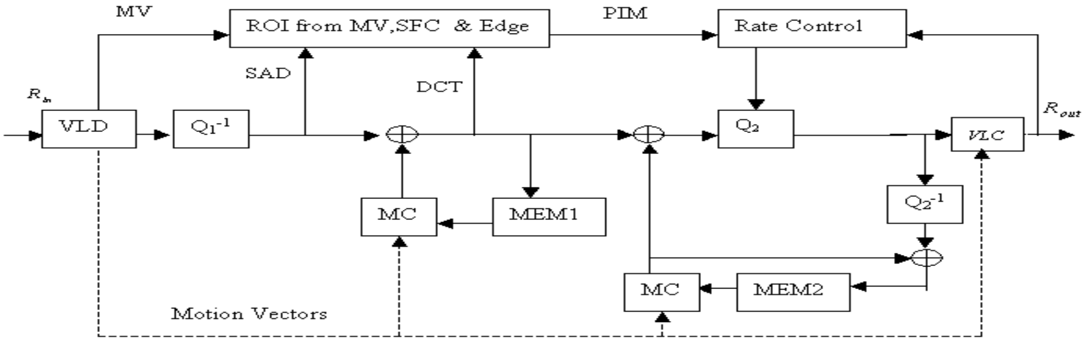

Fig.1 Region of Interest based Compressed Domain Video Transcoder

inverse quantization of the first stage and second stage respectively. MEM1 and MEM2 denote the reconstructed frames.

The MPEG standard encodes the video frame in units of 16x16 Macroblocks (MBs). The bitstream has information about the motion vectors and the DCT of the residue for every MB. The bitstream can be partially decoded to obtain the motion vector and DCT of each MB. In our algorithm, the following information from the bitstream (Rin) is used:

i) Reconstructed image in the first stage in the DCT domain.

ii) Normalized SAD (Sum of Absolute Difference) of each decoded blocks in the first stage. This is same as the DC values of the DCT coefficients of the blocks in the first stage

iii) Motion vectors decoded from first stage (MV).

This extracted information is used to find the motion, spatial frequency, edge and color contents of the MB. Based on these features, we find the following Importance Maps (IM), for every encoded video frame, that rate the perceptual importance of every macroblock

1. Motion Importance Map (MIM)
2. Edge Importance Map (EIM)
3. Spatial frequency content map (FIM)

A final Perceptual IM (PIM) for each frame is in turn computed by a weighted addition of these three IMs. Fig.2 shows the block diagram for the computation of the final PIM. The PIM is an input to the rate control, which uses it to assign more bits to perceptually important regions.

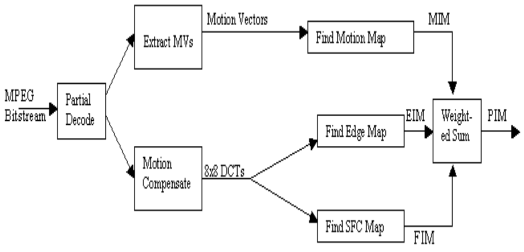

Fig.2. Block Diagram for Finding the PIM

3.1. Motion Importance Map (MIM)

The MIM is the normalized sum of two maps:

1. Motion Magnitude Map (MM)
2. Motion Center Surround Map (MC)

The motion magnitude map is computed as follows:

For each MB, i (1 <= i <= 99, for QCIF)

$$MM(i)= (|MVx(i)|+|MVy(i)|) * (|MVx(i)| + |MVy(i)|) / SAD(i) \quad \ldots(1)$$

where MVx(i) and MVy(i) denote the x and y components of the motion vectors of the i th MB. SAD (i) used here is normalized.The motivation for this map is that regions with higher motion are more important. At the same time, a high value for SAD indicates an unreliable motion vector.

Therefore, the division by the SAD value reduces the contribution of these unreliable motion vectors.

The Motion Center Surround Map (MC) identifies those regions in the image that have a motion different from their surroundings. To obtain the map, we do an operation similar to Itti's [1] center surround technique, on an image termed Motion_Image, whose pixel values are the magnitudes of the motion vectors of the MBs. The MC is the Motion Image map at scale 2. Thus,

$$MC = |\{Motion\_Image - (LPF\ (Motion\_Image)\ \Downarrow 2)\Uparrow 2\}| \quad \ldots(2)$$

3.2. Edge Importance Map (EIM)

It is known that regions with more edges are perceptually important [1]. To find the edge content of a macroblock, edge detection is done on the compressed domain data .We use a 9x9 isotropic LoG operator, because it is rotationally symmetric and thereby reduces the number of computations needed. The details of this method could be found in [9]

3.3. Spatial Frequency Content (SFC) Map (FIM)

Reinagel and Zador [4] found the spatial frequency content at the eye-fixated locations in an image to be significantly higher than, on average, at random locations. Therefore, we seek to obtain the spatial frequency content of each macroblock. The operation is simplified since we have the DCT values of the MBs. For each MB, the number of DCT values that are higher than a threshold is computed and the MBs are rated correspondingly . The details of this method could be found in [9].

3.4. Perceptual Importance Map(PIM)

The three importance maps are now added to obtain the final Importance Map. It is obvious that the motion map should get the highest weight, since, a region with a motion different from the surroundings is perceptually very important [5]. We experimentally determined that the weight ratio for motion, SFC and edge should be 4:2:1, for good results.

$PIM = 4*MIM + 2*FIM + EIM$ ...(3)

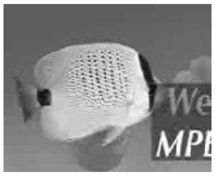

Fig.3a Original

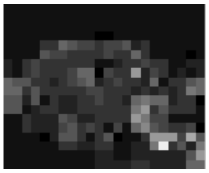

Fig.3b Perceptual Map (PIM)

4. RATE CONTROL

The transcoder shown in Fig. 1 changes the bit-rate of MPEG4 bit-stream from input bit-rate (I_{br}) to the desired output bit-rate (O_{br}). A new rate control scheme is proposed to satisfy the output bit budget while relatively improving the quality of ROIs. This rate-control is a modification of TM5 algorithm, and takes the PIM as an additional control parameter.

The quantization scale factor of each macroblock (MB) is determined to meet a given bit budget. Frame bit allocation (B_{cf}) in current frame is determined from the bits spent till previous frame (B_{pf}) and average bits per frame (B_{av}). If N^{th} frame is encoded then,

$B_{cf} = N * B_{av} - B_{pf}$ … (4)

$B_{av} = O_{br} / F_r$ … (5)

where F_r is the frame rate.

Appropriate MB bit allocation is done to meet B_{cf}. ROI patterns are used to modulate the quantization parameters for the MBs to selectively improve the quality. Experimental results show that the proposed scheme gives visually improved quality as compared to TM5. In MPEG4 Simple Profile(SP) transcoder the TM5 rate control algorithm is modified to handle only I and P frames. Initial quantization value found from TM5 is given as,

$mquant = Q_j \times N\_act_j$, … (6)

$N\_act_j = (((2 \times act_j) + avg\_act) / (act_j + (2 \times avg\_act)))$ …(7)

where: N_act_j is the normalized activity

Q_j is the quantization parameter of the j^{th} MB.

act_j is the activity in the j^{th} macroblock

avg_act is the average value of act_j in the last encoded picture.

The final MB level quant value is derived by the following equation

$quant = mquant – (PIM[j] - 7)$ …(8)

where, $PIM[j] = \{x, 1<=x<=15\}$, j = 1 to 99 for QCIF sequence.and $2<=quant<=31$,

The above equation denotes that a macroblock whose PIM[j] is 7 is not affected.

In MPEG4 SP the quantization value can at most vary by +/-2 between MBs if the MB is not the start of a slice (packet). In packet boundaries, absolute quant values can be transmitted in the bit-stream; hence any quant value within the standard limit can be used to adjust the number of bits generated by the MB and meet the bit-budget. Thus if a MB falls under high PIM[j] then a new packet is started to accommodate the drastic change in quant indicating the start of ROI zone. This packet ends once the ROI zone gets over. Usually the PIM is generated such that the important zone gets over in the same row of MB. Within the ROI zone, quant can be varied by +/-2 between MBs depending on the bits generated. Once the ROI zone is over, a new slice is started and then the quant is increased based on the (8). Once the full frame is encoded and it has generated actual bits (B_{ac}), then excess frame bits (E_f) is used to derive the next frame's expected bits, where,

$E_f = B_{cf} – B_{ac}$ … (9)

To reduce the variation of the bits generated in each frame, E_f is reduced by maintaining the sum of PIM[j] for all the MBs in each frame equal to or less than 6*M*N, where M is the width and N is the height of the frame. This is a constraint imposed on the generation of PIM as the value of 7 in PIM[j] is treated as neutral point as seen in (8).

5. RESULTS

We tested the proposed algorithm on various standard and general sequences for converting QCIF MPEG4 SP bitstream, from 256kbps to 128 kbps, using the proposed DCT domain transcoder. The improvement in the MSE of the ROIs are listed in Table1. Fig 4 and Fig 5 compares the visual quality of the transcoded bitstreams with and without ROI usage.Fig 6a and Fig6b show the reduction in the Frame level MSE's for the first 30 frames. It can be seen the average MSEs has decreased in the important regions adding to the overall visual quality of the video perceptually without degrading the other regions.

Sequence	Normal MSE (db)	ROI MSE (db)
MPEG Fish	15.90	14.14
Coastguard	7.45	7.22
Ship	4.11	0.93
Office Surveillance	5.13	4.12
Table Tennis	10.97	9.22
Highway	5.04	2.76
Talking Lady	6.52	4.17
Talking Man	1.82	1.73
Super Market	11.11	9.57
James Bond	7.95	5.32
Man Entering	0.83	0.65
Man and tree	11.01	10.81
Professor in Library	5.88	5.21
Soccer	11.82	11.48
Medley	13.39	12.66
Mother & Daughter	7.51	6.28

Table1. MSE comparison for the ROI.

Fig.3 shows the original image (reconstructed QCIF, encoded at 128 Kbps), the Perceptual Importance Map (PIM), for frame no.10 of 'fish' sequence. Fig. 5 compares the final result at 128 Kbps. There is a need of more experimentation to get a better relative weighting factor of the three importance maps.

6. COMPLEXITY ANALYSIS

We wish to determine theoretically the saving in complexity achieved by our algorithm in finding the ROIs as compared to Itti's saliency based scene analysis technique [1].

Itti [1] uses Gaussian pyramids at scales of 1 to 8 to find color and intensity maps. The orientation maps are found using Gabor pyramids, at orientations of 0, 45, 90 and 135 degrees.

For an MxN size image, the number of computations needed, on a floating-point processor, for the various steps in [1] are:

- Low Pass Filtering for all scales (assuming a 5x5 window): 8.3 *M*N multiplications and 8 *M*N additions
- Difference operation (at Scale 2): (M*N)/4 additions
- Modulation (at Scale 2): (M*N)/4 multiplications

For a 176x144 image, the total number of computations needed for finding 6 intensity maps, 12 color maps and 24 orientation maps is approximately $5.5*10^6$ multiplications and $5.5*10^6$ additions.

In the proposed algorithm, for a MxN frame, assuming a 8x8 MB, the finding of SFC Map requires at most M*N/64 multiplications and MN additions. The finding of the motion. map requires 0.17*M*N additions and 0.16*M*N multiplications. The edge importance map requires 9.1*M*N multiplications and 10*M*N additions.

Therefore, for a 176x144 frame(M=176, N=144), the computation of the final Perceptual importance map for each frame requires *9.28*M*N = $2.4*10^5$ multiplications and 13.17*M*N = $3.3*10^5$ additions.*

Thus it can be seen that compared to [1], our algorithm achieves a speed-up of 15-20 times in the number of basic arithmetic operations needed. This computation is only around 8% of the total transcoder complexity.

7. CONCLUSION

We have proposed a MPEG-4 video transcoding system based on HVS based region-of-interest identification. It is observed that the ROI computation in the compressed domain is much faster than the widely used saliency based feature extraction algorithm, operating in the pixel domain. The perceptual quality of the ROI transcoder output is significantly better than that of the normal transcoder, while maintaining the overall bitrate requirements. The proposed scheme for finding the ROI can also be used in other applications such as scene-dependent video encoding and post processing, defining new MPEG-7 descriptors, etc.

8. REFERENCES

[1] L. Itti, C. Koch, and E. Niebur, "*A model of saliency-based visual attention for rapid scene analysis*," IEEE Trans. Pattern Anal. Machine Intelligence, vol. 20, pp. 1254--1259, 1998.

[2] Osberger, W. and Maeder, A.J., "*Automatic Identification of Perceptually Important Regions in an Image*", ICPR 98, pp 701-704,1998.

[3] L. Jiebo, A. Singhal*, "On Measuring Low-Level Self and Relative Saliency in Photographic Images"*, PRL (22), No. 2, pp. 157-169, February 2001.

[4] Reinagel, P., and Zador, A. M., "*Natural scene statistics at the center of gaze",* Network: Computation in Neural Systems, 10:1—10, 1999.

[5] G.Boccignone "Vision between action and perception", http://www-dii.iing.unisi.it/aiia2002/paper.htm

[6] B.S Manjunath, et al, "*Color and Texture Descriptors",* IEEE Trans. On Circuits Syst. Video Technol., vol. 11 no. 6, pp. 703-715, 2001.

[7] *Vetro, A.; Christopoulos, C.; Huifang Sun;"**Video transcoding architectures and techniques: an overview"*** Signal Processing Magazine, IEEE , Volume: 20 Issue: 2 , March 2003 Page(s): 18 -29

[8] J.K. Tsotsos, et al, "*Modelling Visual Attention via Selective Tuning*," Artificial Intelligence, vol. 78, no. 1-2, pp. 507–545, Oct. 1995.

[9] Agarwal, G.; Anbu, A.; Sinha, A. "*A fast algorithm to find the region-of-interest in the compressed mpeg domain*". ICME, 2003, Volume: 2 , 6-9 July 2003, Page(s): 133 –136

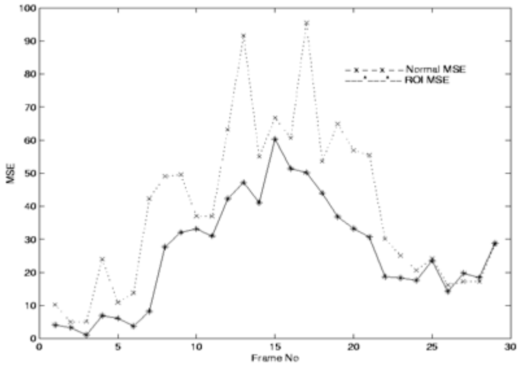

Fig 4a Comparison of ROI and Normal MSE for fish sequence

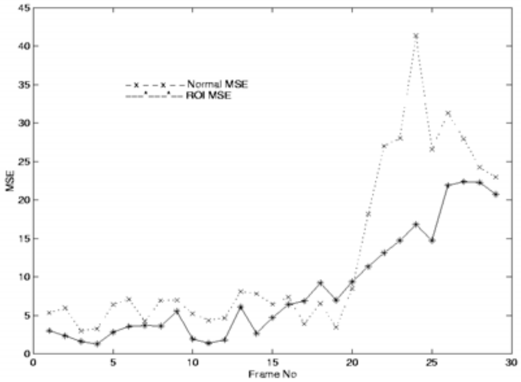

Fig 4b Comparison of ROI and Normal MSE for Table Tennis sequence

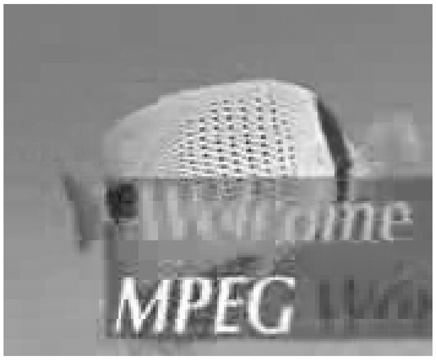

Fig 5(a) ROI Transcoding at 128 Kbps

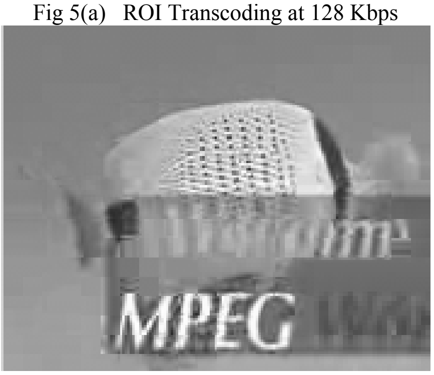

Fig 5(b) Normal Transcoding at 128 Kbps