# MPEG-21 DIGITAL ITEM ADAPTATION BY APPLYING PERCEIVED MOTION ENERGY TO H.264 VIDEO

*Zhao Gang[1], Liang-Tien Chia[2], Yang Zongkai[1]*

[1]Department of Electronics & Information Engineering,
Huazhong University of Science & Technology, Wuhan City, China.
E-mail: zhaogangping@163.com, Zkyang@public.wh.hb.cn

[2]Center for Multimedia and Network Technology,
School of Computer Engineering, Nanyang Technological University, Singapore.
E-mail: asltchia@ntu.edu.sg

## ABSTRACT

*The new MPEG-21 standard defines a multimedia framework to enable transparent and augmented use of multimedia resources across heterogeneous networks and devices used by different communities. In this paper, we incorporated the Perceived Motion Energy (PME) Model into the proposed MPEG-21 Digital Item Adaptation Framework for frame dropping in H.264 encoded video adaptation. There are two advantages of this work, one is the use of PME model to reduce the viewer's perceived motion jitter due to frame dropping to a minimum. The other is the adaptation nodes can easily apply frame dropping operations without knowledge of detailed encoding syntax of H.264 videos.*

## 1. INTRODUCTION

In the context of video adaptation, dropping frame is a simple and efficient method for rate adaptation. However users will perceive the loss as motion jitter and motion jitter is heavily dependent on the motion patterns in the video sequence. To minimize the effort of motion jitter, it is reasonable to analyses motion information and drops the frames that are visually less important in the human perception system.

In another aspect intermediate video adaptation nodes with limited processing capabilities can hardly handle complex adaptation and may not support some video formats especially new standards such as H.264. Generic Bitstream Syntax Description (gBSD), a component of the new standard MPEG-21 Digital Item Adaptation (DIA), provides a format-independent multimedia adaptation mechanism which can also be used to convey information such as results of motion analysis on video sequences to intermediate adaptation nodes.

In this paper, we use the Perceived Motion Energy (PME) [1-2] Model, which provides a good representation of motion information to determine the importance of frames to human perception, for H.264 encoded videos. We incorporated this model into MPEG-21 DIA Framework by generating the gBSD embedded with PME information for H.264 videos. A gBSD Adaptation Engine interprets this description, dropping the less important frames to satisfy the bandwidth constraint or limited display capability of devices. Our work focuses on dropping B-frames in H.264 because of its simplicity and the fact that B-frames makes up two-third of all frames and about 30-40% of the total size in H.264 video sequences. (Although B-frames can be used as reference frames in H.264, it is not used in the test model).

There are two advantages in this work. One is the use of PME to reduce to a minimum the viewer's perceived motion jitter due to frame dropping. The other is that adaptation nodes can easily apply frame dropping without knowledge of detailed encoding syntax of H.264 videos.

The remainder of this paper is organized as follows. Section 2 briefly introduces the Perceived Motion Energy Model and its modification for H.264 videos. Section 3 is the description of MPEG-21 framework. Then we describe our integration of PME model and MPEG-21 DIA framework in Section 4. Experiment results will be shown in Section 5. Finally, we conclude our work and outline future work directions.

## 2. PERCEIVED MOTION ENERGY

Many methods such as dominant motion based method [3], Camera motion based method [4] make use of motion information for video retrieval. However, these methods cannot provide sufficient motion information for general users and are not suitable for video adaptation. So we use the PME model to evaluate perceived motion jitter for

This work is partially supported by Chinese National Science Foundation (Project No.60202005).

0-7803-8554-3/04/$20.00 ©2004 IEEE.

frame dropping. We will briefly introduce this model and our PME extraction for H.264 videos as follows.

### 2.1 Perceived Motion Energy

In compressed video, motion information between frames is captured in motion vectors. However, motion vectors are just rough and sparse approximation to the real optical flows and are more sensitive to noise in magnitude than in direction [5]. Therefore a median filtering based on magnitude is usually necessary to smooth out the inherent noise.

Let $W_s$ be the width of the modified median filter window, the filtered magnitude of motion vector is computed by:

$$Mag_{i,j} = \begin{cases} Mag_{i,j} & (if\ Mag_{i,j} \le Max4th(Mag_k)) \\ Mag4th(Mag_{i,j}) & (if\ Mag_{i,j} > Max4th(Mag_k)) \end{cases} \quad (1)$$

where $Mag_{i,j}$ is the magnitude of the motion vector at the micro block $MB_{i,j}$, function $Max4th(Mag_k)$ returns the fourth value in the descending sorted list of magnitude elements in the filter window.

The filtered magnitudes are averaged in temporal energy filter such as an alpha-trimmed filter within a 3-D spatial-temporal tracking volume, with the spatial size of $W_t^2$ and the temporal duration of $L_t$. The mixture energy $MixEn_{i,j}$, which includes the energy of both object and camera motion, is calculated as:

$$MixEn_{i,j} = \frac{1}{(M-2\cdot\lfloor \alpha M \rfloor)\cdot W_t^2} \sum_{m=\lfloor \alpha M \rfloor+1}^{M-\lfloor \alpha M \rfloor} Mag_{i,j}(m) \quad (2)$$

where M is the total number of magnitudes in tracking volume, and $\lfloor \alpha M \rfloor$ equals to the largest integer not greater than $\alpha M$ ;and $Mag_{i,j}(m)$ is the magnitude values in the sorted list of tracking volume. The trimming parameter $\alpha(0 \le \alpha \le 0.5)$ controls the number of data samples excluded from the accumulating computation.

To evaluate motions in one frame, the average magnitudes $Mag(n)$ of motion vectors for P and B frames are computed by (3) and (4) respectively:

$$Mag(n) = \sum MixFEn_{i,j}(n)/N \quad (3)$$

$$Mag(n) = (\sum MixFEn_{i,j}(n)/N + \sum MixBEn_{i,j}(n)/N)/2 \quad (4)$$

where $MixFEn_{i,j}(n)$ and $MixBEn_{i,j}(n)$ represent forward and backward motion vectors in frame $n$. N is the number of macroblocks in the frame.

To match the fact that the human eyes tend to track dominant motion in the scene, the percentage of dominant motion directions is calculated as:

$$\alpha(n) = \frac{\max(AH(n,k), k \in [1,m])}{\sum_{k=1}^{m} AH(n,k)} \quad (5)$$

The total angle of $2\pi$ is quantized into $m$ angle ranges and $AH(n,k)$ is an angle histogram with $m$ bins. So $\max(AH(n,k), k \in [1,m])$ is the dominant direction bin among all motion directions.

The PME is defined as the product of the average magnitude of motion vectors and the percentage of dominant motion direction:

$$PME(n) = Mag(n) \times \alpha(n) \quad (6)$$

$Mag(n)$ reflects the fact that dropping frames of low motion intensity is less perceptible than dropping frames of high motion intensity. Multiply by $\alpha(n)$ will add a significant factor to PME if there are camera motions or other dominant object motions.

### 2.2 PME extraction in H.264 video

In our work, the PME model is applied in the new H.264 standard. Because of the new features in H.264 standard, we have to make some modifications to the traditional method of extracting PME information in previous MPEG encoded videos.

One of the new features in H.264 is the variable block-size motion compensation with small block sizes. There may be motion vectors for 16x16, 8x16, 16x8, 8x8, 4x8, 8x4 or 4x4 block sizes. So it is not appropriate to calculate the PME at the macroblock level. Instead, we extract PME at 4x4 block level because we can get motion vectors for all 4x4 blocks by dividing large-size predicted blocks such as 16x16, 8x16, 16x8 blocks into numbers of 4x4 blocks and assigning the same motion vectors within the predicted blocks for these 4x4 blocks.

Another important new feature in H.264 is multiple reference picture motion compensation. If the reference picture is very far away from current frame, we note that the motion vector for this reference frame should contribute less to PME because viewers are sensitive to motion jitter occurring recently. So we assign weights for every block based on the distance from its reference picture when calculating PME. We replace $MixFEn_{i,j}(n)$ in (3), (4) with $\overline{MixFEn_{i,j}}$ which is defined as:

$$\overline{MixFEn_{i,j}(n)} = \beta(d)\ MixFEn_{i,j}(n) \quad (7)$$

where $\beta(d)$ is the weight for the block for which the distance between current frame and the reference frame is $d$ .In simplicity, we defined $\beta(d)$ as:

$$\beta(d) = \frac{\lambda}{d} \quad (8)$$

In (8) $\lambda$ is a constant to avoid largely decrease of weights for different encoded pattern of I, P, B.

Additionally, there are different prediction types for every block in H.264 such as inra prediction, forward prediction, backward prediction and bi-direction prediction

types. The influence of these different prediction types to the motion jitter should be carefully studied. We found that intra prediction block should be regarded as having very large motion because it contains new information and we assigned the motion vectors with the maximum search area for example (16, 16). We also noted that the loss of forward prediction will cause more motion jitter than the loss of backward prediction because of the display order of the video sequence. Thus we assign large weights for forward motion vectors and small weights for backward motion vectors in our PME extraction algorithm for H.264 video. The weight $\gamma(t)$ for different prediction types is defined as:

$$\gamma(t) \in \left\{1,1,\frac{1}{2},\frac{1}{4}\right\}$$

$$t \in \{Intra, Forward, Backward, Bi-direction\} \qquad (9)$$

That means $\gamma(t)$ will be 1, 1, 1/2, 1/4 for the prediction type of Intra, Forward, Backward and Bi-direction.

By combining (7),(8),(9),we rewrite (3),(4) as PME extraction algorithm for H.264 video:

$$\overline{Mag(n)} = \sum d_e \frac{\lambda}{d} \cdot \gamma(t) \cdot MixEn_{i,j}(n) / N \qquad (10)$$

where $d_e$ is a leverage constant for the decreasing effect of weighting for multiple reference picture feature in forward prediction type compared to other prediction types.

## 3. MPEG-21 DIGITAL ITEM ADAPTATION

The new MPEG-21 standard defines a multimedia framework to enable transparent and augmented use of multimedia resources across a wide range of networks and devices used by different communities [6]. The fundamental unit of distribution and transaction in this framework is Digital Item (DI), which is a structured digital object with standard representation, identification, and associated metadata.

gBSD is a part of MPEG-21 Part 7 Digital Item Adaptation (DIA) that provides a flexible description method based on XML for multimedia resources and format-independent multimedia adaptation mechanism. An adaptation engine has been proposed for both resource adaptation and gBSD descriptor adaptation [7]. When DIs embedded with gBSD are subject to such an adaptation engine, the engine will determine the optimal adaptation given the constraints as provided by the usage environment description. Based on this decision, description transformation using XSLT can be applied to the original description. Using the new transformed description from these transformations; an adapted DI can be easily generated. More details can be found in [7].

## 4. PME BASED VIDEO ADAPTATION

HP Research Labs has proposed the Structured Scalable Meta-formats (SSM) for Fully Content Agnostic Adaptation [8] as a MPEG-21 reference software module. This software can dynamically adapt multimedia contents to network conditions and terminal constraints. However; the software cannot dynamically generate gBSD to assist the adaptation decision process in the adaptation engine.

Based on the automatic generation module of gBSD in [7] and the reference software of H.264, we developed a revised automated PME based gBSD generator for H.264 videos and established an adaptation framework based on HP's Model. The architecture of this framework is shown in Fig 1.

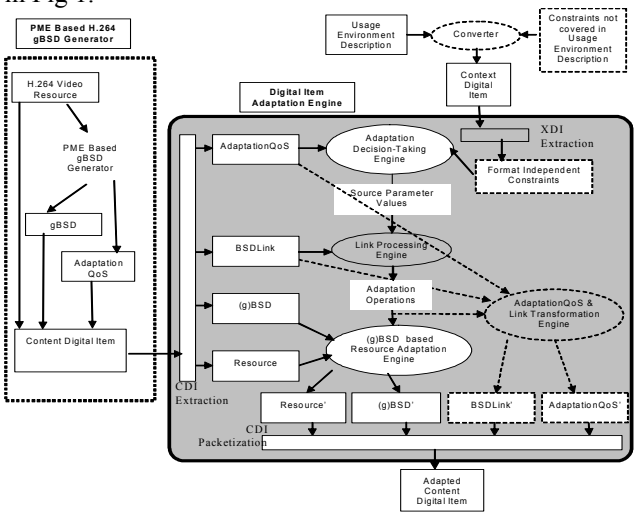


Fig 1. System Architecture of PME based H.264 video adaptation for Digital Item Adaptation

PME based gBSD generator firstly calculates PME values of all frames. Then it dispatches these frames to several quality layers according to their PME values. These quality layers correspond to different importance of frames. When dropping different quality layers, viewers will get different amount of motion jitter. For example, if there are 10 quality layers and the PME value of one frame belongs to the least 10% category, this frame will be put into Quality Layer 1 or if it belongs to the least 20% category, its quality layer should be 2. gBSD will be generated for this video while the layered information for every frame will be labeled using the "Marker" component in gBSD.Additionally an adaptationQoS descriptor with the information about the bit rates after dropping some quality layers has also been generated by the PME based gBSD generator. These two types of off-line generated descriptors will be packaged with the video itself into a DI, which will be subject to HP's adaptation engine.

The adaptation engine will determine the optimal adaptation based on the relationship of current network bandwidth $BW$ and current transmission bit rate $R_k$, where $k$ is the minimum number of the current transmitting quality layer that means all the layers from layer $k$ to top

layer $n$ are transmitted with $R_1$ having the highest bit rate because all the layers are transmitted. The objective of this adaptation is to retain higher quality layers, whenever possible by dropping lower quality layers.

- If $R_k < BW < R_{k-1}$ : This means the current network bandwidth can only support transmitting layer $k, k+1, ...n$ . Adding $k-1$ layer to current transmission will exceed the capacity of the network. So no adaptation occurs.
- If $R_{k+1} < BW < R_k$ : This means network bandwidth decreases to only support the transmission of layer $k+1$ and higher layers $k+2$ etc. Therefore, the adaptation engine should drop all frames belonging to the layers which are below layer $k+1$ .
- If $BW > R_{k-1}$ : This means the network can support the higher transmitting bit rate, thus a lower quality layer than $k$ can be transmitted. The adaptation engine will allow addition frames from layer $k$ to $k$-1 to be transmitted.

## 5. EXPERIMENT RESULTS

To test the effectiveness of our video adaptation algorithm, we apply subjective user study to evaluate the perceived quality of the adapted video after dropping different percentages of B-frames. The testing environment conditions are according to the ITU Recommendation P.910 [9]. We selected the test sequences from H.264 standard ftp site. Ten human subjects were invited to assign scores to the adapted versions of test sequences. Every subject was asked to grade the adapted videos in five-level scale: 5.Excellent, 4.Good, 3.Fair, 2.Poor, 1.Bad and the results are shown in Table 1.

Table 1.Subjective User Study Evaluation –Frame Dropping Rate

| | Excel-lent | Good | Fair | Poor | Bad |
|---|---|---|---|---|---|
| **Foreman** | 0% | 10-20% | 30-60% | 70-80% | 90-100% |
| **Akiyo** | 10% | 20-30% | 40-60% | 70-90% | 100% |
| **Container** | 0% | 10% | 20% | 30-60% | 70-100% |
| **Hall** | 0% | 10-20% | 30-60% | 70-80% | 90-100% |
| **Mad** | 0% | 10-20% | 30-50% | 60-80% | 90-100% |

From table 1, we noticed that video sequences, with an average up to 50-60% of dropped frames will get a grading of "Fair" and this will result in a corresponding reduction of an average of 24% in the transmitting bit rate. This means that our framework dropped most of the visually less important frames in the human perception system.

## 6. CONCLUSIONS AND FUTURE WORK

The use of PME model will allow us to label the frames according to the level of motion activity and therefore perceived visual quality. In this paper, we integrated this model to the upcoming MPEG-21 DIA adaptation framework for H.264 video adaptation. Our work provides a standard and flexible way to adapt H.264 videos to actual usage environments, while the viewers' perceived qualities are retained as high as possible.

One aspect of our future work is revising the PME model to get more accurate representation of motion information for H.264 encoded videos. Another considerable research direction is how to realize the video adaptation in distributed environment, in such a situation several consecutive adaptations should be executed in distributed intermediate nodes.

## 7. REFERENCES


[1] Y. Ma, H.J. Zhang. "A New Perceived Motion based Shot Content Representation," *ICIP 2001*, Greece, October 2001.

[2] T. Liu, H. J. Zhang, F. Qi, "Perceptual Frame Dropping in Adaptive Video Streaming," *Proc. of ISCAS2002*, Arizona, May 26-29, 2002.

[3] E. Ardizzone, et al, "Video Indexing Using MPEG Motion Compensation Vectors," *Multimedia Computing and Systems*, pp.725-729, 1999.

[4] E. Ardizzone, M. La Casica, D. Molinelli, "Motion and Color-Based Video Indexing and Retrieval," *ICPR'96*, pp.135-139, 1996.

[5] H. Wang, A. Divakaran, A. Vetro, S.-F. Chang, H. Sun, "Survey of Compressed-Domain Features Used in Audio-Visual Indexing and Analysis," *Journal of Visual Communication and Image Representation*, vol.14, pp.150-183, June 2003.

[6] J. Bormans, J. Gelissen, A. Perkis, "MPEG-21 The 21st century multimedia framework," *IEEE SIGNAL PROCESSING MAGAZINE*, MARCH 2003.

[7] G. Panisa, A. Huttera, J. Heuera, H. Hellwagnerb, H. Koschb, C. Timmererb, S. Devillersc, and M. Amielhc, "Bitstream syntax description: a tool for multimedia resource adaptation within MPEG-21," *Signal Processing: Image Communication*, EURASIP, vol. 18, no. 11, 2003.

[8] D. Mukerjee, G. Kuo, S. Liu, and G. Beretta, "Motivation and Use cases for Decision-wise BSDLink, and a proposal for Usage Environment Descriptor-AdaptationQoSLinking," in *ISO/IEC JTC 1/SC 29/WG 11, Hewlett Packard Laboratories*, April 2003.

[9] "Subjective video quality assessment methods for multimedia applications", *ITU-T Recommendation P.910*, ITU-T, Geneva, 1996.