MERL – A MITSUBISHI ELECTRIC RESEARCH LABORATORY
http://www.merl.com

Coding Mode Optimization for MPEG-2 Transcoding with Spatial Resolution Reduction

Hao-Song Kong, Anthony Vetro, and Huifang Sun

TR-2003-99 August 2003

Abstract

This paper presents a coding mode decision algorithm for MPEG-2 spatial transcoding. The optimization for coding mode and quantization scale are formulated in an operational rate distortion sense and solved by Lagarange multiplier method. The experimental results show that the proposed transcoder with optimized coding mode and quantizer can achieve better quality and lower bit rate than those obtained using cascaded transcoder or MPEG-2 TM5 encoder.

Publication History:

1. First printing, TR-2003-99, August 2003

Coding Mode Optimization for MPEG-2 Transcoding with Spatial Resolution Reduction

Hao-Song Kong, Anthony Vetro, and Huifang Sun

Mitsubishi Electric Research Laboratories
201 Broadway, Cambridge, MA 02139, USA
hkong@merl.com; avetro@merl.com; hsun@merl.com

ABSTRACT

This paper presents a coding mode decision algorithm for MPEG-2 spatial transcoding. The optimization for coding mode and quantization scale are formulated in an operational rate distortion sense and solved by Lagarange multiplier method. The experimental results show that the proposed transcoder with optimized coding mode and quantizer can achieve better quality and lower bit rate than those obtained using cascaded transcoder or MPEG-2 TM5 encoder.

Keyword: MPEG-2 transcoding, spatial resolution reduction, coding mode decision, Lagrangian minimization, motion vector re-sampling

1. INTRODUCTION

Video transcoding is an efficient mechanism used to convert video bitstreams from one coding format to other formats, including syntax, bit rate and resolution conversions. Attempts have been made to put transcoders in the video servers or in the middle of network routers to deliver visual content to a variety of users who have different network connections or terminal devices with different display capabilities. MPEG-2 video coding standard has been widely used in digital broadcast and DVD applications and its transcoder has been under investigation for several years [1]-[5]. In transcoding techniques, re-quantization of DCT coefficients, spatial resolution reduction and temporal resolution reduction are the three major tools dealing with bitrate conversion. The re-quantization tool has been used for bitrate conversion with the same spatial-temporal resolution. The spatial resolution reduction tool is often useful when the target bitrate is too small or when the devices only have small display capabilities that are not suitable for playing full resolution video. The temporal resolution tool is applied when the bitrate reduction is further required or processing on a terminal needs to be reduced.

This paper emphasizes on MPEG-2-to-MPEG-2 transcoding with spatial resolution reduction, since there are increasing demands for video transcoding with spatial resolution reduction. Such requirements come from High-Definition TV (HDTV) broadcasting and DVD applications, as well as mobile devices with universal multimedia accessing. Transcoding with spatial resolution reduction can be performed either in the pixel domain (spatial domain) or in the compressed domain (DCT frequency domain). In the spatial domain, the simplest way for the transcoding is through cascaded architecture, i.e. to decode the video bitstream into pixel pictures, downscale the pictures to the desired size and perform the motion estimation based on the spatially downscaled pictures, and then re-encode them into lower spatial resolution bitstream. Cascaded pixel domain transcoding has some desirable advantages. For example, it provides more dynamic control and operations of the bitstream. By using a new quantization scale in the encoder, the bit rate of the output bit stream can be dynamically adjusted to match the available network bandwidth condition. In addition, editing functionalities, such as a logo or a digital watermark insertion can be embedded into the video stream. However, the major problem with the cascaded transcoder is that its motion estimation operation requires an intensive computation.

In DCT domain transcoding, the incoming bitstream is first decoded into DCT coefficients. Then spatial reduction is performed in the DCT domain. Since the motion compensation is realized in the DCT domain [6]-[7], no DCT and IDCT operations are needed, which simplifies the DCT domain transcoding architecture. However, computational complexities of motion compensation and motion estimation in the DCT domain are very high, even though several efficient algorithms have been proposed using matrix operation [8]-[9]. A much more powerful processor is expected compared to that of spatial domain transcoding, as stated in the previous studies [4] [13].

Since motion estimation requires intensive computational complexity, most of the reported transcoding architectures, such as those described in [4]-[5], [10]-[12], do not perform motion estimation. These architectures achieve transcoding bitstreams by re-using motion information embedded in the original video sequences to avoid the motion estimation operation. Utilizing the motion information to decide coding modes and estimate the motion vectors for the downscaled video is very challenging. Usually, the incoming motion information, such as coding modes and motion vectors, cannot be used directly. They have to be re-sampled and downscaled. The reported methods only focus on the motion vector re-sampling problem, without fully utilizing the coding mode in the transcoding to guarantee the coding efficiency and video quality.

The MPEG-2 video standard provides very rich macroblock level coding modes to achieve coding gains. These coding modes are divided into two groups: intra coding mode and inter coding mode. The intra coding mode involves the coding of a macroblock using only information from the macroblock itself. The inter coding mode, on the other hand, involves the coding of a macroblock using information from macroblocks occurring at different times. There exist intra mode, frame/field/dual-prime motion compensation inter mode, no motion compensation mode, forward/backward/interpolative inter mode, and field/frame DCT mode in P and B pictures. The advantage of the multi-mode coding approach is that its inherent adaptability lays a foundation for better coding efficiencies. To achieve the best coding performance, it is critical to select the most efficient coding mode for each macroblock. This paper presents an optimization algorithm for macroblock level coding mode selection.

The rest of the paper is organized as follows. In section 2, previous approaches to the video transcoding with spatial resolution reduction are reviewed. In section 3, our proposed optimization algorithm is described in detail. Section 4 shows the experimental results and compares the results with those produced by different algorithms, e.g. MPEG-2 TM5 encoder and cascaded transcoder. Finally, conclusions are given in section 5.

2. REVIEW OF PREVIOUS TRANSCODING APPROACHES

In transcoding with spatial resolution reduction, the original image size is normally downscaled by 2, although the downscale factor can be any rational number [14]. The operation of downscaling by 2 reduces four macroblocks of size 32x32 in the original video to a new macroblock of size 16x16 in the downscaled video. Therefore, the motion vectors extracted from the incoming bitstream cannot be applied to the reduced image directly. They need to be downscaled accordingly to form new motion vectors, as shown in Figure 1. Based on the new motion vectors, a predictive residue is calculated, DCT transformed, re-quantized and encoded into a new bitstream. The key issue is how to calculate the new motion vectors from the input bitstream and make sure the new motion vectors are accurate for the downscaled video image. The same problem occurs for the macroblock coding mode, as shown in Figure 2.

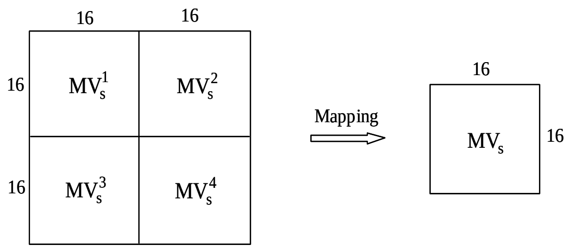

Figure 1. Macroblock motion vector mapping

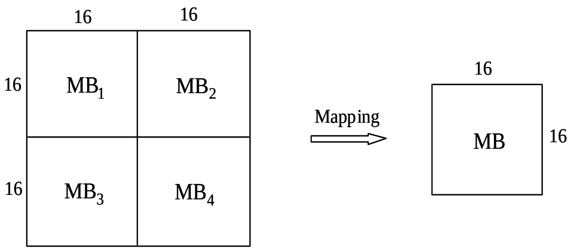

Figure 2. Macroblock coding mode mapping

Several algorithms have been proposed for re-sampling motion vectors using the extracted motion information from the original video bitstreams. The simplest method to estimate the downscaled motion vectors is to average the original motion vectors in the four macroblocks [5][15]:

$$\vec{v} = \frac{1}{2N}\sum_{i=1}^{N}\vec{v}_i$$

where $\vec{v}_i$ denotes the motion vector of macroblock *i* in the original video, $\vec{v}$ denotes the motion vector for the downscaled macroblock and *N* is the total number of motion vectors in four macroblocks. This method would lead to poor results if the input motion vectors were not well aligned; for example, the magnitude of one of the input motion vectors is significantly larger than the rest.

Another method proposed in [5] and [15] is to generate a median vector as the downscaled motion vector. This method calculates the sum of the distance between each vector and its neighbors as follows:

$$d_i = \sum_{\substack{j=1 \\ j\neq i}}^{4} \| \vec{v}_i - \vec{v}_j \| .$$

The median vector is defined as the vector with the least distance from all vectors. The above method extracts the motion vector situated in the middle of all motion vectors. The magnitude of the selected motion vector is then downscaled by 2 to reflect the reduction in the spatial resolution.

For the macroblock coding mode decision, the following options are also given in [5] and [15]:

- Derive a new MB mode in the encoder;
- Use a majority-voting mechanism; i.e. using the majority of the input macroblock type in the output bitstream;
- If there is at least one Intra type among the 4 MBs, then output Intra mode;
- If there is at least one Inter type MB and no Intra MB, then output Inter mode;
- If all MBs are of the Skip type, then output Skip mode.

Both the above schemes would be good only if all the input motion vectors were well aligned, however, this is not the case in most videos. Therefore, in [10] the authors proposed an adaptive motion vector re-sampling (AMVR) method. It takes into account the spatial activity measurement to generate the new motion vector by

$$\vec{v} = \frac{1}{2}\frac{\sum_{i=1}^{4}\vec{v}_i A_i}{\sum_{i=1}^{4} A_i}$$

where A_i denotes the activity measurement of residual macroblock *i*. The results obtained by this method are slightly better than those obtained by above-mentioned methods.

Based on the AMVR algorithm, various weighting methods have been proposed. These methods weight the motion vectors by their different measurements. For example, in [11] a maximum average correlation method is introduced. The motion vector with the maximum weighted average correlation among the input motion vectors is selected as the downscaled motion vector. In [4] quantization parameters are used as weighting factors to represent the spatial activity measures. In [12] extended neighboring macroblocks are involved in the motion vector re-sampling, and the weighting factors are set according to the input motion vectors. If the four input motion vectors are equal, each weight is equally set and equal to 1/4. This is equivalent to the average mean algorithm. If the four input motion vectors are different, the weights are set as in [10]. This is equivalent to the weighted mean algorithm (AMVR). Otherwise, the weights for the four input motion vectors are set to 1/4 and multiplied by 0.8 and the weights for the neighboring input motion vectors are set to 1/8 and multiplied by 0.2. In [16] a weighted median method is used to weight the input motion vectors in order to calculate the downscaled motion vector.

All the above-mentioned algorithms are effective in a small motion situation. With high motion video, however, they do not work properly since the re-sampled motion vectors do not accurately describe the motion in the downscaled video. These inaccurate motion vectors result in larger predictive errors and lead to higher transmission bit rates. Besides, most of these methods determine the coding mode for the downscaled video by using the majority-voting mechanism. The resulting modes are certainly not optimal. Other criteria for making the mode decision have also been identified in some of these methods, but the coding mode is only limited to intra and inter decision, so the optimal coding performance cannot be achieved since the rich coding modes provided by the MPEG-2 coding scheme are not fully utilized. In this paper, we propose a transcoding method that lays emphasis on video quality and low bit rate. In contrast to the above methods, which focus on weighting input motion vectors, our method considers input motion vector re-sampling and refinement, as well as quantization scale selection and coding mode decision.

3. PROPOSED APPROACH

The architecture of the proposed transcoder is shown in Figure 3. The input video stream is received by the video decoder for bitstream syntax decoding. The decoded image sequence is downscaled to the sub-sampled image sequence by the downscaling filter, according to the spatial resolution requirement. The macroblock information extracted from the input video stream is used for motion vector re-sampling. To overcome the impulse noise problem caused by the re-sampling process, motion vector refinement is applied, generating accurate motion vectors for the downscaled image sequence. The resulting motion vectors are then used for macroblock prediction. As stated earlier, different coding modes are available and the proper coding mode will produce optimal coding performance.

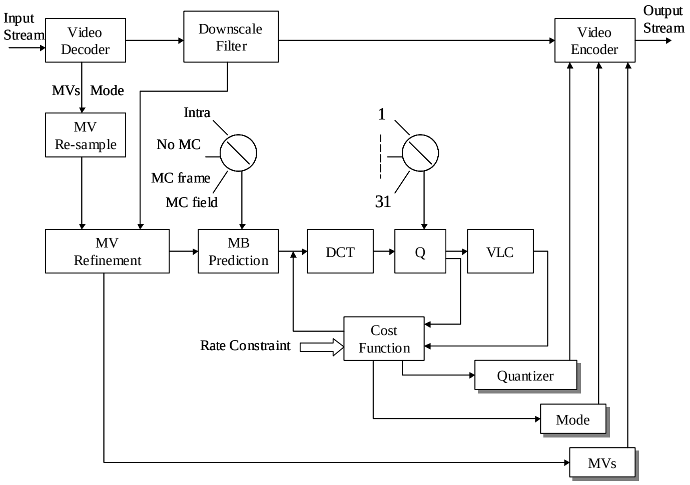

Figure 3. Quantization scale and mode decision optimized transcoding architecture.

3.1 Combined quantization scale and coding mode optimization

It is important to realize that coding modes should be determined jointly with rate control because the best coding mode depends on the operating point for the bit rate. The optimization must choose the most efficient coding mode for each macroblock in the rate-distortion (R-D) sense. This task is complicated by the fact that the various coding modes show varying efficiency at different bit rates. Intuitively, improved R-D performance is expected if the modes can be applied judiciously to different macroblocks. In the proposed algorithm, the macroblock prediction for each coding mode combined with each quantizer is calculated. The optimal quantizer and mode decision are obtained by minimizing the Lagrangian cost function, which is based on the rate and distortion calculation:

$$J_i(\lambda, M_k, q_i) = \min_{M_k}\{D_i(M_k, q_i) + \lambda R_i(M_k, q_i)\} \qquad (1)$$

where the M_k is varied over the coding mode set (7 modes for P picture, 11 modes for B picture), q_i is the quantizer step size $\in \{q_{1,} q_2, ... q_N\}$, $\forall i = 1,...N$, and N is the macroblock numbers of each frame. The mode decision and quantization parameter that are assigned to the macroblocks generate different R-D characteristics. Our goal is to determine a set of quantizer step sizes for all macroblocks of each frame, such that the total distortion *D* is minimized and the total number of bits *R* complies with the target budget imposed by the constraint, $R_{picture}$. The constrained problem is then formulated as:

$$\min D \quad \textit{subject to} \quad R < R_{picture} \qquad (2)$$

with *D* and *R* given by

$$D = \sum_{i=1}^{N} d_i(q_i) \qquad R = \sum_{i=1}^{N} r_i(q_i) \qquad (3)$$

For a particular value of the Lagrange multiplier, λ, if a set of $q_i^*(\lambda)$ minimizes the following expression:

$$\min_{q_i}\{d_i(q_i) + \lambda r_i(q_i)\} \ \ \forall i = 1,...N \qquad (4)$$

then this set of $q_i^*(\lambda)$ corresponds to an optimal solution to equation (2).

To find the optimal operating point on the R-D curve, we searched for an optimal slope, λ^*, in equation (4), such that, $R(\lambda^*) < R_{picture}$. A fast convex search algorithm [17] has been implemented in this paper and is outlined in the following steps.

Step-1) Initialize two values of λ, λ_1 and λ_2, with $\lambda_1 < \lambda_2$ which satisfies the relation:

$$\sum_{i=1}^{N} R_i(\lambda_1) < R_{picture} < \sum_{i=1}^{N} R_i(\lambda_2)$$

Step-2) $\lambda_{next} = \dfrac{\lambda_1 + \lambda_2}{2}$.

Step-3) Substitute λ_1 and λ_{next} into expression (4), minimize the expression and derive $q_i^*(\lambda_1)$ and $q_i^*(\lambda_{next})$, $\forall i = 1,...N$, respectively.

Step-4) If $[R(\lambda_1) - R_{picture}][R(\lambda_{next}) - R_{picture}] < 0$ substitute λ_2 by λ_{next}, otherwise substitute λ_1 by λ_{next}.

Step-5) If $|\frac{R(\lambda_{next}) - R_{picture}}{R_{picture}}| < \varepsilon$, where ε is a preset small positive number, the optimal slope λ^* is found and $q_i^*, \forall i = 1,...N$ is the optimal quantizer step size for each macroblock; else, go to Step-2.

3.2 Coding mode optimization after obtaining quantizer

Since the optimal $q_i^*, \forall i = 1,...N$ is derived for each macroblock subject to the constraints of $\sum_{i=1}^{N} r_i(q_i^*) < R_{picture}$, equation (1) becomes:

$$J_i(\lambda, M_k \mid q_i) = \min_{M_k}\{D_i(M_k \mid q_i) + \lambda R_i(M_k \mid q_i)\} \qquad (5)$$

The minimum of the Lagrangian rate distortion function is now obtained by setting its derivative to zero, i.e.,

$$\frac{\partial J}{\partial R} = \frac{\partial D}{\partial R} + \lambda = 0$$

which yields

$$\lambda = -\frac{\partial D}{\partial R}$$

Since q_i is given to each macroblock, therefore λ can be solved by the following approximation:

$$\lambda = -\frac{\partial D}{\partial R} \approx -\frac{\Delta D}{\Delta R} = \frac{D(q) - D(q-1)}{R(q-1) - R(q)}$$

For each candidate mode, the cost function (5) is calculated and the one that has the minimum cost is selected as the coding mode for the macroblock. In the proposed method, motion vectors are simply averaged to form the candidate motion vector and a small search window further refines the candidate motion vector. The impulse noise in the input motion vectors can be indicated by a high cost value. Therefore, the coding mode with such a motion vector hardly has a chance to be selected.

The quantizer selection in the proposed transcoding architecture can in fact be achieved by any known means. For example, the well-known TM5 quantizer selection process or an optimal quantizer selection process may be used. The main point is that this quantizer selection process can be separated from the mode decision to lower the computational complexity, yet high quality is still achievable.

4. EXPERIMENTAL RESULTS

For simplicity, P frames are used in the experiments for discussions. There are seven modes for P frame coding. Only four modes (intra, no MC coded, MC frame, MC field) are used as mode estimation for transcoding. Figure 4 illustrates the macroblock type for P frame. To evaluate the effectiveness of the proposed optimal quantization scale selection and mode decision algorithm, experiments are conducted with the HDTV video test sequence “Sprink” which has a resolution of 1920x1080i (interlaced video) and is encoded at 30Mbps. The encoded bit stream is transcoded down to a resolution of 720x480i with average target bit rates between 5Mbps and 9Mbps. In order to show the efficiency of the transcoding, the proposed transcoder is compared to a cascaded transcoder and an MPEG-2 TM5 encoder.

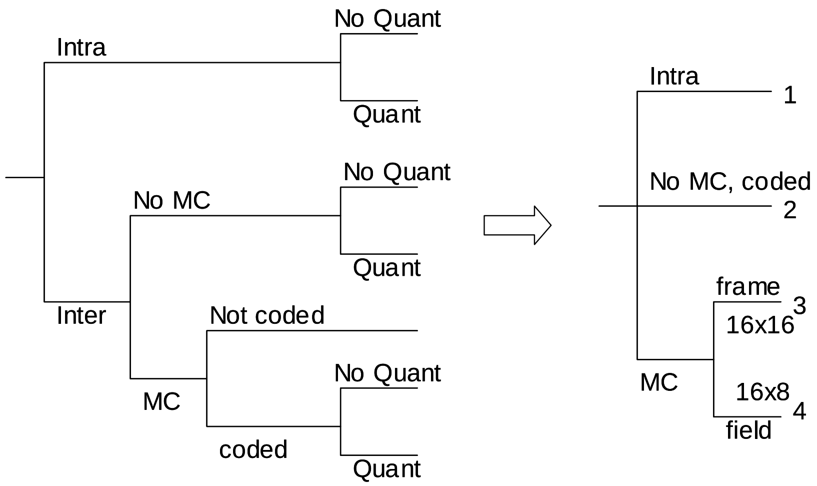

Figure 4. Transcoding modes for P frame.

Figure 5 shows the peak signal to noise ratio of 100 frames transcoded from 30Mbps to 9Mbps. The results demonstrate that our transcoding scheme provides nearly constant quality and achieves higher signal to noise ratio gain compared to the cascaded transcoder and the MPEG-2 TM5 encoder.

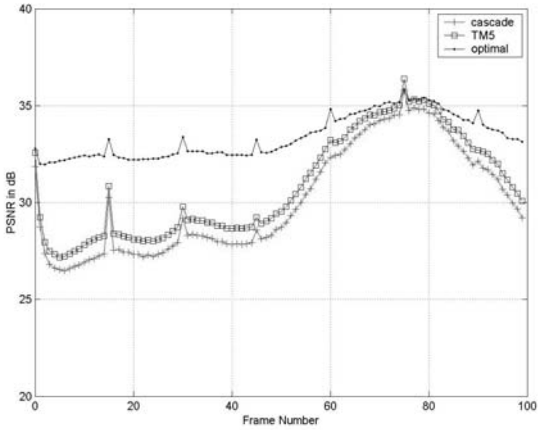

Figure 5. PSNR comparison among three coding schemes on “Sprink” test sequence.

Figure 6 shows the PSNR of 100 frames for different conversion ratios. The average PSNR obtained by our optimally transcoded video is better than that obtained by the other two schemes.

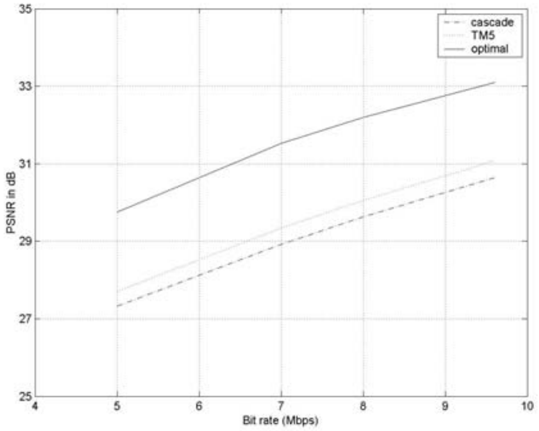

Figure 6. PSNR vs. bit rate of "Sprink" test sequence.

Figures 7, 8 and 9 show the visual results of the 20^{th} frame from the sprink test sequence. These images are transcoded from 30Mbps to 6Mbps or directly encoded at 6Mbps, using cascaded transcoder, MPEG-2 TM5 encoder and our proposed transcoder, respectively. The coding artifacts are obvious in Figure 7 and 8. The blocky effect almost disappears in Figure 9.

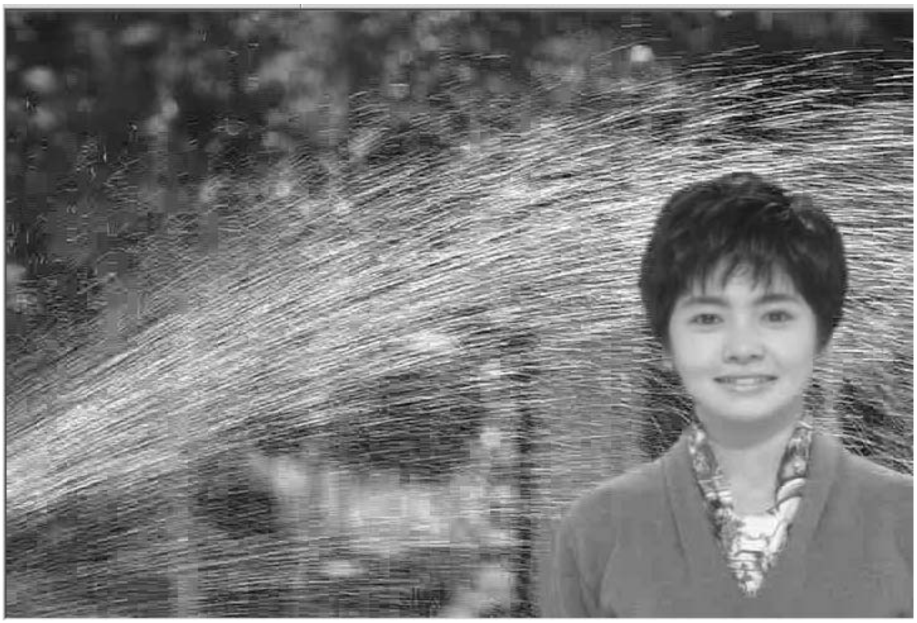

Figure 7. The 20^{th} frame transcoded by cascaded transcoding method.

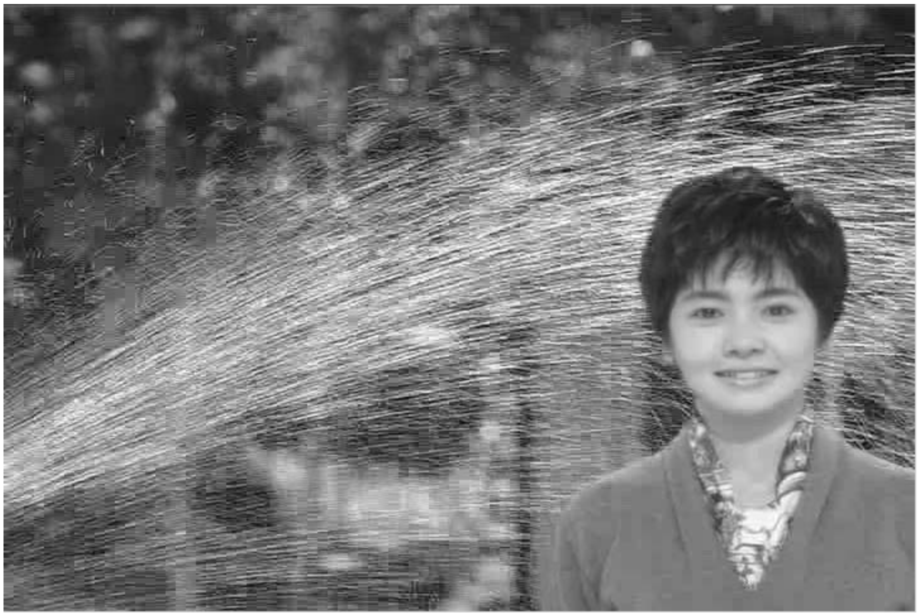

Figure 8. The 20th frame encoded by MPEG-2 TM5 method.

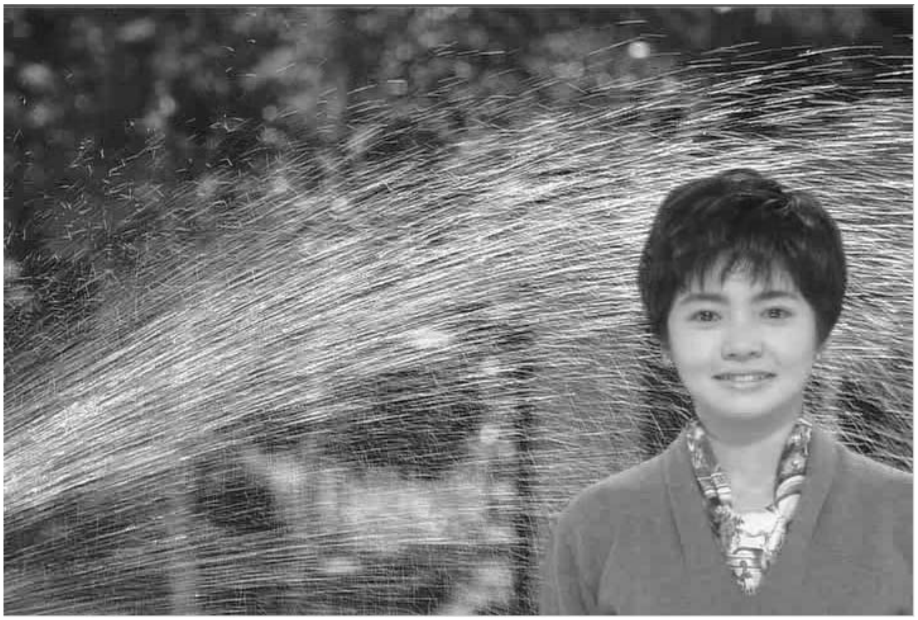

Figure 9. The 20th frame transcoded by proposed method.

Apart from the experiments mentioned above, we have also performed test on other video sequences. Figure 10 shows the result of "Mobile" test sequence originally encoded at 6Mbps. The resolution is transcoded from 704x480 to 352x240.

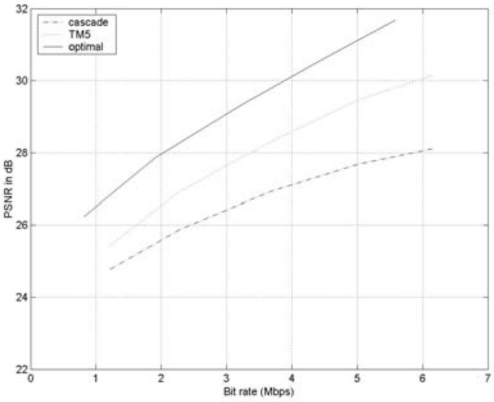

Figure 10. PSNR vs. bit rate of "Mobile & Calendar" sequence.

5. CONCLUSIONS

In response to demands for high-quality and low bit rate transcoding for HDTV broadcasting, DVD application and mobile communications, we propose an optimal transcoder that adapts the bit rate and spatial resolution from both MP@HL to MP@ML and from MP@ML to MP@ML. This approach optimizes the coding mode and quantization scale using Lagarange multiplier algorithm. The experimental results show that the proposed transcoder can achieve higher quality and lower bit rate performance compared to other methods, giving rise to the promise to provide high quality recording of HDTV broadcast in the DVD recording system, or to downscale video to different sizes for various devices with low bandwidth capacity in the network video server.

6. REFERENCES

[1] G. Keesman, R. Hellinghuizen, F. Hoeksema, and G. Heideman, "Transcoding of MPEG bitstreams," *Signal Processing:Image Communication*, Vol. 8, pp. 481-500, 1996.

[2] H. Sun, W. Kwok, and W. Zdepski, "Architectures for MPEG compressed bitstream scaling," *IEEE Trans. Circuits Syst. Video Technol.*, Vol. 6, No. 2, pp. 191-199, April 1996.

[3] P.A.A. Assuncao, and M. Ghanbari, "Post-processing of MPEG2 coded video for transmission at lower bit rates," *IEEE ICASSP 1996*.

[4] M. Sugano, Y. Nakajima, H. Yanagihara, and A. Yoneyama, "An efficient transcoding from MPEG-2 to MPEG-1," *IEEE ICIP 2001*, pp. 417-420.

[5] N. Bjork, and C. Christopoulos, "Transcoder architectures for video coding," *IEEE Trans. Consumer Electronics*, Vol. 44, No. 1, pp. 88-98, Feb. 1998.

[6] P.A.A. Assuncao, and M. Ghanbari, "A frequency-domain video transcoder for dynamic bit-rate reduction of MPEG-2 bit streams," *IEEE Trans. Circuits Syst. Video Technol.*, Vol. 8, No. 8, pp. 953-967, Dec. 1998.

[7] S. F. Chang, D. G. Messerschmitt, "Manipulation and compositing of MC-DCT compressed video", *IEEE JSAC special issue on intelligent signal processing, 1994.*

[8] N. Merhav and V. Bhaskaran, "Fast algorithms for DCT-domain image down-sampling and for inverse motion compensation", *IEEE Trans. Circuits Syst. Video Technol.*, Vol. 7, No. 3, pp. 468-476, June 1997.

[9] J. Song and B. L. Yeo, "Fast extraction of spatially reduced image sequences from MPEG-2 compressed video", *IEEE Trans. Circuits Syst. Video Technol.*, Vol. 9, No. 7, pp. 1100-1114, Oct. 1999.

[10] B. Shen, I. K. Sethi, and B. Vasudev, "Adaptive motion-vector resampling for compressed video downscaling," *IEEE Trans. Circuits Syst. Video Technol.*, Vol. 9, No. 6, pp. 929-936, Sep 1999.

[11] M.R. Hashemi, L. Winger, and S. Panchanathan, "Compressed domain motion vector resampling for down-scaling of MPEG video," *IEEE ICIP 1999*.

[12] P. Yin, M. Wu, and B. Liu, "Video transcoding by reducing spatial resolution," *IEEE ICIP 2000.*

[13] A. Yomeyama, Y. Hizume, and Y. Nakajima, "Fast dissolve operations for MPEG video contents," *IEEE ICIP 2000.*

[14] G. Shen, B. Zeng, Y. Q. Zhang, and M. L. Liou, “Transcoder with arbitrarily resizing capability,” *IEEE Symp. Circuits and System*, vol. 5, pp. 25-28, 2001.

[15] T. Shanableh, and M. Ghanbari, “Heterogeneous video transcoding to lower spatio-temporal resolutions and different encoding formats,” *IEEE Trans. Multimedia*, Vol. 2, No. 2, pp. 101-110, June 2000.

[16] J. Xin, M. T. Sun, B. S. Choi and K. W. Chun, “An HDTV-toSDTV spatial transcoder”, *IEEE Trans. Circuits Syst. Video Technol.*, Vol. 12, No. 11, pp. 998-1008, Nov. 2002.

[17] Y. Shoham, and A. Gersho, “Efficient bit allocation for an arbitrary set of quantizers,” *IEEE Trans. Acoustics Speech and Signal Processing*, Vol. 36, No. 9, pp. 1445-1453, September 1988.