Caption Processing for MPEG Video in MC-DCT Compressed Domain*

Jongho Nang, Ohyeong Kwon, Seungwook Hong
Dept. of Computer Science and Engineering, Sogang University
1 Shinsoo-Dong, Mapo-Ku, Seoul 121-742, Korea
+82-2-705-8494
jhnang@ccs.sogang.ac.kr

ABSTRACT

The (cinema) caption processing that adds descriptive texts on the sequence of frames is an important video manipulation function that video editor should support. This paper proposes an efficient MC-DCT compressed domain approach to insert the caption into the MPEG-compressed video stream. It basically adds the DCT blocks of the caption image to the corresponding DCT blocks of the input frames one by one in MC-DCT domain as in [5]. However, the strength of the caption image is adjusted in the DCT domain to prevent the resulting DCT coefficients from exceeding the maximum value that is allowed in MPEG. In order to adjust the strength of caption image adaptively, we should know the exact pixel values of input image that is a difficult task in DCT domain. We propose an approximation scheme for the pixel values in which the DC value of a block is used as the expected pixel value for all pixels in that block. Although this approximation may lead some errors in the caption area, it still provides a relatively high image quality in the non-caption area, while the processing time is about 4.9 times faster than the decode-captioning-reencode approach.

Keywords

Caption Processing, MC-DCT Domain, MPEG Video, MPEG Editing

1. INTRODUCTION

The (cinema) caption processing (or *subtitl*ing) which adds some descriptive information in a text form on the top of the sequence of the frames is an important video manipulation function that the video editor should support. The simplest approach to insert the caption for the MC-DCT (Motion Compensation – Discrete Cosine Transform) compressed video is to fully decode the video first, add the caption into the video data in a raw format, and then re-encode the resulting captioned video to the MC-DCT compressed form again. Since this decode-captioning-reencode approach requires a huge amount of computational and storage resources, there have been some researches [1, 2, 3, 4, 5, 6, 9] that directly manipulate the video in the MC-DCT domain. Among these researches, Meng's scheme [5] which inserts the visible watermark mask into the MC-DCT compressed video is a candidate that could be directly used for the caption processing, because basically the caption processing and embedding visible watermark require the same functionality that overlaps one image to another transparently or non-transparently. However, since usually the strength (or distinctness) of the caption is higher than that of visible watermark, the direct application of this scheme to caption processing may cause a problem in the referencing frame in MC-DCT compressed videos. This problem comes from the fact that they did not consider the case in which the strength of the visible watermark mask is too high so that the sum of original pixel value and watermark mask value may exceed a maximum value that is allowed in the MC-DCT compressed video. In this case, the resulting value is normalized to its maximum value and stored in the anchor frame. However, when this resulting (normalized) anchor frame is reconstructed to be referenced by the successive frames for their encodings, the watermark mask value should be subtracted from the stored value in order to build the original anchor frame. Since the watermarked value in the anchor frame is already normalized to its maximum value, the original image could not be reconstructed precisely so that there might be some data blocks with errors in the referencing frames.

* This work is supported by KOSEF (Project Title : "*Development of Non-Linear Editor for MPEG Video Stream*," 1998.9 ~ 2000.8)

This paper proposes an efficient caption processing scheme that inserts the descriptive information into the MPEG compressed video frames in the MC-DCT compressed domain directly. Basically, it adopts the visible watermark embedding scheme proposed in [5] so that the DCT blocks of caption image are added to the DCT frames of the input video in the DCT or MC-DCT domain. In this scheme, for I frame or intracoded blocks of P- and B-frame, the DCT blocks of caption images are added directly to the DCT blocks of input frames. On the other hand, for the intercoded blocks in P- and B-frame, the DCT blocks of caption image added in the anchor frames are subtracted before adding the DCT block of current caption image to the DCT error residual blocks of current frame as in [5]. However, since just adding the DCT blocks of caption image to the DCT block of input anchor frame may cause a problem as mentioned before, the DCT coefficients of caption block are adjusted with respect to the DCT coefficients of corresponding target image block. This adjustment is designed to have a property that the sum of luminance values of pixel at (n,m) in the caption image and the input frame does not exceed the maximum allowed value. However, since we could not figure out the exact luminance value for each pixel in the input frame in DCT domain, the average of the luminance values of all pixels in the block of input video frame is used as an approximated luminance value for all pixels in that block. This average value is actually the DC coefficient of the block in the DCT domain, and could be obtained in the DCT domain easily. The proposed adjustment and approximation scheme contribute to minimize the errors in the caption processing for the both of intracoded and intercoded blocks in MC-DCT domain, while keeping the decoding and encoding overhead to be minimal. Upon our experiments on the caption processing for the MPEG video streams, although this adjustment may lead some errors in the caption areas because of the approximation of pixel luminance value, it still provides a relatively high image quality in the non-caption area, while the processing time is about 4.9 times faster than the decode-captioning-reencode approach. The caption processing scheme proposed in this paper could be used to insert the strong visible watermark or caption into the MPEG compressed video streams efficiently while keeping the video quality as high as possible with minimal decoding overheads.

2. Previous Works

The caption used in the TV programs and movies is either opaque or transparent and usually surrounded with black borders to maximize the visibility of the characters as shown in Figure 1. The process that inserts this caption to video frame in the spatial domain is formalized as follows [2, 9];

$$P_{new}(i,j) = \alpha(i,j) \times P_a(i,j) + (1-\alpha(i,j)) \times P_b(i,j) \quad (1)$$

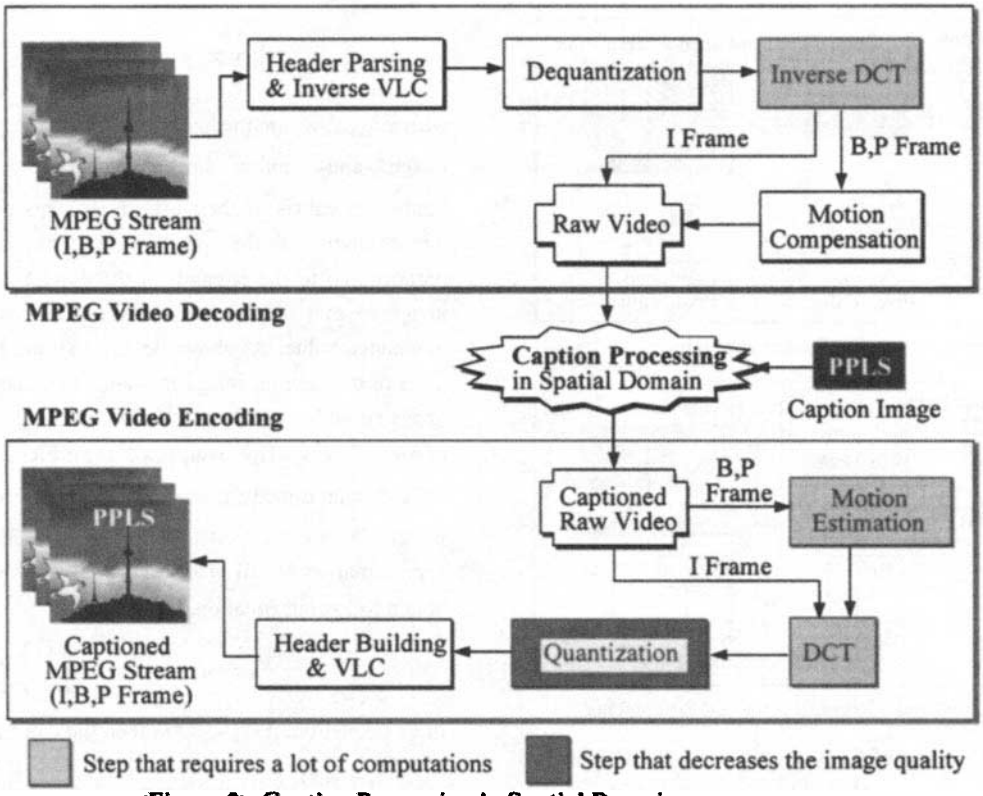
where P_{new}, P_a, P_b, and $\alpha(i,j)$ are the captioned video frame, the input video frame, the caption image, and the transparency factor $(0 \le \alpha(i,j) \le 1)$, respectively. If $\alpha(i,j)$ is 1, it is an opaque caption. Otherwise, it is a transparent caption.

<Figure 1> An Example of Captioned Video Frame

The simplest way to insert the caption into an MPEG video is to fully decode the MPEG video stream to raw data, add the captions using Eq. (1) in the spatial domain, and reencode the resulting captioned video to MPEG video streams as shown in Figure 2. However, this approach requires not only a huge amout of computational resources for DCT/IDCT and motion estimation but also a large storages for storing the temporal raw video data. Furthermore, the image qualities of resulting captioned MPEG streams may be degraded because of the repeated quantization and dequantization processes. It has stimulated several researches [2, 3, 9, 10] on the manipulation of images and video streams in DCT or MC-DCT domain directly. Among these researches, there was an approach [8, 9] based on the convolution that directly mainipulated the compressed image and videos in DCT or MC-DCT domain. This approach could be formulated as follows;

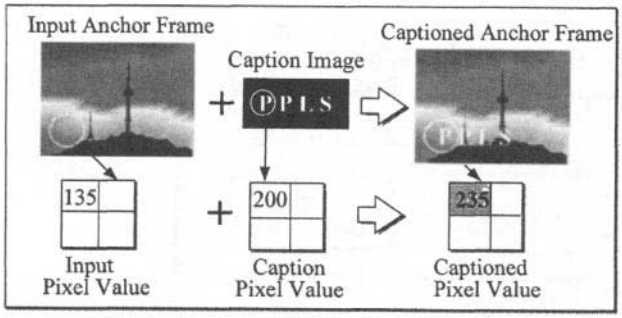
$$DCT(P_{new}) = DCT(\alpha) \otimes DCT(P_a) + DCT(1-\alpha) \otimes DCT(P_b) \quad (2)$$

where *DCT(A)* represents the result of the DCT transformation of block A, and $\otimes$ represents the convolution in the DCT domain. It is actually a DCT domain representation of Eq. (1). Although there are some researches [9] to optimize the convolution process in DCT domain, it still requires too many computations to be used in the real environments.

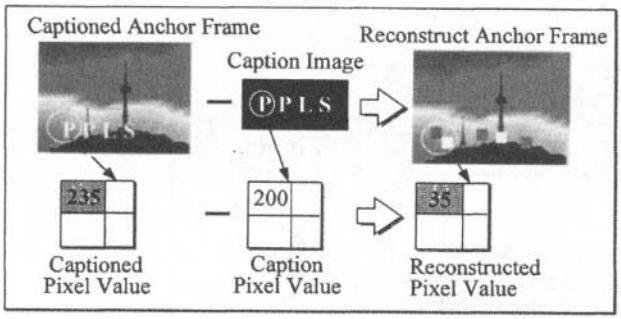


<Figure 2> Caption Processing in Spatial Domain

Recently, there was an approach [5] to embed the visible watermark to MPEG compressed video in the MC-DCT without the complex convolution operations. The key idea of this scheme is to adjust the strength of the watermark block by block (not pixel by pixel) adaptively with respect to the strength of the blocks in the input video frames so that the resulting watermark will have constant visibility. They show that this operation could be performed in MC-DCT domain directly using the motion compensation technique in DCT domain [2]. Since the embedding of the visible watermark and the caption processing requires the same functionality that overlaps one image to another transparently or non-transparently, we can use this scheme to insert the caption to MPEG video streams in MC-DCT domain. However, since usually the strength of the caption is stronger than that of watermark mask, a direct application of this scheme to caption processing may cause an artifact in the resulting captioned video streams. Let us explain this problem in more detail via an example in the MC-Spatial domain as shown in Figure 3. If the sum of luminance values of pixel at (n,m) of target video frame and the watermark image gets out of the luminance range allowed in MPEG which is [16, 235], the luminance value at (n,m) of resulting watermarked frame is normalized to its maximum value 235 (Figure 3-(a)). It causes a problem when this pixel is referenced by the successive frame in which the added watermark luminance value is being subtracted again to reconstruct the original anchor frame without watermark. Since the resulting luminance value of the pixel in watermarked image is already normalized to its maximum value 235, the original reference frame could not be reconstructed properly with just subtracting the watermark value from the watermarked frame (Figure 3-(b)). If the strength of the watermark is weak enough (as in embedding visible watermark) so that the result of the addition does not exceed the maximum value, then there would be no problem. However, if the strength of the watermark is strong enough (as in the caption processing), the strength of the watermark should be adjusted adaptively so that the result of the addition would not exceed the maximum value. This problem is the same in the inverse motion compensation in the MC-DCT domain. This paper proposes a scheme to adaptively adjust the strength of the caption image in DCT domain so that the summation does not exceed the allowed range. Of course, Meng and Chang [5] proposed a scaling scheme that adaptively adjusts the strength of the watermark mask based on the local image content, but it was to provide a constant visibility of resulting watermark, not to prevent the errors in the anchor frame reconstruction process.



(a) Caption Processing for Anchor Frame



(b) Anchor Frame Reconstruction for Referencing

<Figure 3> An Example of Caption Processing in MC-Spatial Domain

3. A NEW CAPTION PROCESSING in MC-DCT DOMAIN with DC IMAGE

In this section, we propose a caption strength adjustment scheme according to the strength of the input frame in DCT or MC-DCT domain to minimize the errors in the motion compensation process.

3.1 Adaptive Caption Scaling with DC Image

For the errorless reconstruction of the anchor frames, the summation of the luminance values of the pixels in the captioned and input images should not exceed the maximum value allowed in MPEG that is 235 as explained in Figure 3. One way to adaptively adjust the strength of caption image is to decrease the strength of pixel in the caption image if the luminance of the corresponding pixel in the input image is high enough so that the summation would not exceed the maximum value. On the other hand, if the luminance of the corresponding pixel is low enough, the strength of the pixel in caption image is not changed. Since the luminance of pixel in the text area of the caption image usually takes the highest value (that is 235 in MPEG), this mechanism would produce the captioned image in which the luminance of the pixels in the caption area have their maximum values and others have their original values in the input image. One way to satisfy these requirements could be formulated as follows:

$$x'_i = x_i + w_i * \left(1 - \frac{x_i}{235}\right) \tag{3}$$

where x', x, w are the luminance block in captioned image, input image, and caption image, respectively, and x'_i, x_i, w_i are luminance values of their i-th pixels. This equation implies that the strength of the pixel in caption image is decreased proportional to the strength of the corresponding pixel in input image so that the summation would not exceed the maximum luminance value. As shown in Eq. (3), the strength of the each pixel in the caption image that are added adaptively to the input image could be calculated exactly if we know the luminance value of each pixel x_i. However, since it is difficult to know its value in DCT domain directly, the average luminance value of the pixels in the block x, $e(x)$, is used in the proposed scheme as an approximation for all pixels in x. Using this approximation, we obtain following equation from Eq. (3);

$$x'_i = x_i + w_i * \left(1 - \frac{e(x)}{235}\right) \tag{4}$$

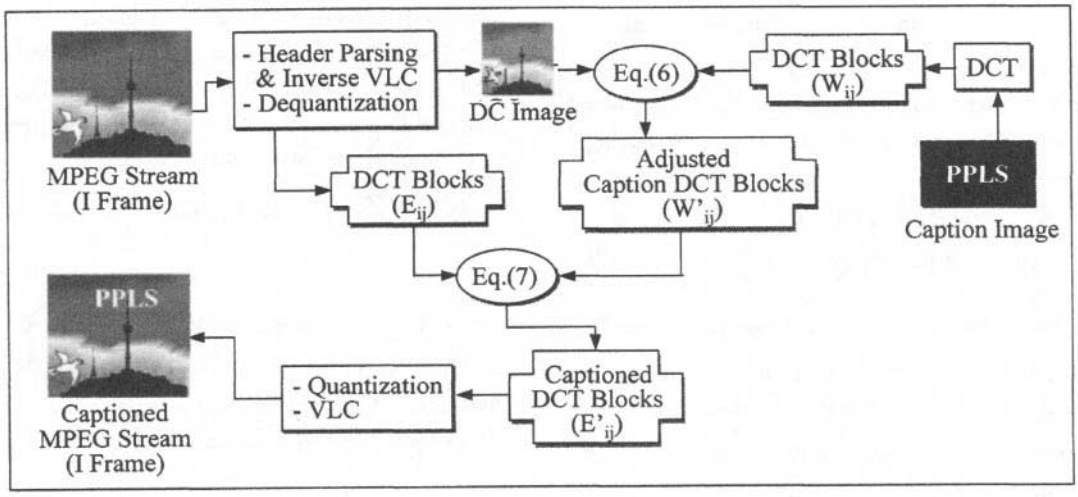
In this equation, if $x_i < e(x)$, then the computed x'_i is as less as $w_i * \frac{x_i - e(x)}{235}$ than the exact value computed using Eq. (3). However, since the resulting value still does not exceed the maximum value, it would be no problem in the reconstruction phase. On the other hand, if $x_i > e(x)$, then the actual x'_i is as bigger as $w_i * \frac{x_i - e(x)}{235}$ than the exact value computed using Eq. (3). In this case, since the resulting value exceeds the maximum value, it may cause an error again in the reconstruction phase. The formulae in Eq. (4) could be transform to DCT domain as follows;

$$X' = X + W * \left(1 - \frac{DC(X)}{1880}\right) \tag{5}$$

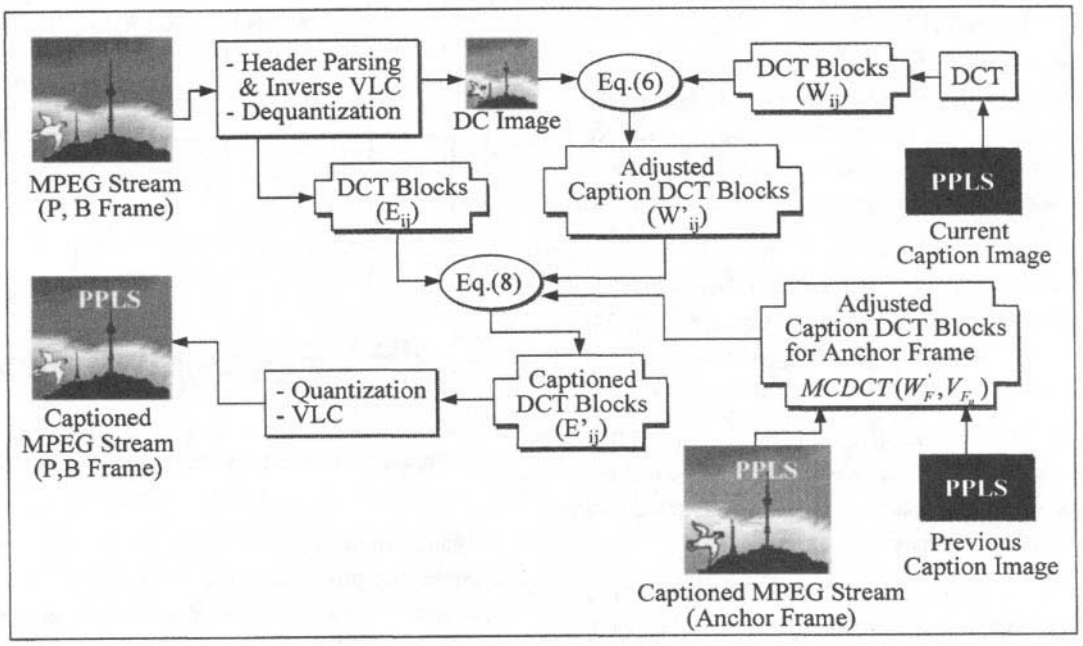
where X', X, W are the DCT block of x', x, w, respectively, and $DC(X)$ is the DC coefficient of block x that can be easily obtained in the DCT domain.

3.2 Inserting Caption in DCT and MC-DCT Domain

The caption image is converted to a gray scale image because the caption is only added to luminance channel of the input video frame. All pixels in the transparent region are set to 0 so that the addition of their values to target image pixels has no effects on the captioned image. Let W' be the DCT coefficients of the block in caption image that are adjusted and added to the input image;



<Figure 4> Caption Processing in DCT Domain : I-Frame



<Figure 5> Caption Processing in MC-DCT Domain : P- and B-Frame

$$W' = W * \left(1 - \frac{DC(X)}{1880}\right) \tag{6}$$

Once these DCT coefficients are computed for each block, they are inserted into the DCT frames of the input video differently for each of three macroblock types. This process is the same as the one for embedding the visible watermark in MC-DCT domain proposed in [5]. Let us explain this process roughly. For I-frame or intracoded blocks in B- or P-frames, the DCT of scaled watermark mask is added directly,

$$E_{ij}^{'} = E_{ij} + W_{ij}^{'} \tag{7}$$

where $E_{ij}^{'}$, E_{ij}, and $W_{ij}^{'}$ are the ij-th DCT blocks of the captioned frame, the input frame, and the adjusted caption image, respectively. This process is performed in DCT domain directly as shown in Figure 4.

For the blocks with forward motion vector in P frame (or backward motion vector in B frame), the caption image added in the anchor frame needs to be subtracted before adding the current caption image in the current frame. The resulting residual is,

$$E_{ij}^{'} = E_{ij} - MCDCT(W_F^{'}, V_{F_{ij}}) + W_{ij}^{'} \tag{8}$$

where $MCDCT()$ is the motion compensation function performed in DCT domain as described in [2]. $W_F^{'}$ is the caption DCT used in the forward anchor frame, and $V_{F_{ij}}$ is the motion

vector. E_{ij} and E'_{ij} are the original and new motion compensation residual errors, respectively. For bidirectional predicted blocks in B frame, both forward and backward motion compensation needs to be averaged and subtracted while adding the current caption as follows;

$$E'_{ij} = E_{ij} - (MCDCT(W'_F, V_{F_{ij}}) + MCDCT(W'_B, V_{B_{ij}}))/2 + W'_{ij} \quad (9)$$

where $V_{F_{ij}}$ and $V_{B_{ij}}$ is the forward and backward motion vector respectively. This process is performed in the MC-DCT domain as shown in Figure 5. Note that these processing mechanisms are basically the same as the one proposed in [5] except that W'_{ij} is computed differently.

4. Experimental Analyses

We have implemented the proposed caption processing scheme, and evaluated the performance with respect to the produced image quality and its speed. For the performance comparisons, we have also implemented the following schemes that could be used to insert the caption to MPEG compressed video;

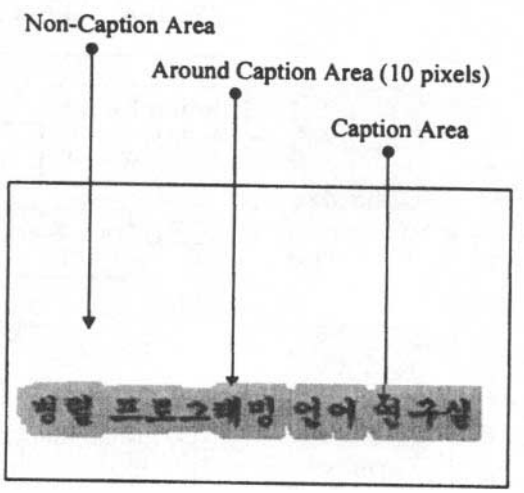
- *Processing in totally Spatial domain* (Spatial Scheme): It is a decode-captioning-reencode method in which the MPEG compressed video is fully decoded, the caption is inserted in the spatial domain, and reencoded to MPEG video.
- *Processing in DCT domain* (DCT Scheme) : The MPEG compressed video with I-, P-, and B-frames is converted into MPEG video with only I-frames, and the caption processing is performed in the DCT domain.
- *Processing in MC-DCT domain* (MC-DCT Scheme): The caption is inserted into MPEG compressed video in MC-DCT domain using the method proposed in [5].
- *Processing in MC-DCT domain with DC Image* (Ours : MC-DCT(DC) scheme) : It is the same as the one proposed in [5], but the strength of the caption image is adjusted according to its DC image to prevent the errors in the motion compensation using Eq. (5).

All of above four implementations read the same MPEG stream, process the captions, and output the captioned MPEG stream. The GOP pattern of MPEG stream used in the experiments is IBBPBBPBBPBBPBB.

Let us first compare the image qualities of captioned MPEG streams produced by above four schemes. The resulting four captioned MPEG streams are decoded to PPM files with the public domain decoder [7], and compared with the captioned images that are generated by decoding the input MPEG streams to PPM files and inserting the caption in the spatial domain. For the image quality comparison, we have used PSNR (Peak Signal-to-Noise Ratio) values that are defined as follows;

$$PSNR(X,Y) = 20\log_{10}\frac{255}{\sqrt{\frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N}(X_{ij}-Y_{ij})^2}} \; dB$$

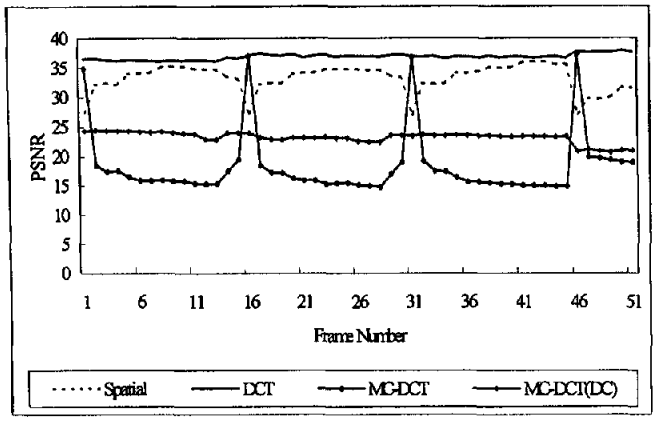
where M and N are the sizes of the image, and X_{ij} and Y_{ij} are the (i,j)-th pixels in the image X and Y, respectively. The comparison of the image qualities is performed with respect to the areas that are influenced by the caption processing as shown in Figure 6.



<Figure 6> Example of Image Divisions for Quality Comparisons

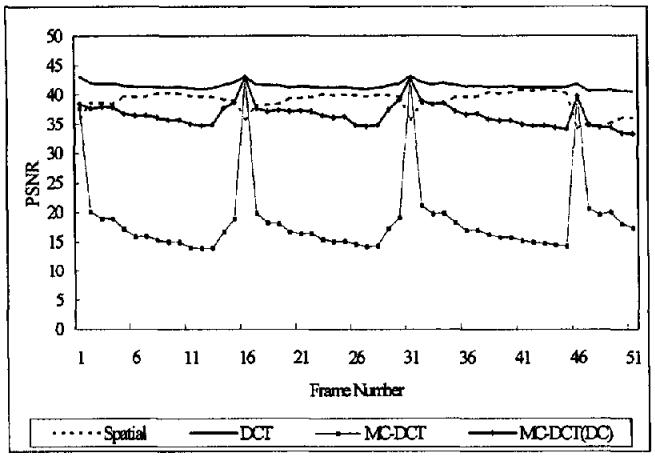
Figure 7 shows an experimental image quality comparison of the four caption processing schemes with respect to the three areas, while Figure 8 shows their examples visually. Let us analyze these experimental results in more details. Figure 7-(a) shows the experimental PSNR values of four caption processing schemes in the caption area. As shown in this figure, our approach (MC-DCT (DC)) produces a lower PSNR than the spatial and DCT domain approaches in caption area because it adaptively decreases the strength of the caption image according to the contents of the input image frame. On the other hand, it produces a higher PSNR value except I-frame than MC-DCT approach because it adjusts the strength of the caption image not to exceed the maximum value so that it could minimize the errors in the anchor frame reconstruction process. The image quality of the around caption area (10 pixels far from the caption) is also compared, and its experimental results is shown in Figure 7-(b). As shown in this figure, all processing schemes except Meng's scheme (MC-DCT) produce almost the same PSNR values. It is mainly because the

Meng's scheme did not adjust the caption strengths with respect to the strength of the input video frame image as explained in Figure 3. However, for the I-frames, both of the MC-DCT and MC-DCT(DC) schemes produce the same PSNR values because

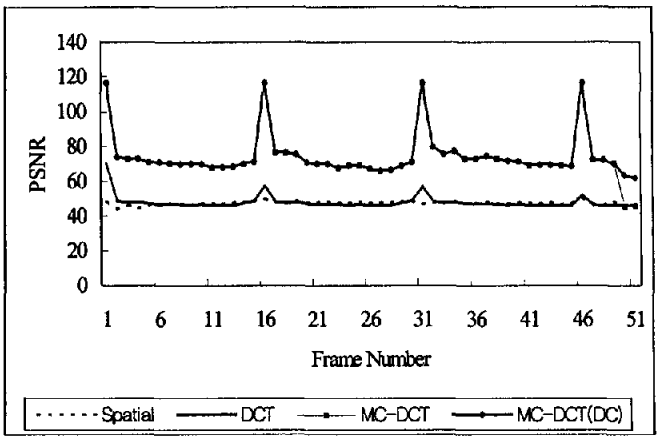


there is no motion compensation process in these frames.

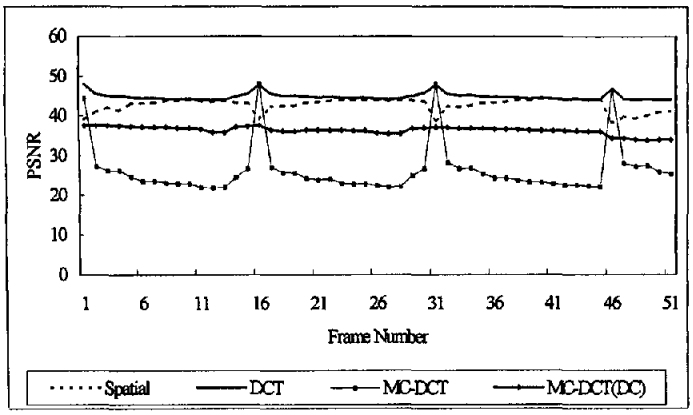
(a) Image Quality in Caption Area



(b) Image Quality in Around Caption Area



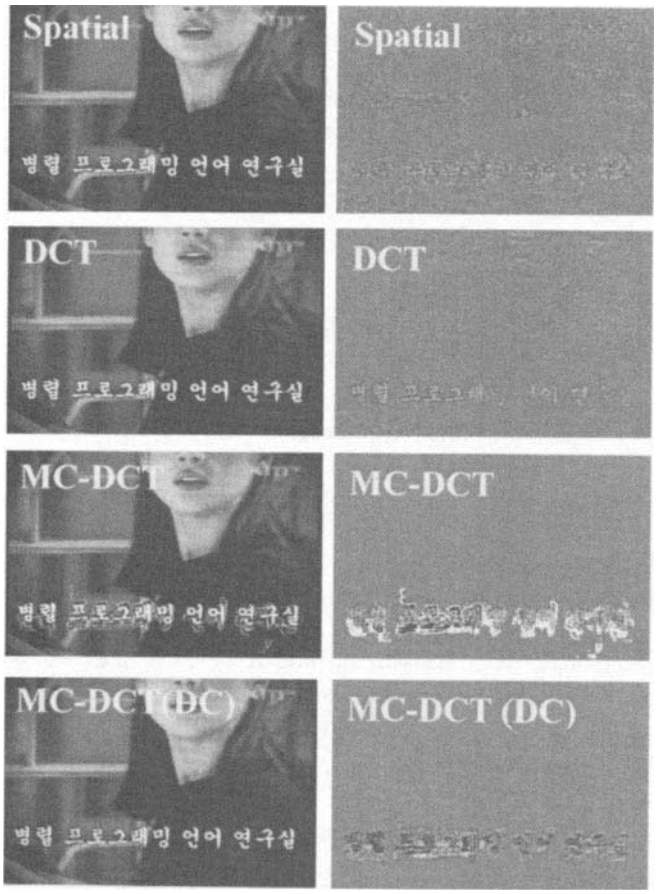
(c) Image Quality in Non Caption Area



(d) Overall Image Quality

<Figure 7> An Experimental Image Quality Comparison

Since the non-caption area is not affected by the caption processing, theoretically there should be no image quality degradation in this area. However, since the spatial domain approach requires a full decoding of MPEG stream for the caption processing (VLC-Dequantization-MC-IDCT-CaptionProcessing-DCT-MC-Quantization-VLC), the image quality of the non-caption area is also degraded as shown in Figure 7-(c). Furthermore, the DCT domain approach also requires to convert the P- and B-frame to I-frame for caption processing and it requires to decode the video to the spatial domain, it also introduces a quality degradation. On the other hand, MC-DCT and MC-DCT(DC) approaches do not require the full decoding of MPEG stream for caption processing, the image quality could be kept in the non caption area as shown in Figure 7-(c). Actually, all steps except the quantization/dequantization are theoretically lossless compressions so that there would be no more errors in the other processes, but practically small errors in the DCT/IDCT could introduce a non-negligible error in quantization/ dequantization process. It is the reason why MC-DCT and MC-DCT(DC) approach could produce a high image quality in the non-caption area. Figure 7-(d) shows the overall image qualities of the four caption processing schemes. As shown in this figure, the proposed scheme (MC-DCT(DC) approach) may produce a higher image quality than MC-DCT approach, and a lower image quality than spatial and DCT approaches. But since it processes the caption in MC-DCT domain so that could skip the DCT/IDCT and motion estimation process, the execution time is about 4.9 times faster than that of spatial domain approach as shown in Table 1. It shows the average caption processing time per frame of four schemes on Pentium PC.



<Figure 8> Caption Processing Results (Left : Captioned Image, Right : Embedded Error Image)

<Table 1> Experimental Average Processing Time (Sec/Frame)

	Spatial	DCT	MC-DCT	MC-DCT(DC)
Time	0.53	0.41	0.11	0.11

5. CONCLUDING REMARKS

This paper proposes a caption processing scheme that could be directly applied in the MC-DCT domain while keeping the image quality as high as possible. This scheme basically follows the MC-DCT domain approach proposed in [5], however, it adjusts the strength of the caption image in order not to exceed the maximum value allowed in MPEG. To adjust the strength of the caption image, it uses the DC image of the input frame as an approximation of the source image. Since this DC value can be obtained in MC-DCT domain easily, the proposed scheme could adjust the caption strength easily while keeping the decoding overhead as small as possible. Upon on the experimental results, we could argue that although there are still some errors in the captioned image because of the approximation scheme, the proposed scheme could produce a higher quality captioned MPEG stream than the other MC-DCT domain scheme, while keeping the caption processing time to be about 4.9 times faster than the spatial domain approach.

6. References

[1] Soam Acharya and Brian Smith, "Compressed Domain Transcoding of MPEG," *Proc. of IEEE International Conference on Multimedia Computing and Systems*, Jul. 1998.

[2] Shih-Fu Chang and David G. Messerschmitt, "Manipulation and Compositing of MC-DCT Compressed Video," *IEEE Journal on Selected Areas in Communications*, Vol. 13, No. 1, Jan. 1995.

[3] Vrkrant Kobla and David Doerman, "Compressed Domain Video Indexing Techniques Using DCT and Motion Vector Information in MPEG Video," *Proc. of SPIE Conf. on Storage and Retrieval for Image and Video Databases V*, Vol. 3022, Feb.1997.

[4] Jianhao Meng and Shih-Fu Chang, "CVEPS: A Compressed Video Editing and Parsing System," *Proc. of ACM Multimedia 96 Conference*, Nov. 1996.

[5] Jianhao Meng and Shih-Fu Chang, "Embedding Visible Video Watermarks in the Compressed Domain," *Proc. of ICIP International Conference on Image Processing*, Oct. 1998.

[6] Neri Merhav and Vasudev Bhaskaran, *A Fast Algorithm for DCT-Domain Inverse Motion Compensation*, HPL Technical Report #HPL-95-17, Sep. 1995.

[7] MPEG Software Simulation Group, MPEG-2 Video Codec, Available from http://www.mpeg.org/MPEG/MSSG.

[8] Bo Shen and Ishwar K. Sethi, "Inner-Block Operations On Compressed Images," *Proc. of ACM Multimedia 95 Conference*, Nov. 1995.

[9] Bo Shen, Ishwar K. Sethi and Vasudev Bhaskaran, "DCT Convolution and Its Application in Compressed Video Editing," *Proc. of SPIE 3024 in Visual Communications and Image Processing*, Feb. 1997.

[10] Bo Shen, Ishwar K. Sethi, and V. Bhaskaran, "Closed-loop MPEG Video Rendering," *Proc. of IEEE Conference on Multimedia Computing Systems*, Jun. 1997.