Proceedings of the 28th IEEE
EMBS Annual International Conference
New York City, USA, Aug 30-Sept 3, 2006

Synthesizing Multi-View Video Frames for Coding Patient Monitoring Video

Renbin Peng, *Student Member*, *IEEE*, Robert J. Sclabassi, *Senior Member*, *IEEE*, Qiang Liu, *Member*, *IEEE*, Gusphyl Justin, *Student Member*, *IEEE*, and Mingui Sun, *Senior Member*, *IEEE*

Abstract—**This paper presents a method to synthesize multi-view video frames to facilitate coding and transmission of patient monitoring video. The synthesis is carried out in the DCT domain by means of interlacing. The synthesized video provides a higher video coding efficiency, better synchronization of the video streams from multiple cameras, as well as the improved data loss resilience and protection of the video content. The viability of the presented method was demonstrated by experimental results on patient monitoring video.**

I. Introduction

VIDEO recording via multiple cameras has become a popular practice for many applications. Patient monitoring, for example, is directly benefited from this technology. In this application, the patient images are recorded from different viewpoints at the same time to reduce occlusions. However, an immediate issue raised from multi-camera recording is the video coding scheme for both archiving and transmission purposes. Coding the multiple video streams independently can be a tremendous waste of bandwidth, and the synchronization of these streams may also be problematic. Research has been conducted targeting multi-view video coding. A variety of methods have been proposed based on free viewpoint video communications[1][2], using a hierarchical light field to facilitate immersive viewing [3], and using a wavelet-based codec to improve rate-distortion performance [4]. These previous works were either intended to utilize the MPEG-2/MPEG-4/H264 multi-view profile to achieve higher compression efficiency, or tried to formulate a new coding structure specifically designed for multi-view images. In these methods, the video frames from different cameras are cross-referenced and the disparities between them are explicitly utilized to reduce the redundancies between cameras. The resulting approach in this respect can be highly complex in its implementation, and requires accurate disparity maps that are often hard to obtain.

Against the backdrop of these concerns, we explore a “mono-view” coding scheme that codes multi-view video frames after synthesizing the views into a single image. The regular video codecs can be directly applied to the synthesized video frames. Additionally, by using this method, the synchronization problem regarding the multiple cameras is automatically resolved, and the reconstructed images are more resilient to certain losses of data during transmission. This scheme is accomplished by synthesizing patient monitoring video frames in a transform domain. These video frames, recording the same patient from different viewpoints, possess common features in the frequency domain. Therefore, we employ a DCT-coefficient interlacing technique to synthesize the frames without requiring an accurate geometric mapping between the cameras. In addition, this process rearranges the contents contained in the multiple images, achieving an improved content protection. We have conducted experiments of this method using patient monitoring video clips.

II. Methodology

In order to show the input data forms, we demonstrate in Fig.1 four images of a patient monitoring scene recorded from different viewpoints.

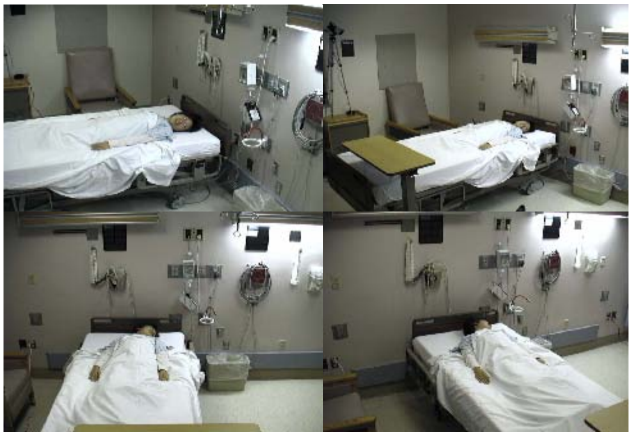

Fig.1. Patient monitoring images recorded from four viewpoints

It is clear that these multiple views possess a high content similarity, which is visually justifiable. Examining in the frequency domain, one can observe a similarity between their spectrum, shown in Fig. 2. This implies that the correlations between cameras can be reflected by their spectrum. Therefore, instead of explicitly registering the images in the spatial domain, which is often computationally expensive, one may exploit the image spectrum to manage the camera correlation in an implicit manner. To realize this concept, we explore a simple method that synthesizes the multiple images in the DCT domain. Firstly, we aim at synthesizing the images so that the decorrelation between cameras can be performed by the motion estimation in the regular mono-view codec. This is motivated by previous works [1][2] that carries out motion estimation between video frames from different cameras. After synthesizing the multiple views into one, block matching (in the motion estimation process) between the synthesized video frames is potentially carried out between

1-4244-0033-3/06/$20.00 ©2006 IEEE.

different cameras, therefore leading to a camera decorrelation as well as temporal decorrelation for the combined single video stream. Secondly, we exploit the similarities in spectral components among different images. This is because that the spectrum provides the power spectral density (PSD) information of the image if it is modeled as a random process, and a smoother PSD suggests a "sharper" covariance function, which implies a decorrelation between pixels. To smooth the spectrum, we simply interlace the transform coefficients of different images into one, taking advantage that their spectrum has a strong similarity. The details of interlacing is exemplified in Fig. 3. We should mention that besides the coding efficiency that may benefit from the synthesizing approach, the synchronization problem of the multiple video streams is also solved by this method. In addition, the manipulation in the frequency domain rearranges the information in the spatial domain, which also benefits content protection (shown in Fig. 4).

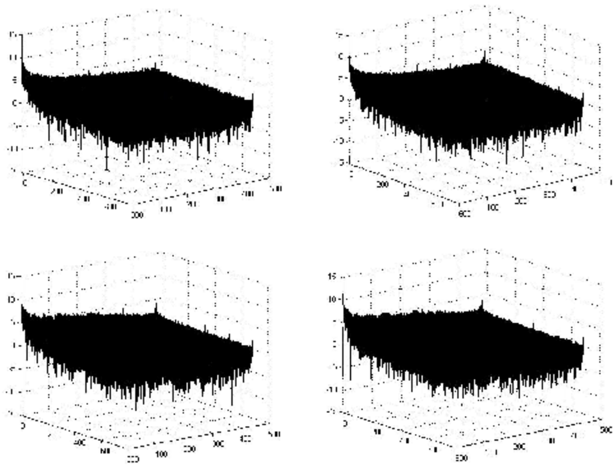

Fig.2. Frequency spectrum of four images, where we rescale the amplitudes of the spectral elements.

A_{00}	B_{00}	A_{01}	B_{01}	A_{02}	B_{02}	…	B_{0n}
C_{00}	D_{00}	C_{01}	D_{01}	C_{02}	D_{02}	…	D_{1n}
A_{10}	B_{10}	A_{11}	B_{11}	A_{12}	B_{12}	…	B_{2n}
C_{10}	D_{10}	C_{11}	D_{11}	C_{12}	D_{12}	…	D_{2n}
...	…	…	…	…	…	…	…
C_{m0}	D_{m0}	C_{m1}	D_{m1}	C_{m2}	D_{m2}	…	D_{mn}

Fig. 3. Interlacing method in the DCT domain, where A_{ij}, B_{ij} , C_{ij} and D_{ij} are the entries in the i^th^ row and j^th^ column of the DCT coefficients matrices A, B, C, and D, respectively, and A, B, C and D denote the DCT coefficient matrices of the four images.

III. Analysis

In this section, we show how the interlacing method may benefit video coding. First, we prove, for the most general case, that the presented method will not cause information loss or introduce any redundancy to the original multi-view video streams. According to Data Processing Theorem [5], reversible transforms do not result in information loss.

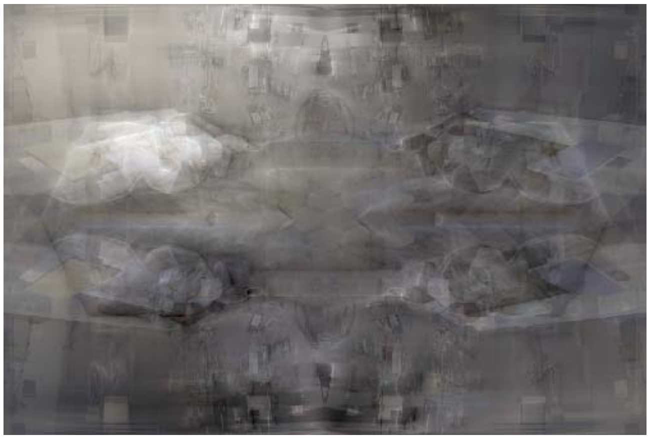

Fig.4. Synthesized 960×1440 image containing the information in the four images in Fig. 1. It is difficult to visualize the information in the original multi-view images, providing both synchronization and content protection.

Proof :

For aggregates *X*, *Y* and *Z*, let $y = f(z)$ .Then

$$\begin{aligned} I(X;YZ) &= I(X;Y) + I(X;Z \mid Y) \\ &= I(X;Z) + I(X;Y \mid Z) \end{aligned} \tag{1}$$

where $I(X;Y)$ is the mutual information with regard to *X* and *Y*. $I(X;Z \mid Y)$ is the condition mutual information with regard to *X* and *Z* given *Y*. So

$$I(X;Z) = I(X;Y) + I(X;Z \mid Y) - I(X;Y \mid Z) \tag{2}$$

For the systems in Fig.5, $I(X;Y \mid Z) = 0$. So,

$$I(X;Z) = I(X;Y) + I(X;Z \mid Y) \tag{3}$$

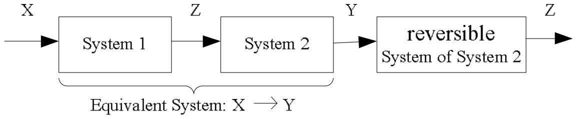

Fig.5. Reversible transforms don't result in information loss.

Because the conditional mutual information is nonnegative, i.e. $I(X;Z \mid Y) \geq 0$ [6], one has

$$I(X;Z) \geq I(X;Y) \tag{4}$$

If there exists an inverse transform f^{-1} , such that $z = f^{-1}(y)$, then similar to the above discussion, we have $I(X;Z) \leq I(X;Y)$. Thus, $I(X;Z) = I(X;Y)$, that is, the reversible transforms do not result in the loss of information. The presented interlacing method is just a kind of reversible transform. Specifically, the DCT is an orthogonal transform, which will not introduce redundancy to the transformed data. Therefore, the amount of information contained in the original and transformed signal remains equivalent.

In the following, we derive the DCT domain interlacing result. For simplicity, we utilize DFT to induce the DCT result. Rigorously speaking, DCT is not the real part of the DFT, but they have close connections [7][8]. Although the following discussion is based on the DFT, an analogical conclusion applies to the DCT case, which has been validated by our experiments.

Using the DFT, the synthesized image resulting from the interlacing technique illustrated in Fig. 3 is given by

$$u(m,n)=\frac{1}{4MN}\{a(m,n)+b(m,n)e^{j\pi n/N} + c(m,n)e^{j\pi m/M} + d(m,n)e^{j\pi(m/M+n/N)}\} \tag{5}$$

if $0 \le m \le M-1$, and $0 \le n \le N-1$;

$$u(m,n)=\frac{1}{4MN}\{a(m,n-N)-b(m,n-N)e^{j\pi(n-N)/N} + c(m,n-N)e^{j\pi m/M} - d(m,n-N)e^{j\pi[m/M+(n-N)/N]} \tag{6}$$

if $0 \le m \le M-1$, and $N \le n \le 2N-1$;

$$u(m,n)=\frac{1}{4MN}\{a(m-M,n)+b(m-M,n)e^{j\pi n/N} - c(m-M,n)e^{j\pi(m-M)/M} - d(m-M,n)e^{j\pi[(m-M)/M+n/N]}\} \tag{7}$$

if $M \le m \le 2M-1$, and $0 \le n \le N-1$;

$$u(m,n)=\frac{1}{4MN}\{a(m-M,n-N) - b(m-M,n-N)e^{j\pi(n-N)/N} - c(m-M,n-N)e^{j\pi(m-M)/M} + d(m-M,n-N)e^{j\pi[(m-M)/M+(n-N)/N]}\} \tag{8}$$

if $M \le m \le 2M-1$, and $N \le n \le 2N-1$, where $u(m,n)$ is the image in spatial domain with the size of $2M \times 2N$.

The synthesized image can be divided into four quadrants: $\{Q_1(m,n): 0 \le m \le M-1, 0 \le n \le N-1\}$, $\{Q_2(m,n): 0 \le m \le M-1, N \le n \le 2N-1\}$, $\{Q_3(m,n): M \le m \le 2M-1, 0 \le n \le N-1\}$, and $\{Q_4(m,n): M \le m \le 2M-1, N \le n \le 2N-1\}$

From (5), we find that, due to signs in the combination of image pixel values, more power of the synthesized image is concentrated in the first quadrant Q_1. When coding the synthesized image in Q_1, we exploit not only the autocorrelation of one image but also the cross correlation between the four images, which may improve the coding efficiency. The subtraction operation in the other three quadrants has a high-pass filtering effect [7] which may facilitate motion estimation in the standard codec that follows. This phenomenon has been observed in our experiment (Experiment 2)

Three sets of coding experiments were carried out to evaluate the performance of the interlacing method: image coding and reconstruction, interframe coding, and data loss resilience.

Experiment 1: Multi-view image coding and reconstruction

The experimental results of the multi-view image coding are shown in Fig. 6, for the original four images and the corresponding synthesized image. The JPEG codec [9] was employed, and a quality factor q_JPEG [9] was used to control the quantization matrix.

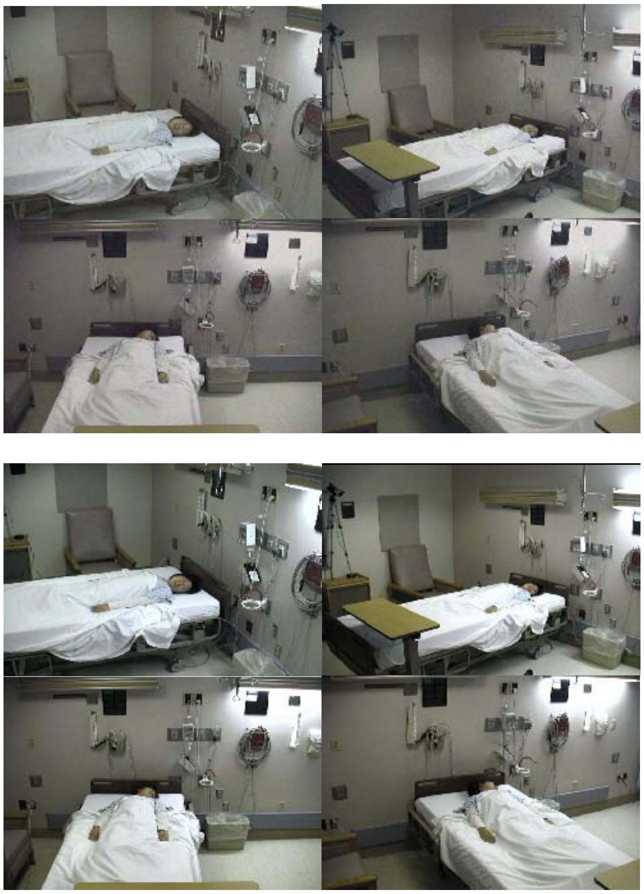

Fig.6. Images reconstruction. The upper four images are the reconstructed image from the synthesized one, where the JPEG codec is used with the q_JEPEG value of 65. The bottom four images are those coded separately, with the same q_JEPEG value.

As we can see that the reconstructed images from synthesization and the individually coded images hardly have any visual differences.

Experiment 2: Interframe coding

In this part, we discuss the interframe coding based on the DCT-interlacing technique.

In the general inter-coding techniques, only the reference frames, residuals after motion compensation, and motion vectors are coded with the video codec. The overall bits allocated for coding reference frames are usually much less than those for the residuals and motion vectors. Therefore, we only discuss the efficiency for the residuals and motion vectors. From (5) ~ (8) and Fig. 4, we find that there is a close relationship among the distributions of the pixels in the four quadrants, and the changes in one quadrant correspond to those in other quadrants. The same applies to the motion vectors associated with the changes. This may lend merit to video coding. In contrast, the residuals and motion vectors of the several video streams in the multi-view video set are considered to be irrelevant if they are coded independently.

In the video coding experiments, we compared the performances of the two coding methods. The resulting coding efficiencies of the presented method over those of the separate coding method are shown in Fig. 7, where both 4×4 and 8×8 blocks were used in the video coding procedure, and the efficiency improvements were tested under various quantization stepsizes, from 16 to 64. From the figure, we can see that the interlacing technique results in substantial improvements on the video coding efficiency.

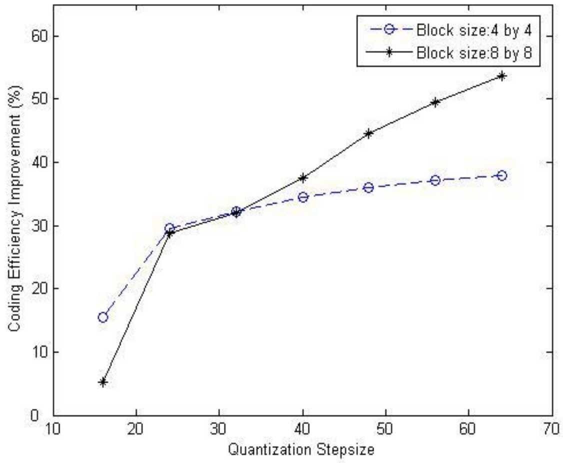

Fig. 7. Coding efficiency improvements of the interlacing method over the separate coding method. The experimental multi-view video was recoded at the University of Pittsburgh Medical Center (UPMC) including four viewpoints by camcorders Type JVC DV3000. The 480×720 video frames were in the format of RGB32. During the video coding process, the block sizes utilized were 4×4 and 8×8, and the quantization stepsizes were from 16 to 64.

Experiment 3: Data loss resilience

The presented method may provide an error resilience function by averaging the risk on each original image. As seen in Fig. 8, the reconstructed images from an experimental simulation of data loss still maintain satisfactory quality, which suggests that the interlacing technique is robust to occasional data loss. The quality of the reconstructed images is relevant to which coefficients are lost. Based upon the analysis given in the previous section, we see that the signal energy is smaller in quadrants Q_2 , Q_3 , and Q_4 . Therefore, if data loss happens in these three quadrants, the distortion of the reconstructed images is much less significant than that in Q_1 . In our experiments, the cases of DCT coefficient block lost in various locations were simulated, and what Fig. 8. shows is a typical result, where the lost 16×16 block lies in the center of the DCT coefficient matrix of the synthesized image.

IV. Conclusion and Future Work

In this work, we have presented a synthesizing scheme for coding multi-view patient monitoring video. Mono-view video codecs can be directly applied to the synthesized video, therefore leading to a simplified implementation as compared to other multi-view coding schemes. In addition, synchronization of the multiple video streams is obtained directly, and the resulting coded video becomes resilient to certain losses of data. We have investigated a simple interlacing algorithm to synthesize four frames into one, and demonstrated the viability of the presented approach by a theoretical analysis and experimental results. In future research, we hope to accurately determine the mapping relation between the pixels in the four quadrants, and equalize the power in the four quadrants so that the influence of data loss in the first quadrant can be decreased.

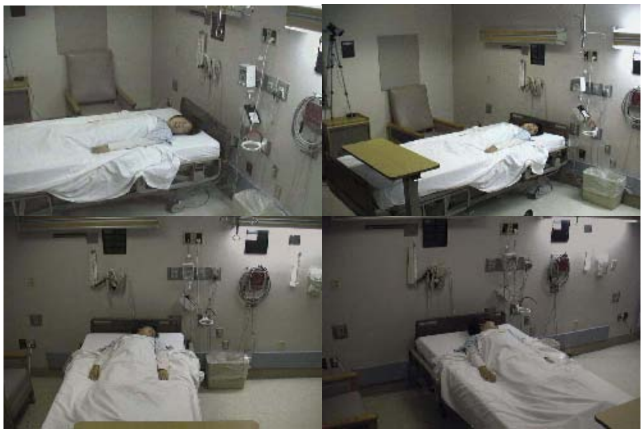

Fig.8. Reconstructed images with data loss. We simulated a case in which a 16×16 DCT coefficient block is lost during transmission.

References

[1] H. Kimata, "Preliminary results on multiple view coding for the sparse Ray Space representation (3DAV EE2.2.1)," *document M10327 MPEG Meeting,* Waikoloa, Hawaii, U.S.A., Dec. 2003.

[2] H. Kimata, "Preliminary results on inter-view coding for the dense Ray Space representation (3DAV EE2.2.1)," *document M10652 MPEG Meeting,* Munich, Germany, Mar. 2004.

[3] Xin Tong and Robert M. Grny, "Coding of multi-view images for immersive viewing," *Proceeding of IEEE Acoustics, Speech, and Signal Processing,*.vol.4pp Page(s):1879 – 1882, 5-9, Jun, 2000.

[4] N. Anantr asitichai, C. Nishan Canagarajah, and David R. Bull, "Lifting-Based Multi-Vi ew Image Coding," *Processing of Circuits and Systems*, vol. 3 pp:2092 - 2095, 23-26, May, 2005.

[5] Yumin Wang and Chuanjia Liang, *Information and Coding Theory*, Xi'an, China: Xidian Univeriisty, 1985.

[6] Fazlollah M. Reza, *An Introduction to Information Theory*, New York: Mcgraw-hill Bokk Company, Inc., 1961.

[7] Anil K. Jain, *Fundamental of Digital Image Processing*, NJ: Prentice Hall, Englewood Cliffs, 1989.

[8] Zhongde Wang, "Fast algorithms for the Discrete *VV* Transform and for the Discrete Fourier Transform," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. assp-32, NO. 4, Aug, 1984.

[9] Mohammed Ghanbari, *Video coding: an introduction to standard codecs,* IEE,1999.