Multiview Video Coding Using Video Epitomes

Narasak Boonthep[1], Kosin Chamnongthai[1], Werapon Chiracharit[1] and Yo-Sung Ho[2]

[1]Department of Electronic and Telecommunication Engineering

Faculty of Engineering

King Mongkut's University of Technology Thonburi

126 Pracha-utit Road, Bangmod, Tung-kru, Bangkok 10140, Thailand

Tel: +66-2470-9064, Fax: +66-2470-9070

[2]Gwangju Institute of Science and Technology (GIST)

1 Oryong-dong, Buk-gu, Gwangju, 500-712, Korea

Phone: (+82) 62-970 2258 Fax: (+82) 62-970 3164

E-mail: Narasak.b@hotmail.com, kosin.cha@kmutt.ac.th, werapon.chi@kmutt.ac.th, hoyo@gist.ac.kr

Abstract

This paper proposes a method for bit-rate-reduction in multiview video coding by using video epitomes. Multiview video contains a large amount of data to be stored or transmitted to the user. Because of limitation of transmission channels, it takes a long processing time. Epitomes are to compile small patches from input images into a condensed image model that represents many of the high-order statistical data in the training set. However, it still contains most constitutive elements needed for reconstruction of the image. The size of the epitomes is smaller than the size of the video. The video's essential textural, shape, and motion component are retained in the represented output. Epitomes provide a PSNR gain 2.5–3 dB or equivalent a half of the bit rate relative to simulcast.

Keywords: Multiview Video Coding (MVC), Multiview Video (MVV)

1. Introduction

Nowadays, the world's technologies are developing and changing all the time. Visual Communication Technology is one of the most important technology that have be improve in many fields such as 3D image processing, Holography, Multi View Video – MVV. Especially, Multiview video have been applied into many applications like Free Viewpoint Video (FVV), Free Viewpoint Television (FVT), Video-Teleconference and 3 Dimensional TV (3DTV).

In the past, users are limit to see only 2D so they can access to the image only one side but 3D technology allow them to access the image with freedom view so it seem more realistic video to users. Multiview video sequences are captured by several cameras at the same time but different locations. Because of the increased number of cameras, the multiview video contains a large amount of data. Since this system has serious limitations on data distribution applications, such as broadcasting, multimedia streaming services, and other commercial applications, we need to compress the multiview sequence efficiently without sacrificing its visual quality significantly. In this approach is aimed to improve compressing efficiency. The basis of our approach is using the latest standard coding H.264 combine with temporal and inter-view prediction. The result of using H.264 coding standard the output of video signal is better than previous coding standard (H.263) with the lower bit rate.

This paper is organized as follows: Section 2 introduces video epitomes. Section 3 describe about requirements, test data, and conditions. Section 4 present experimental results and section 5 concludes this paper.

2. Video Epitomes

In this paper we present novel simple appearance and shape models that called "epitomes". The epitome of an image is its miniature, condensed version containing the textural and shape components of the image. As opposed to templates or basis functions, the size of the epitome is considerably smaller than the size of the image or object it represents, but the epitome still contains most constitutive elements needed to reconstruct the image of Fig. 1 for an example. The epitome of an image consists of two parts: the epitome shape and the epitome appearance. A collection of images often shares an epitome, when images are a few consecutive frames from a video sequence. A

particular image in a collection is defined by its epitome and a smooth mapping from the epitome to the image pixels.

The epitome of a video sequence is a spatially and/or temporally compact representation of the video that retains the video's essential textural, shape, and motion components. More specifically, in this paper the epitome is built from the patches of various sizes from the input image. By choosing the size of the epitome and the sizes of the patches from which the epitome is constructed it is possible to strike a better balance between the quality of the fit and the ability of the model to generalize than what histogram and template approaches achieve. In fact, templates and histograms can be seen as special cases of the epitome. If the constitutive patches and the epitome are chosen to be of the same size as the input image, the epitomic representation becomes equivalent to a template. On the other hand, if patches consisting of a single pixel are used, and the epitome is very small, than the epitomic representation reduces to modeling the color map, and the probabilities of using a certain pixel in the epitome capture the color histogram. The Fig. 1 shows the manner in which a video epitome is learnt from a video. Viewing the input video as a 3D space-time volume, a large number of 3D training patches are drawn from the video. The learning algorithm is used to compile these patches into an "epitome". We derive the epitome learning algorithm by specifying a generative model, which explains how the input video can be generated from the epitome. By stacking the video frames together of video is considered to be a two-dimensional (2D) image with a time dimension that is a three-dimensional construct is obtained. Three-dimensional (3D) patches of varying spatial and temporal sizes from the video are used to learn the video epitome in an unsupervised manner. The video epitome itself is a 3D construct that can represent the video in both a spatially and temporally compact form. Under a probabilistic generative model, the video patches are considered to have come from a smaller video sequence - the video epitome.

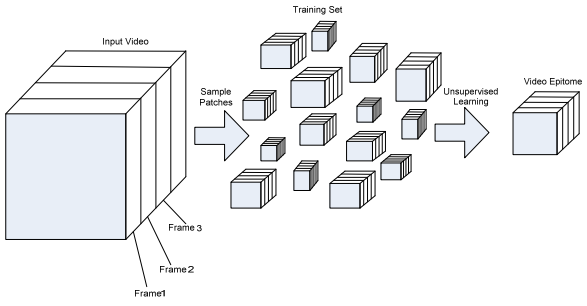

Fig. 1 Learning the epitome of a video.

3. Requirements and Conditions for MVC

In this paper we have been developed for MVC. Therefore, most of the requirements as well as test data and evaluation conditions are defined by the MVC project as presented in Section 3.1-3.3

3.1 Requirements

The central requirement for any video coding standard is high compression efficiency. In the specific case of MVC, this means a significant gain compared to independent compression of each view. Compression efficiency measures the tradeoff between cost (in terms of bit rate) and benefit (in terms of video quality). However, compression efficiency is not the only factor under consideration for a video coding standard. Some requirements of a video coding standard may even be contradictory such as compression efficiency and low delay in some cases. Then a good tradeoff has to be found. General requirements for video coding such as minimum resource consumption (memory, processing power), low delay, error robustness, or support of different pixel and color resolutions, are often applicable to all video coding standards.

For MVC, additionally view scalability is required. In this case a portion of the bit-stream can be accessed in order to output a limited number out of the original views. Also backward compatibility is required for MVC. This means that one bit-stream corresponding to one view that is extracted from the MVC bit-stream shall be conforming to H.264/AVC. Quality consistency among views is also addressed. It should be possible to adjust the encoding for instance to provide approximately constant quality over all views. Parallel processing is required to enable efficient encoder implementation and resource management. Camera parameters (extrinsic and intrinsic) should be transmitted with the bit-stream in order to support intermediate view interpolation at the decoder.

3.2 Test Data and Test Conditions

The proper selection of test data and test conditions is crucial for the development of a video coding standard. The test data set must be representative for the targeted area of applications, and therefore cover a wide range of different content properties. The Interactive Visual Media Group of Microsoft Research for providing the Breakdancers data set by eight different multiview test data sets have been used with 8 cameras views with 20 cm. spacing; 1D/arc. Picture resolutions are 1024 × 768 samples, and picture rates are 15 fps. The applications rather target high-quality TV-type video than limited

channel communication-type video. Therefore, smaller resolutions like CIF or QCIF are not considered. The MVC test data set covers a wide range of different content types, moving camera systems, and different complexities of motion and spatial detail. Fig. 2 shows some examples.

In order to perform comparative evaluations, the test conditions also have to be specified. For each test sequence, three bit rate s have been chosen corresponding to low but acceptable, medium and high quality, depending on the properties and content of the particular sequence. These fixed bit rate s allow a fair comparison of different approaches for MVC in objective and subjective tests as described below.

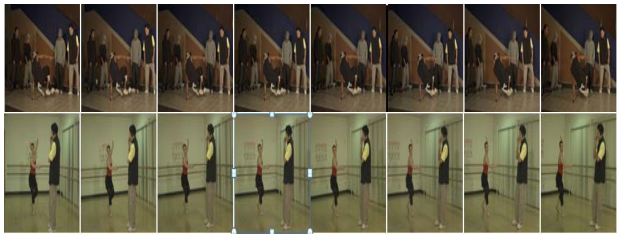

Fig.2 Examples of multiview video test data [7]

The main goal of MVC is to provide significantly increased compression efficiency compared to individually encoding all video signals. Therefore, encoding all views using H.264/AVC with the same test conditions was considered as the reference for coding performance comparison. The resulting decoded video signals (anchors) serve as reference for objective and subjective comparison. Encoding was done using typical settings.

3.3 Evaluation

Evaluation of video coding algorithms can be done using objective measures. The most common methods for measuring the quality of compressed images are mean squared error (MSE), between the original and decoded video samples and peak signal-to-noise-ratio (PSNR) of the signal, which is defined as:

$$PSNR_Y = 10.\log_{10}\left(\frac{255^2}{MSE}\right) \quad (1)$$

PSNR is an engineering term for the ratio between the maximum possible power of a signal and the power of corrupting noise that affects the fidelity of its representation. Because many signals have a very wide dynamic range, PSNR is usually expressed in terms of the logarithmic decibel scale. The PSNR is most commonly used as a measure of quality of reconstruction in image compression etc. It is most easily defined via the mean squared error (MSE) which for two images where one of the images is considered a noisy approximation of the other.

Typically, PSNR values are plotted over bit rate and allow then comparison of the compression efficiency of different algorithms. This can be done in the same way for MVC.

Some types of distortions that result in low PSNR values do not affect the human perception in the same way. One example is a shift of the picture by one sample side wards. Therefore, any video coding algorithm can finally only be judged in subjective evaluations. The formal MVC tests were conducted by MPEG using a Single Stimulus Impairment Scale (SSIS) test.

4. Experimental Results

Fig. 3 compares the rate-distortion performance of different multiview video coding method average over 100 frames (at 15 frames per second) and 8 view of Breakdancers sequence. The lower curve (Labeled “Simulcast”) represents independent coding of each view using version JMVM 8.0 of the H.264 reference software. The Upper curve represents “Epitomes” as described in section 2. Epitomes provide a PSNR gain of 2.5 – 3 dB or equivalently half the bit rate relative to simulcast.

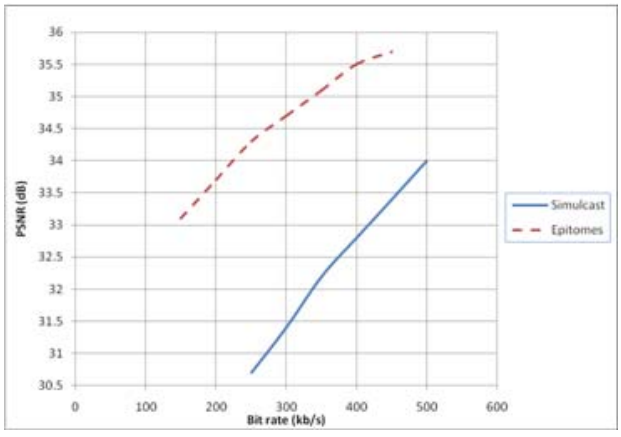

Fig.3 Result for Multiview video coding .

5. Conclusions

We have proposed a multiview video coding based on the H.264 which utilize both temporal/interview prediction by adaptive use

video epitomes. This method can compress the multiview sequence efficiently with representation of the multiview video that retains the video's essential textural, shape, and motion component. From experimental results, it was shown that PSNR gains of 2.5 – 3 dB or equivalently more than half the bit rate relative to simulcast.

References

[1] Vincent Cheung, Brendan J. Frey, Nebojsa Jojic, "Video epitomes", CVPR 2005. IEEE Computer Society Press, Los Alamitos, CA, June 2005.

[2] Philipp Merkle, Aljoscha Smolic´, Karsten Müller, and Thomas Wiegand . "Efficient Prediction Structures for Multiview Video Coding", IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, VOL. 17, NO. 11, NOVEMBER 2007.

[3] K. Yamamoto, M. Kitahara, H. Kimata, T. Yendo, T. Fujii, M. Tanimoto, S. Shimizu, K. Kamikura, and Y. Yashima. "Multiview Video Coding Using View Interpolation and Color Correction", IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, VOL. 17, NO. 11, NOVEMBER 2007.

[4] Philipp Merkle, Aljoscha Smolic, Karsten Müller, and Thomas Wiegand, "MULTI-VIEW VIDEO PLUS DEPTH REPRESENTATION AND CODING"

[5] N. Jojic, B. J. Frey, and A. Kannan. Epitomic analysis of appearance and shape. In Proc. IEEE Intern. Conf. on Computer Vision, 2003.

[6] ISO/IEC MPEG & ITU-T VCEG, "Joint Multiview Video Model (JMVM) 3.0," JVT-V207, Jan. 2007

[7] ftp://ftp.research.microsoft.com/users/sbkang/