PERFORMANCE MODELING AND EVALUATION OF PREDICTION STRUCTURES IN MULTI-VIEW VIDEO CODING

Chao Chen[1], *Yebin Liu*[2], *Qionghai Dai*[1,2], *IEEE Senior Member, Xiaodong Liu*[1]

[1]Graduate School at Shenzhen, Tsinghua Universty
[2]Broadband Networks and Digital Media Lab, Department of Automation, Tsinghua University
Beijing 100084, P.R.China

ABSTRACT

So far, lots of prediction structures have been proposed to exploit the temporal and inter-view redundancy of multi-view video. However, none of them can adapt themselves to different video contents. In this sense, it is desirable to adaptively select different prediction structure for different multi-view video contents. In this paper, a solution is proposed to evaluate the performance of prediction structures. With this solution, the performance of any kind of prediction structure can be evaluated before actually applying them. The solution can also be used for the tradeoff optimization between compression efficiency and other functionalities such as random-access ability and view scalability.

1. INTRODUCTION

With the advancement in computer vision and graphics, 3D video and free-viewpoint video have gained more and more interests. For all these applications, multi-view video coding (MVC) is a key technology. Different from the 2D video, multi-view video is generated by many cameras simultaneously capturing the same scene from different directions. Because of the increased number of cameras, the multi-view video data is extremely large. It is desirable to compress the multi-view sequence efficiently to make it feasible for storage or real-time transmission [1].

Multi-view video contains not only temporal redundancy but also large degree of inter-view redundancy. So far, lots of prediction structures for MVC have been proposed [2]. Among them, the most widely known ones are "GoGOP", "Sequential View Prediction (SVP)" and "Checkerboard Decomposition (CD)". In addition, "Hierarchical B Pictures" proposed by HHI has been selected by MPEG as reference for its excellent performance [3].

However, because of the non-stationary property of video stream, we cannot expect a prediction structure to be universally effective for any scene and any time. Furthermore, like other types of video media, the contents of multi-view video vary slightly during a short time. If we could estimate the performance of all kinds of prediction structures according to the short-time character of a multi-view sequence, it is possible to adaptively select the optimal prediction structure for compression.

No matter in what kind of prediction structure, all the frames, except for I frames, are encoded using one or several reconstructed frames as references. In this sense, the efficiency of motion-compensated prediction determines the performance of the prediction structure. Bernd Girod has made a qualitative analysis of this issue [4] [5]. The theoretical performance bounds were derived based on rate-distortion theory.

In this paper, a solution is presented for the quantitative evaluation of prediction structures. As for single-hypothesis prediction, a logarithmic model was proposed to estimate the prediction efficiency. As for the more complicated case of multi-hypothesis prediction, a normalized exponential factor model is employed. The parameters of the models are adjusted according to the short-time character of the multi-view video. With this solution, we can compare the estimated performance of different prediction structures and select the best one. The models we presented can also be used in the tradeoff analysis between compression efficiency and other functionalities such as random-access ability and view scalability.

The rest of the paper is organized as follows. Section 2 describes the framework of our solution. In section 3, as a key point of our solution, the normalized exponential factor model is detailed. Section 4 gives the experimental results. Finally, our conclusions are presented in section 5.

2. FRAMEWORK OF THE SOLUTION

A frame in a GoP is either encoded with intra mode(I frames) or inter mode(P and B frames). Accordingly, the estimation for the performance of prediction structures can be divided into two sub-problems: The estimation of the output bits number of I frames and that of P or B frames. Our solutions for them both are described below.

This work is supported by the Distinguished Young Scholars of NSFC (No.60525111) and the key project of NSFC (No.60432030)

2.1. Estimation of the output bit number of I frames

We use the average number of output bits of a series of I frames as an estimation of the output bit number of I frames within a GoP. 20 consecutive frames in each view of three multi-view video sequences are encoded with H.264 codec and the output bit number of each frame is shown in Fig.1.

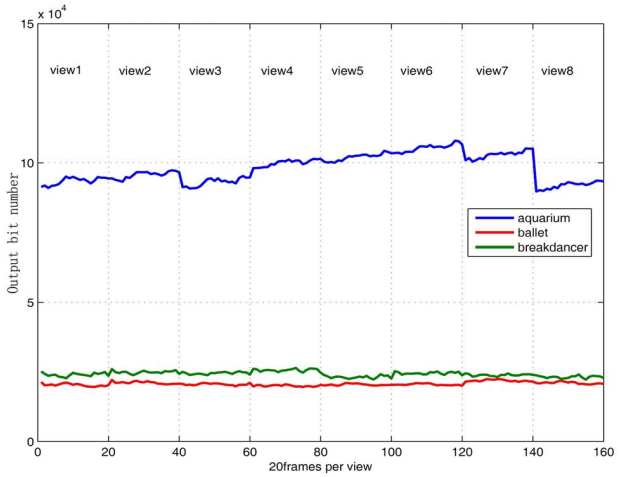

Fig. 1. The output bit number of 20 consecutive I frames in each view.

As shown in the fig.1, the output bits of I frames vary slightly within a GoP which guarantees the rationality of the estimation.The stability of the output bit number is caused by the similarity of the frames because all the cameras are capturing the same scene simultaneously and the scene varies slightly in a short time.

2.2. Estimation of the output bit number of P or B frames

P or B frames are encoded with inter mode. Let B_{inter} denote the output bit number of a frame encoded with inter mode and let B_{intra} denote the output bit number when it was encoded with intra mode. We define the prediction efficiency (PE) by B_{inter}/B_{intra}. We have

$$B_{inter} = B_{intra}\frac{B_{inter}}{B_{intra}} = B_{intra} \cdot PE \tag{1}$$

In the equation above, B_{inter} is determined by B_{intra} and PE. B_{intra} reflects the redundancy within the encoded frame and PE reflects the correlation between the encoded frame and its reference frames. B_{intra} can be estimated by the method described in 2.1. If we could estimate PE, B_{inter} can be calculated with equation (1).

Single-Hypothesis Prediction As for the estimation of PE, the simplest case is that only one reference frame is employed. Let $PE_{temporal}$ denote the efficiency of temporal prediction and let $PE_{spatial}$ denote the efficiency of inter-view prediction. Define the prediction distance D by the temporal or spatial interval between the encoded frame and the reference frame. Assuming that PE is only relative to D, we use an logarithmic function $PE(D) = m + nlog_s D$ to model the relationship between PE and D where m, n and s are model parameters. Let K denote the width or length of the GoP. We first use H.264 codec to get the real $PE(1)$, $PE(2)$, $PE(3)$, $PE(K)$ and then fit them to the model to get m, n and s. As shown in fig.2, the model works well for all the tested sequences.

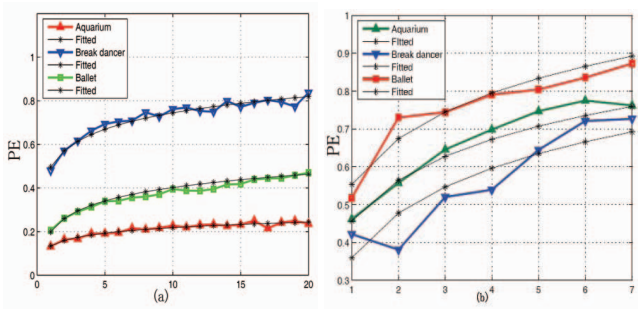

Fig. 2. (a) Real $PE_{temporal}$ and fitted $PE_{temporal}$ vs. temporal interval; (b) real $PE_{spatial}$ and fitted $PE_{spatial}$ vs. inter-view prediction interval.

Multi-Hypothesis Prediction As for the multi-hypothesis prediction in which several reference frames are employed, the joint PE are determined by the property of all reference frames. We use $F_1, F_2, \cdots, F_N$ to denote N reference frames. The PE when the ith reference frame was solely referred is denoted by PE_i. With the logarithmic model described above, we can easily estimate all the $PE_i s$. Actually, PE_i reflects the property of the ith reference frame and every reference frame contributes to the performance of multi-hypothesis prediction. In this sense, the joint PE can be modeled with a function of $PE_1, PE_2, \cdots, PE_N$.

We use a factor-normalized exponential function to model the relationship between joint PE and PE_i. As a key point of the solution ,this model is detailed in the following section.

3. NORMALIZED EXPONENTIAL FACTOR MODEL

Sort the PE_is as $PE_1 > PE_2 > \cdots > PE_N$. The normalized exponential factor model is

$$PE_J = A \cdot \prod_{i=1}^{N} PE_i^{\lambda_i} \tag{2}$$

where PE_J is the joint prediction efficiency, $\lambda_i$and A are model parameters. The choice of $\lambda_i$satisfies

$$\sum_{i=1}^{N} \lambda_i = 1 \tag{3}$$

3.1. Prediction analysis & algorithm modeling

There are three key principles of the model.

1. Among the reference frames, some of them are strongly correlative to the encoded frame while the others are not. Hence the contributions of different reference frames to PE_J are different. Furthermore, as discussed in section 2, the PE_i of the ith frame reflects its contribution to PE_J. In this sense, we assign different power λ_i to different PE_i and λ_i satisfies(3).

2. The prediction efficiency of the reference frames is different from GoP to GoP. To model this temporal difference, we introduce a amplitude parameter A. What's more, the number of employed reference frames will also affect PE_J. With the parameter A, such effect is taken into account as well. Hence the model is expressed as (2).

3. Define the hierarchy of multi-hypothesis prediction by the number of reference frames used for prediction. In the case of bi-hypothesis prediction, the model is$PE_J = A \cdot PE_1^{\lambda} PE_2^{1-\lambda}$. When another reference frame is employed as reference, we simply use the bi-hypothesis model to calculate the prediction efficiency $PE_J^{'}$ as

$$PE_J^{'} = \begin{cases} A \cdot PE_J^{\lambda} PE_3^{1-\lambda} & : \quad if PE_J \geq PE_3 \\ A \cdot PE_3^{\lambda} PE_J^{1-\lambda} & : \quad if PE_J < PE_3 \end{cases} \tag{4}$$

Note that

$$\begin{aligned} PE_J^{'} &= \begin{cases} A \cdot [A \cdot PE_1^{\lambda} PE_2^{1-\lambda}]^{\lambda} PE_3^{1-\lambda} \\ A \cdot PE_3^{\lambda} [A \cdot PE_1^{\lambda} PE_2^{1-\lambda}]^{1-\lambda} \end{cases} \\ &= \begin{cases} A^{1+\lambda} PE_1^{\lambda^2} PE_2^{\lambda-\lambda^2} PE_3^{1-\lambda} \\ A^{2-\lambda} PE_1^{\lambda-\lambda^2} PE_2^{1-2\lambda+\lambda^2} PE_3^{\lambda} \end{cases} \end{aligned}$$

in both cases, the assigned powers still satisfy (3). In the same sense, the PE_J of higher hierarchy can be calculated in the same way. This attribute of normalized exponential factor model makes it feasible to calculatePE_J no matter how many frames are employed for reference.

3.2. Approach to get the model parameters

As demonstrated above, all the model parameters of higher hierarchy can be derived from the model of bi-hypothesis prediction. As for bi-hypothesis prediction, all the possible arrangement of the encoded frame and the reference frames is classified into 5 modes as shown in fig.3. Fitting A and λ for all the 5 modes respectively, the joint PE_J of higher hierarchy can be calculated with the following steps.

1. Divide the reference frames into 4 groups according to their relative position to the encoded frame as shown in fig4.

2. Using equation (4), calculate the joint prediction efficiency PE_1 for the frames in Group1 with the model parameters of the prediction mode 'co-view'.

3. In the same way, calculate joint PE_2 PE_3 and PE_4 for Group2,3 and 4.

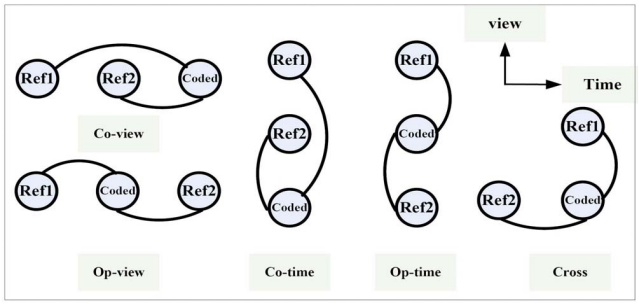

Fig. 3. All arrangements of the encoded and the reference frames in the case of bi-hypothesis prediction.

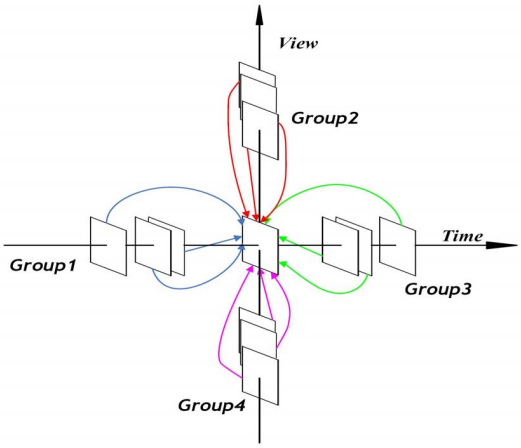

Fig. 4. Classification of the reference frames

4. Calculate the joint PE_{time} using PE_1 and PE_3 with the parameters of mode 'op-view'

5. Calculate the joint PE_{view} using PE_2 and PE_4 with the parameters of mode 'op-time'

6. Calculate PE_J using PE_{time} and PE_{view} with the parameters of mode 'cross'.

With this approach, we calculated the joint PE_J when 4 reference frames are employed. Experimental results in all kinds of prediction modes are shown in figure 5. The prediction modes are named according to the number of frames in each group. For example, if there is 3 frames in Group2 and 1 frame in Group4, the mode is named as 0301. All these results are tested on the first 16x8 GoP of the sequence 'race1'.

4. EXPERIMENTAL RESULT

Using our proposed model, we estimated the performance of Sequential View Prediction (SVP) and Checkerboard Decomposition (CD) when they are applied to the first 8 views of the sequence 'aquarium'. Fig.6 shows the experimental result. All the experiments were conducted using h.264/jm86.The option of Rate Control and Weighted Prediction is set 'off'

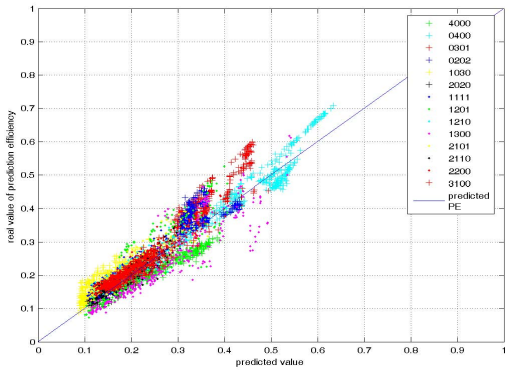

Fig. 5. Predicted PE_J vs. real PE_J when 4 reference frames are employed

and QP is 28.

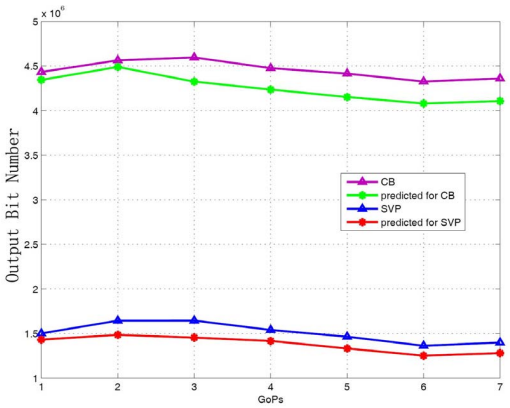

Fig. 6. Predicted performance of SVP and CD on sequence 'Aquarium'

The size of the GoPs in the experiment is 8x8 and the model parameters are fitted for every GoP respectively. As shown in fig.6, for all the GoPs, our model works well.

The predicted performances of SVP are averagely 8.52% lower than the real value SVP while the predicted performance of CD are 4.62% lower. This deviation is caused by the nature of SVP and CD. In both of the prediction structures, all of the prediction distances are 1. However, the model parameters are fitted for all possible distances. Hence, the prediction is not an unbiased estimation for those prediction structures in which the prediction distances are concentrated on several values. Because the model parameters fitted for each GoP vary slightly, we can apply the same parameter to more GoPs and thus sharply reduce the computational load for fitting parameters. We fitted the model parameters for the first GoP of the sequence 'ballet' and apply them to all the tested GoPs. The experimental result is shown in fig.7.

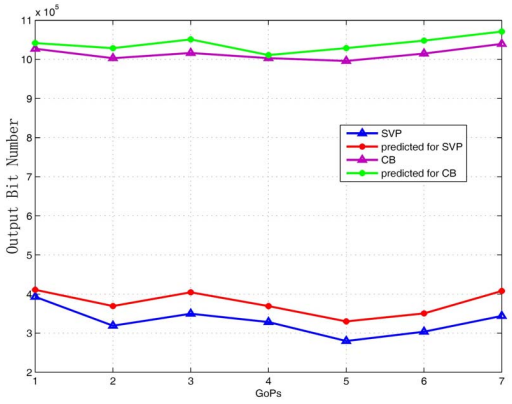

Fig. 7. Predicted performance of SVP and CD on sequence 'Ballet'

5. CONCLUSIONS

Our proposed solution can give an estimation for the performance of prediction structures before actually applying them. As a part of our solution, a normalized exponential factor model was proposed. The model parameters of higher hierarchy can be deduced from the parameters of lower hierarchy and thus it is possible to calculate the prediction efficiency no matter how many reference frames are employed.

With our proposed solution and model, we could select the most efficient prediction structure for the encoded GoPs. In the tradeoff analysis between different functionalities, the proposed method can measure the compression efficiency of complicated prediction structures.

6. REFERENCES

[1] A.Smolic, and P. Kauff, "Interactive 3D Video Representation and Coding Technologies," *Proceedings of IEEE*, vol. 93, no. 1, Jan. 2005.

[2] ISO/IEC JTC1/SC29/WG11 N6909, "Survey of algorithms used for Multi-view Video Coding (MVC)," Hong Kong, China, 2005.

[3] K. Mueller, P. Merkle, etal., "Multi-View Video Coding based on H.264/MPEG4-AVC Using Hierarchical B Pictures," *PCS2006*, Beijing, China, April 2006.

[4] Bernd Girod, "The Efficiency of Motion-Compensating Prediction for Hybrid Coding of Video Sequence," *IEEE Journal on Selected Areas in Communication*, Vol. Sac-5. No. 7. August 1987.

[5] Bernd Girod, "Efficiency Analysis of Multihypothesis Motion-Compensated Prediction for video coding," *IEEE Trans. Image Processing*, Vol. 9. No. 2. February 2000.