LOW-DELAY RANDOM VIEW ACCESS IN MULTI-VIEW CODING USING A BIT-RATE ADAPTIVE DOWNSAMPLING APPROACH

Erhan Ekmekcioglu, Stewart T. Worrall, Ahmet M. Kondoz, Member, IEEE

Centre for Communication Systems Research, University of Surrey, GU2 7XH Guildford, Surrey, United Kingdom
{E.Ekmekcioglu, S.Worrall, A.Kondoz}@surrey.ac.uk

ABSTRACT

In this paper, a new multi-view coding (MVC) scheme is proposed and evaluated. The scheme offers improved low-delay view random access capability and at the same time comparable compression performance with respect to the reference multi-view coding scheme currently used. The proposed scheme uses the concept of multiple-resolution view coding, exploiting the trade-off between quantization distortion and downsampling distortion at changing bit-rates, which in turn provides improved coding efficiency. Bi-predictive (B) coded views, used in the conventional MVC method, are replaced with predictive coded downscaled views, reducing the view dependency in a multi-view set and hence reducing the random view access delay, but preserving the compression performance at the same time. Results show that the proposed method reduces the view random access delay in a MVC system significantly, but has a similar objective and subjective performance with the conventional MVC method.

Index Terms— Multi-view video coding, view random access, multi resolution video coding

1. INTRODUCTION

The issue of multi-view video coding has grown in significance following recent advances in 3D capture and display technologies, making applications like 3D TV [1] and Free viewpoint TV (FTV) [2] possible, where the user is given the chance to navigate between viewing angles freely. Due to increased visual content, it is essential to use advanced techniques for compressing multi-view data, where temporal as well as inter-view correlations are exploited.

A large amount of the research carried out so far on MVC is aimed at improving the compression performance [3-4]. Though it is important, there are other requirements of MVC, such as low-delay random view access, enabling fast view switching in free-view video applications. In [5], a multi-view light field coding technique is proposed, based on Wyner-Ziv coding. It provides enhanced random access capability, although the performance of the Wyner-Ziv coding used is poor compared to H.264/AVC [6]. Similarly, in [7], an MVC approach based on distributed source coding is proposed to tackle the free-view switching problem, where the results are compared against conventional closed loop and intra coding. In [8], an MVC approach based on 3D point clouds is proposed providing easy free-view access. However, it uses a different picture representation which is not suitable for use with block based video coders. In [9], SP/SI frame coding, interleaved coding and multiple representations coding are tested separately. All of the approaches perform worse than the conventional MVC approach which uses hierarchical B prediction [4] in both view and time domain, although they provide reduced delay view random access capability.

In this paper, a different inter-view prediction structure is proposed, which aims to replace bi-predictive (B) coded views with downscaled (using bit-rate adaptive downsampling ratios) and predictive (P) coded views. The definitions of B and P coding are as stated in the standard [6]. The goal is to omit B type inter-view predictions, which inherently introduces view hierarchy to the system and increases random view access delay. The disadvantage is that B coding, which improves the coding efficiency significantly, is avoided. However, the proposed scheme preserves the coding performance by using downsampled and P coded views, reducing the random view access delay remarkably at the same time. A mathematical model is constructed to relate the coding performances of different coding types used within the proposed scheme to each other, which enables us to estimate the relative coding efficiencies of different inter-view prediction structures.

Section 2 describes the bit-rate adaptive downsampling scheme. The proposed inter-view prediction structures are given in section 3. The performance estimation model is derived in section 4, where derived performance graphs are compared to real experimental results. Section 5 provides the subjective evaluation results and section 6 concludes the paper.

This work was developed within VISNET II, a European Network of Excellence (http://www.visnetnoe.org), funded under the European Commission IST FP6 programme.
2. BIT-RATE ADAPTIVE DOWNSAMPLING

The idea behind downsampling a view prior to coding and upsampling the reconstructed samples then after is based on the trade-off between two types of distortion: distortion due to quantization and distortion due to downsampling. Considering a fixed bit-rate budget, increasing the downsampling ratio means that less coarse quantization needs to be used. Thus, more information is lost through downsampling, but less is lost through coarse quantization. Finding the optimum trade-off between the two distortion sources should lead to improved compression efficiency. To observe this, views were downsampled with different downscaling ratios prior to encoding, ranging from 0.3 to 0.9 (same ratio for each dimension of the video, leaving the aspect ratio unchanged). These ratios were tested over a broad range of bit-rates. The results indicated that the optimum trade-off between the two distortion types varies with the target bit rate. Fig. 1 shows the performance curves of downscaled coding with some downscaling ratios and full resolution coding for a particular view of Breakdancer test sequence [10].

The best performance characteristics at medium and low bit-rates, where the quantization distortion is more effective, is achieved with 0.6 scaling ratio (mid-range), whereas at much higher bit-rates, where the effect of the distortion due to downsampling becomes more destructive, larger scaling ratios (0.8-0.9) are suitable, introducing less downsampling distortion. Very low ratios, such as 0.3, are only useful at very low bit-rates, where the reconstruction quality is already insufficient to be considered (less than 32 dB). Results do not change over different data-sets.

In the rest, for simplicity, two predefined downscaling ratios are used, 0.6 for bit-rates less than 300 kbps, 0.8 for bit-rates over 300 kbps targeting VGA sequences (640x480) at 25 fps. Accordingly, up to 20% save in bit-rate is achieved for individual views at certain reconstruction qualities.

3. INTER-VIEW PREDICTION STRUCTURE

The random view access corresponds to accessing any frame in a GOP (group of pictures) of any view with minimal decoding of other views [9]. In Fig. 2 (a), the reference inter-view prediction structure of the current MVC draft is shown (for 8 views and 16 views cases) at anchor frame positions. The random view access cost, defined as the maximum number of frames that must be decoded to reach the desired view, is 8 and 16 for the 8-view and the 16-view cases respectively. The disadvantage is that as the number of cameras increases, the cost increases with the same rate. Furthermore, in some streaming applications only relevant views may be sent to the user, to save bandwidth. With such dependency structure, more views would have to be streamed and hence, the bit-rate would increase. In Fig. 2 (b), the proposed low delay view random access model with downsampling P coding (LDVRA+DP) is given. The Group of Views (GOV) concept is used which is suitable for free viewpoint video (separated by dashed lines) and in each GOV, one view, called a base view, is coded at full spatial resolution, while other views, called enhancement views, are downsampled using the idea described in section 2 and are predictive (P) coded. None of the views are B coded, so that no extra layers of view dependency are present. Every enhancement view is dependent on its associated base view and every base view depends on the same base view, whose anchor frames are intra coded. This obviously necessitates perceptual testing of the downsampled views (later upsampled again) to observe the potential effect of low-pass filtering, which will be discussed in section 5.
4. PERFORMANCE ESTIMATION MODEL AND EXPERIMENTAL RESULTS

In this work it is assumed that the coding performances at anchor frame positions of both techniques reflect the overall multi-view coding performances of the respective techniques. The reason is that most of the coding gain in MVC, compared to simulcast, is achieved at anchor frame positions where there are no means of temporal prediction. Therefore, only the coding efficiency at anchor frame positions is going to be evaluated. In both techniques there is one intra coded (I) view (all GOPs begin with an intra coded frame). Besides, both prediction structures contain a certain number of P views, B views and/or DP views. Another assumption is that for each view, I coding at anchor frame positions at the same time instant would generate similar bit-rate for the same output quality. Same is valid for P coding.

The efficiency metrics of P, B and DP coding are defined as $\alpha_p$, $\alpha_b$, $\alpha_{dp}$ respectively. $\alpha_p$ is set to 1. Accordingly, $\alpha_b$ and $\alpha_{dp}$ change between 0 and 1. The values of $\alpha_b$, $\alpha_{dp}$ are determined experimentally, and their values for different views are found to be consistent. $\alpha_b$ and $\alpha_{dp}$ are determined at different bit-rates by taking the ratios of output bit-rates for B and DP coded views respectively to the output bit-rate of P coded views. A lower efficiency index means higher coding efficiency. It is important to select the bit-rates of B coded and DP coded views such that they generate reconstructed images at similar objective quality, to ensure that the efficiency comparison between them is fair. Let the total number of cameras in a multi-view system be equal to $2n$. The per-view efficiency index can be calculated as:

<table>
<thead>
<tr>
<th>Reference MVC</th>
<th>$\Phi_1 = \frac{n + (n - 1)\alpha_b}{2n}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>LDVRA+DP (GOV=3)</td>
<td>$\Phi_1 = \frac{2n}{2n} - 1 \cdot \frac{4^n}{5} \cdot \alpha_{dp}$</td>
</tr>
<tr>
<td>LDVRA+DP (GOV=5)</td>
<td>$\Phi_1 = \frac{2n}{2n} - 1 \cdot \left(\frac{8^n}{5} \cdot \alpha_{dp}\right)$</td>
</tr>
</tbody>
</table>

In order to make sure that the proposed low delay random access coding scenario performs at least as good as the reference MVC method $\Phi_1$, should be greater than zero. Experimental values for $\alpha_b$, $\alpha_{dp}$ guarantee that this condition is satisfied for the test videos used, at most bit-rates. Similarly, in order to make sure that the proposed scheme performs better with larger GOV sizes, define $\Phi_2$ as the

Fig. 3 (a) and (b) show per-view efficiency versus PSNR graphs drawn for 16-view Rena and 8-view Breakdancer sequences, for experimentally determined values of $\alpha_b$ and $\alpha_{dp}$. At several frame reconstruction quality levels, the output bit-rates of B and DP coded views are divided by the output bit-rate of P coded view to get $\alpha_b$ and $\alpha_{dp}$. These graphs represent rough estimations of real performance graphs although the horizontal axis doesn’t directly stand for the bit-rate. However, since the per-view efficiency indices for each method are calculated with $\alpha$ values obtained at similar output qualities, they are proportional to the actual bit-rate. Actual coding results with JMVM [11] are given in Fig. 4 (a) and (b). Common coding configurations for each view are shown on Table 1. LDRA performs worse than the reference MVC method, since it does not benefit from the coding gains of B coding or DP coding. The proposed LDVRA+DP coding scheme, in which no downsampling is utilized for P coded views, LDVRA+DP performs slightly better than the reference coding technique, in which no downsampling is utilized for P coded views. LDRA performs worse than the reference MVC method, since it does not benefit from the coding gains of B coding or DP coding. The proposed LDVRA+DP coding tends to perform slightly better than the reference coding technique at especially lower bit-rates. The real relative efficiencies of the proposed techniques with respect to the reference coding technique are reflected in the estimated graphs. In order to compare the relative efficiencies of both techniques $\Phi_1$ is defined as the difference between the per-view efficiency indices of the reference and the proposed techniques. Then,

$$\Phi_1 = \left[\frac{n - 2n}{3} - 1 \cdot \frac{4^n}{5} \cdot \alpha_{dp}\right]$$

In order to make sure that the proposed low delay random access coding scenario performs at least as good as the reference MVC method $\Phi_1$, should be greater than zero. Experimental values for $\alpha_b$, $\alpha_{dp}$ guarantee that this condition is satisfied for the test videos used, at most bit-rates. Similarly, in order to make sure that the proposed scheme performs better with larger GOV sizes, define $\Phi_2$ as the
difference between the per-view efficiency indices of the two proposed techniques. Then, the following is obtained:

\[
\Phi_2 = \left[ \frac{2n}{3} - 1 + \frac{4n}{3} + 1 \right] \left[ \frac{2n}{5} - 1 - \frac{8n}{5} + 1 \right] \alpha_{op} \\
\Rightarrow \Phi_2 = \frac{2}{15} (1 - \alpha_{op}) \geq 0 \quad (3)
\]

Since \( \alpha_{op} \) is absolutely below 1, it is certain that the condition in (3) is satisfied. It is verified in both the estimated and the real coding results.

5. SUBJECTIVE EVALUATION

The perceptual quality of the proposed low delay random access scheme with DP coding is compared with the reference technique containing B coded views, using the stimulus comparison - adjectival categorical judgement method described in recommendation [12]. Rena (Nagoya University), Breakdancer and Ballet (Microsoft Research) test sequences are used for evaluations. A differential mean score opinion is calculated at two different bit-rates and plotted on a differential scale, where 0 corresponds to no perceptual difference between the two methods and negative values indicate that the proposed method performs better. 16 subjects are used in the evaluations. Fig. 5 shows the results. Since the Rena sequence is a blurry sequence originally, the downsampled distortion is not visually suitable. Therefore, there is no visual change for Rena sequence. On the other hand, for the other sequences tested, it can be observed that at high bit-rates, the perceptual qualities of both methods do not differ, indicating that the downsampled distortion (blurriness) is not a significant issue. At lower bit-rates, quantization distortion (blockiness) is more visible than the downsampled distortion and hence the proposed method generates visually more satisfactory results.

6. CONCLUSION

It is observed that the random view access performance of multi-view coding systems can be improved significantly with respect to the conventional MVC method, without any loss of coding performance and perceptual quality. The reason is that the performance of efficient B coding, present in the conventional MVC method, can be achieved by the downsampled P (DP) coding. The proposed inter-view dependency structure is more suitable for fast switching free-view systems, due to the utilization of the concept of group of views (GOV). Furthermore, assigning larger GOV sizes can further increase the compression performance, without affecting the overall random view access delay. The proposed approach brings slight increase in the complexity due to the addition of up-conversion and down-conversion blocks, but this is balanced with the reduction in the processing load for the downsampled videos.

One limitation of this technique is with highly textured video sequences, where the inherent low-pass filtering effect might significantly degrade the subjective quality. This can be overcome by transmitting extra residuals for the blocks in the vicinity of object edges, to improve the visual quality, which is of future research interest.

7. REFERENCES