TECHNIQUES FOR MULTIVIEW VIDEO CODING

Anthony Vetro, Emin Martinian, Jun Xin, Alexander Behrens, Huifang Sun

Mitsubishi Electric Research Labs
Cambridge, MA USA 02139

ABSTRACT

This paper presents a brief overview of some techniques used in our current system for multiview video coding. Specifically, we describe the prediction of select camera views from synthesized views for improved coding efficiency. We also describe a unique form of random access for multiview coding systems. Some system-level considerations of our system, such as reference picture management and the importance of maintaining compatibility with single-view video codecs, such as H.264/AVC, are also discussed.

1. INTRODUCTION

Conventional video coding systems are designed to encode the video from a single video source. With the emergence of new interactive applications and display technology that require multiple views of video, there is a need to extend the conventional coding schemes to exploit the redundancy between camera views and provide higher coding efficiency than independent coding of each view [1].

Disparity compensated prediction is one of the more common techniques to exploit inter-view correlations. In this scheme, certain views are encoded independently using a standard video encoder; these independently encoded views are reference views. The remaining views are then encoded using temporal prediction and inter-view spatial predictions based on reconstructed reference views. The prediction can be determined adaptively on a per block basis, e.g., as in [2].

In this paper, we describe the prediction of select views based on view synthesis methods. The potential advantage of this form of prediction over traditional spatial prediction is that the synthesized view attempts to reconstruct the view from neighboring views using camera parameters and depth information. Such an approximation of the view to be coded could have significant advantages in fast moving areas of a scene and for objects that undergo non-translational motion by providing a better predictor of the frame.

This paper also describes unique form of random access for multiview coding systems in which temporal prediction is not allowed, but prediction along the spatial dimension is allowed for improved coding efficiency. Finally, a reference picture management scheme is presented. Compatibility with existing single-view video coding schemes is achieved by the proposed technique.

2. PREDICTION FROM SYNTHESIZED VIEWS

This section describes the use of synthesized frames for prediction in a multiview coding system. Given the pixel values of frames of one or more reference views and the depth values of points in the scene, the pixels in the frames of the target view can be synthesized from the corresponding pixel values in the frames of the reference views. View synthesis (or interpolation) is commonly used in computer graphics for rendering images with multiple views [3]; such methods typically require extrinsic and intrinsic parameters of the cameras. In our system, we generate synthesized frames for prediction.

One difficulty with this approach is that the depth values of the scene are unknown. Therefore, we must estimate the depth values. Known graphics techniques that are based on feature correspondences in the multiple views may be used. Alternatively, for each target' view, we can generate multiple view synthesized frames, each corresponding to a candidate depth value. For each block in the current frame, the best matching block in the set of view synthesized frames can be determined. The view synthesis frame from which this best match is found indicates the depth value of the block in the current frame. This process is repeated for all block in the frame.

The residual between the current block and a view synthesis block is encoded. Besides signaling this coding mode at the block level, some additional side information including the depth value and an optional displacement vector must also be coded. Experiments have shown that the depth maps can be efficiently coded using standard video coding tools and will not incur much overhead. The optional disparity vectors are small and intended to compensate for any misalignments between the block in the current frame and the best matching block in the view interpolation frame to be compensated, so should also not incur much coding overhead.

3. MULTIVIEW RANDOM ACCESS

In order to provide random access to any point in a typical video sequence, intra-frames (I-frames) are usually spaced throughout the sequence at regular intervals. This enables the decoder access to any frame in the decoded sequence, although at a decreased compression efficiency.

In multi-view coding, we provide a new type of frame, which we call a "V-frame" to increase compression efficiency. Specifically, a V-frame is like an I-frame in the sense that the V-frame is encoded without any temporal prediction. However, the V-frame also allows prediction from other cameras or prediction from synthetically generated frames. By inserting V-frames instead of I-frames in the bitstream in a periodic way, we obtain the same temporal random access as with I-frames, but with a better coding efficiency.

In the H.264/AVC video coding standard, IDR frames imply that all reference frames are removed from the decoder picture buffer. In this way, the frame before the IDR frame cannot be used to perform prediction for frames after the IDR frame. In the multiview decoder, V-frames imply that all temporal reference frames are removed from the decoder picture buffer, however spatial reference frames remain in the decoder picture buffer. In this way, a frame in a given view before the V-frame cannot be used to perform temporal prediction for a frame in the same view after the V-frame.

4. REFERENCE PICTURE MANAGEMENT

Before encoding a frame in our multiview coding system, reference frames to be used for multiview coding and decoding are inserted into the multiview reference picture list. All frames inserted into this picture list are initialized and marked as usable for reference using an appropriate syntax. For example, according to the H.264/AVC standard and reference software, the used_for_reference flag is set to 1. All frames in the multiview reference picture list are stored into the decoded picture buffer (DPB) and the relevant reference picture lists in such a way that the frames are treated as normal reference frames by the encoder or decoder. After the current frame is encoded or decoded, all multiview reference frames are removed from the DPB and from any other short or long term reference lists that a single-view video encoder or decoder would maintain.

To ensure that the encoder and decoder are operating consistently, a multiview frame convention is established where a fixed pattern of frame insertion operations are performed at each step. In our current system, we use a convention that reserves a portion of each reference picture list for temporal reference frames and reserves the remaining portion for multiview reference frames including neighboring or synthesized frames that may also be used for prediction.

Many conventions to insert multiview frames into the multi-frame reference picture list are possible. Therefore, the particular convention that is used should either be directly coded in the bitstream or provided as sequence level side information, e.g., configuration information that is communicated out of band. The means to represent the convention should be general enough to allow for different types of camera configurations, e.g., both 1D and 2D arrays, as well as different prediction structures.

With the above reference picture management, the encoding and decoding processes for multiview coding are compatible with single-view processing, which allows for existing codec designs to be easily extended for multiview video coding and decoding. We believe that this is a very important aspect to speed the deployment of multi-view coding systems.

5. CONCLUDING REMARKS

We presented several techniques used in our multiview video coding system, including prediction from synthesized views, spatial random access and multiview reference picture management. Using these techniques, our system achieves high coding efficiency and maintains compatibility with existing single view video coding schemes.

Experimental results that validate the effectiveness of the proposed techniques will be submitted in response to MPEG's Call for Proposal on Multiview Video Coding [4]. The results will also be reported in a future publication.

REFERENCES

[1] A. Smolic, and P. Kauff, "Interactive 3D Video Representation and Coding Technologies", Proceedings of the IEEE, Special Issue on Advances in Video Coding and Delivery, vol. 93, no. 1, Jan. 2005.

[1] S. C. Chan, et al., "The data compression of simplified dynamic light fields," Proc. IEEE Int. Acoustics, Speech, and Signal Processing Conf., April 2003

[2] C. Buehler et al., "Unstructured Lumigraph Rendering," Proc. ACM SIGGRAPH, 2001.

[3] ISO/IEC JTC1/SC29/WG11, "Updated Call for Proposals on Multi-View Video Coding", Doc. N7657, Nice, France, Oct. 2005.