21st International Conference on Pattern Recognition (ICPR 2012)
November 11-15, 2012. Tsukuba, Japan

RGB and Depth Intra-Frame Cross-Compression for Low Bandwidth 3D Video

Karthik Mahesh Varadarajan, Kai Zhou, Markus Vincze
Technical University of Vienna
{kv,kz,mv}@acin.tuwien.ac.at

Abstract

With the recent explosion in the development of multimedia hardware capable of 3D display, 3D Picture Coding Sytems have assumed a pivotal role. While encoding techniques for stereo-scopic images is a well researched topic and compression standards such as MPEG provide variants to support it, compression of RGB-D data such as from the Microsoft Kinect sensor offers a number of unsolved challenges. Projected texture based active sensors such as the Kinect offer a number of advantages in comparison with traditional 3D capture systems. While not affecting the visible spectrum of the scene these sensors are capable of producing highly accurate 3D reconstructions of complex scenes (as well as novel viewpoints) - even those with homogeneous surfaces lacking textural features that form the foundation of stereoscopic range measurement systems. Conventional approaches to compressing the RGB and D images separately are suboptimal in terms of compression efficiency, bandwidth usage and scalability. On the other hand, state-of-the-art methods in the field are not suitable for low bandwidth applications, typical of mobile phone devices, on-field civilian and defense robotic systems, especially those operating on unreliable or high-loss wireless networks. In order to address these concerns, we present a novel RGB-D Cross-Compression algorithm that can be used for static 3D scene reconstruction as well as intra-frame coding of 3D videos. The algorithm detects salient edge-like structures in RGB and D images and perform cross-coding across the modalities to yield a scalable system for 3D video coding. Results presented using the Microsoft Ballet and Breakdancers test sequences demonstrate the efficiency of the system in terms of compression rate, reconstruction quality and rate-distortion characteristics. The scalability of the approach also makes it well suited for mobile and wireless applications.

1. Introduction

3D video coding has achieved prominence in recent years due to the large number of 3D display capable hardware. Mobile phones such as the HTC EVO 3D and the LG Optimus 3D are devices that are not just capable of 3D image acquisition, but also 3D display. 3D television (3DTV) has also been in vogue the last few years and the technology has been rapidly growing. These systems for multi-view video (MVV), being predominantly based on stereoscopic image acquisition and display systems or similar approaches show poor performance in regions of homogeneous texture, where stereo-matching fails and are thus restricted in operation to applications which do not require full 3D user interaction or novel 3D viewpoint generation – free viewpoint video (FVV). With the advent of the Microsoft Kinect projected texture based active sensing system as well as other similar devices, such restrictions to the generation of 3D imagery are being largely overcome, resulting in the creation of efficient systems for Depth Image Based Rendering (DIBR) mechanisms. Nevertheless the compression techniques available for encoding such RGB-D imagery remains nascent, in comparison with techniques for encoding stereoscopic imagery/ video that are quite mature.

State-of-the-art techniques in the field of encoding RGB-D imagery include [1], in which an optimal joint-bit allocation model based on the RGB and D images is estimated and used for compression. Other methods such as Locally Adaptive Resolution (LAR) [2] are targeted at providing visually pleasing surface reconstructions. This approach uses layered coding of depth maps using variable block size representations based on quad-trees. On the other hand, sparse representation schemes such as [3], use depth transform domain optimization (in this case, L_1 norm minimization over the representation function) to yield the necessary encoding. Use of disparity maps instead of depth maps and depth propagation/ estimation has been studied in [4]. These approaches, as with

conventional approaches such as MPEG applied to depth maps suffer from edge smoothing and other artifacts. Edge sensitive depth image compression has been presented in [5]. This paper uses 'Wedgelets' and 'Platelets' for compression and is again based on quad-tree decompositions of the maps.

Nevertheless, none of these methods are well suited for Kinect like depth sensors and especially for low bandwidth applications. Since the Kinect sensor suffers from depth edge localization errors, it is necessary to build a scheme that is robust to such errors and can make use of the high fidelity RGB data to localize depth edges. Furthermore, all the above approaches, while being suitable for compressing high bit-rate depth maps independently or in the form of multi-view video, lack features for cross-modal compression that can greatly reduce the bit rate, enabling low bandwidth applications.

2. Algorithm

The main contribution of this paper is in presenting an RGB-D Cross modality compression scheme that provides efficient compression by taking advantage of the structural content co-localization across the two sensing modalities. The presented scheme is also designed to be scalable and targeted at superior performance especially for low bandwidth applications such as for mobile phones or civilian and defense robots operating across unreliable wireless networks. We use the algorithm presented in [8] and applied to compression of infra-red imagery in [9] as the motivational basis of our approach. Similar to [9], we employ an edge-based structural content detection algorithm for pattern driven compression. While edge-based approaches are admittedly expensive in the representation of highly textured scenes with low structural content, the trade-off in efficiency in relation to perceptual quality for edge based compression is particularly suited for our target application – low bandwidth compression with 3D structural fidelity preservation. The key components of the algorithm are presented as follows.

2.1. Cross-modal Edge Detection

The first step in the pipeline involves the detection of salient edges in the RGB-D image. To this end we employ a multi-scale version of the structural tensor [10] based canny edge detector that works across five channels (RG,B,D and DG- the depth gradient) to estimate RGB-D edges. The edge contribution from the depth gradient helps in the estimation of surface curvature or orientation changes while that from the depth images help detect depth jumps. While simply summing the differential structure across the various channels may result in cancellation of the component structures, tensors defined in the range 0 to π provide a mechanism to preserve the components by describing the local orientation of the gradients rather than the overall direction. We represent the RGB-D image as $\boldsymbol{f} = (f_R, f_B, f_G, f_D, f_{DG})^T$, the structure tensor is given by

$$\boldsymbol{G} = \begin{pmatrix} \dfrac{\overline{w_x^2 \boldsymbol{f}_x \cdot \boldsymbol{f}_x}}{\overline{w_x^2}} & \dfrac{\overline{w_x w_y \boldsymbol{f}_x \cdot \boldsymbol{f}_y}}{\overline{w_x w_y}} \\ \dfrac{\overline{w_y w_x \boldsymbol{f}_y \cdot \boldsymbol{f}_x}}{\overline{w_y w_x}} & \dfrac{\overline{w_y^2 \boldsymbol{f}_y \cdot \boldsymbol{f}_y}}{w_y^2} \end{pmatrix}$$

where the subscript notation is used to denote partial derivatives and the weights $w$are associated with per-pixel measurements – in our case, Gaussian scales. It should be noted that the elements of the tensor are invariant with respect to rotation and translation of the spatial axes. While it is possible to also build a photometric, shading invariant version of the tensor, for the given application of compressing scenes with typically large fields of view this is unnecessary. Eigen value analysis of the tensor results in two eigen values along with the prominent local orientation.

$$\lambda_1 = \frac{1}{2}(\overline{\boldsymbol{g}_x \cdot \boldsymbol{g}_x} + \overline{\boldsymbol{g}_y \cdot \boldsymbol{g}_y} + \sqrt{(\overline{\boldsymbol{g}_x \cdot \boldsymbol{g}_x} - \overline{\boldsymbol{g}_y \cdot \boldsymbol{g}_y})^2 + (2\overline{\boldsymbol{g}_x \cdot \boldsymbol{g}_y})^2})$$

$$\lambda_2 = \frac{1}{2}(\overline{\boldsymbol{g}_x \cdot \boldsymbol{g}_x} + \overline{\boldsymbol{g}_y \cdot \boldsymbol{g}_y} - \sqrt{(\overline{\boldsymbol{g}_x \cdot \boldsymbol{g}_x} - \overline{\boldsymbol{g}_y \cdot \boldsymbol{g}_y})^2 + (2\overline{\boldsymbol{g}_x \cdot \boldsymbol{g}_y})^2})$$

The direction of λ_1 indicates the prominent local orientation which is given as

$$\theta = \frac{1}{2} arctan\left(\frac{2\overline{\boldsymbol{g}_x \cdot \boldsymbol{g}_y}}{\overline{\boldsymbol{g}_x \cdot \boldsymbol{g}_x} - \overline{\boldsymbol{g}_y \cdot \boldsymbol{g}_y}}\right)$$

Non-maxima suppression on the λ_1 , the maximal eigen value, yields the required edge image. The composite edges are thus obtained as a combination of the edge gradients from the two sensing modalities. It can be expected that much of the edge contribution from the depth map overlaps with that from the color image. For the case of data from the Microsoft Kinect Sensor, it is expected that localization errors along the structural boundaries in the depth maps might create errors in the cross-modal edge combination process. To this end, we can refine the depth map using the scheme presented in [11] prior to edge detection.

2.2. Contour Characterization

The detected edges are then thinned and cleaned to yield cross-modal contours. Contours are characterized by their start and end points along with the profile of the RGB and D images along the contours. In the case of junctions and branch points, it is necessary to split the contour. Similar to [9] we choose the continuity of contours at junctions along paths that maximize the length of the edge chain, while resulting in a minimum gradient variation thereby enabling smooth transition between edge segments. The encoding of each contour includes the

geometric and intensity (RGB-D) profiles of the structural discontinuities in the image. The spatial profiles of the contours are listed with the primary end point chosen to be closer to the image plane origin. Furthermore, the contours are sorted with contours having starting points closer to the origin being higher on the list. This sorting helps optimize the coding process. In order to improve the coding efficiency of each contour, a piecewise linear approximation of the contour is used. The piecewise linear approximation is calculated using the spatial and the intensity profiles. For each contour, the values of position and intensity - $i, j, r_1, r_2, g_1, g_2, b_1, b_2, d_1, d_2$ and σ (the color channel blur) are stored and encoded at every anchor point (obtained by the piecewise approximation). The blur σ can be computed as

$$\sigma_{i,j} = \sqrt{\left(\frac{d}{2}\right)^2 - s^2}$$

where i, j indicate the point locations, d is the distance in pixels between extrema in 2nd derivative map and s is the critical scale at the point. The subscripts *1* and *2* for the intensity profiles indicate the high and low intensity values across the anchor point. Since the values can suffer from localization artifacts, we employ a linear interpolation on either side of the contour in a direction normal to it in order to obtain stable profile values. These values are selected such that they do not interfere with the neighboring edge pixel values.

2.3. Scalable Contour Coding

To provide for scalability in the encoding process, it is necessary to prioritize the contour list in terms of visual significance. Based on the available number of bits, it is possible to transmit fewer contours based on this priority to enable a scalable/progressive reconstruction at the decoder. Our ranking system uses a combination of multiple cues to prioritize the ranking of the contours. A geometric mean of the individual cue ranks is used to combine the ranks, resulting in the emphasis of contours with balanced ranks rather than those with high imbalance in rankings. The overall rank is based on the intensity difference ranks ($i_r, i_g, i_b, i_d, i_{dg}$), - length of the contour (l) and number of anchor points (p) – in each larger values lead to numerically lower ranks.

$$C_{rank} = \left(i_r + i_g + i_b + i_d + i_{dg}\right).\frac{l}{p}$$

Since the spatial profile values, intensity profiles and the blur values of all contours are expected to be highly correlated within their respective profile domains, the profile sequences are encoded independently. Since the spatial location cannot be encoded in a lossy manner, so we chose to encode it using Differential Pulse Code Modulation (DPCM). Since the spatial lists have already been sorted and selected by the ranking system based on the available bit budget, this results in optimal coding of the spatial profile. A DPCM between edge chains also helps reduce the bandwidth requirements even further. For the case of the RGB and Depth intensity profiles a 1 dimensional Discrete Cosine Transform (1D –DCT) is applied, followed by quantization (with different quantization levels for RGB and D) and truncation of terminating zeros. While the DC values resulting from the DCT are again subjected to DPCM, the resulting AC/DC values are encoded using CABAC (Content Adaptive Binary Arithmetic Coding). This results in optimal encoding of contours. The encoded contours can then be stored or transmitted progressively based on available bandwidth.

2.4. Contour Decoding and RGB-D Reconstruction

The Contour decoding follows the reverse process as the encoder and performs CABAC decoding followed by inverse DCT and inverse DPCM resulting in the generation of the spatial and intensity domain values. Using the profiles obtained, a skeleton image of contours is reconstructed for the R, G, B and D channels. The full image reconstruction from the skeleton is carried out using image inpainting and depth diffusion processes. The algorithm used for the reconstruction is described in [12]. An anisoptropic Laplacian heat diffusion partial differential equation is used for the inpainting based on the Iterative Back Substitution (IBS) algorithm. While the skeleton image provides the Neumann boundary conditions for the process, the Dirichlet boundary conditions are generated based on the image boundaries.

3. Experiments and Results

The designed codec was tested on a number of different RGB-D images/ image sequences, including those obtained from the Kinect. Figure 2 shows sample RGB and Depth images for compression, with an extremely low bit budget of 12 kbits (minimum operating bandwidth for MJPEG) and the reconstruction obtained after decoding. The PSNR values for the RGB and D images were roughly 31dB and 65 dB respectively. It can be seen from the figure that the performance of our codec is superior to traditional approaches such as MJPEG especially under conditions for the targeted application – operation under low bit rate/ unreliable networks. It can also be seen from Figure 1(a) that the system is scalable and rate-distortion characteristics fall off

gracefully at extremely low bit rates. Figure 1(b) shows the compression efficiency for intra-frame coding of two complete benchmark sequences (Microsoft RGB-D: Ballet, Breakdancers). While numbers for PSNR are similar form both schemes across the 100 frames, it can be seen from the results in Figure 2 that visual quality of the reconstructed images using our approach is significantly better using our approach. It should also be noted that our codec does not suffer from blocking, blurring or ringing artifacts typical of conventional coding schemes.

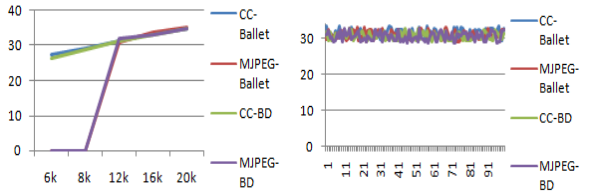

Figure 1. (a) PSNR for a compression target of 12 kB for our codec (CC) and MJPEG (b) Rate distortion characteristics (PSNR vs BR)

4. Conclusion and Future Work

In this paper we have presented a novel algorithm for cross-modality compression of RGB-D images. While the greater proclivity for the usage of active range sensing devices such as the Microsoft Kinect, coupled with the phenomenal growth in 3D display systems have forced the need for better 3D compression, our system tries to address some of these concerns, especially in the domain of low bandwidth applications. Future work involves extending the system to handle inter-frame and predictive coding for enhanced compression of full 3D videos.

References

[1] H.Yuan, et al, "Model-Based Joint Bit Allocation Between Texture Videos & Depth Maps for 3-D Video Coding", Circuits Systems for Video Technology, 2011.

[2] Bosc, E.Pressigout, M, Morin, L., "Focus on visual rendering quality through content-based depth map coding", Picture Coding Symposium, PCS 2010.

[3] Gene Cheung, Akira Kubota, Antonio Ortega, "Sparse representation of depth maps for efficient transform coding", Picture Coding Symposium, PCS 2010.

[4] Mueller K., Merkle, P. Wiegand, T., "3-D Video Representation Using Depth Maps", Proceedings of IEEE, Vol. 99, No. 4, April 2011.

[5] Yannick Morvan, Peter H. N. de With, Dirk Farin, Platelet-based coding of depth maps for the transmission of multiview images, SPIE, Stereo Disp., 2006.

[6] C. Fehn, "3D-TV using depth-image-based rendering (DIBR)", PCS, 2004.

[7] J. van de Weijer, et al, Robust Photometric Invariant Features from the Color Tensor, IEEE Trans. Image Processing, vol. 15 (1): 118-127, January 2006.

[8] Elder Zucker Image Compression (http://www.stanford.edu/class/ee368b/Projects/cnetzer/)

[9] H. Wei, S. Zabuawala, KM. Varadarajan, et al. "Adaptive pattern-based image compression for ultra-low bandwidth weapon seeker image communication". Visual Information Processing 2009: 73410

[10] J. Weijer, Gevers, T. ; Smeulders, A.W.M., "Robust Photometric Invariant Features from the Color TensorTensor", Image Processing, IEEE Trans., 2006.

[11] KM Varadarajan, M. Vincze, "MRF based Kinect data Refinement", ACCV submitted 2012

[12] KM Varadarajan, M. Vincze ''Real-Time Depth Diffusion for 3D Surface Reconstruction', ICIP 2010

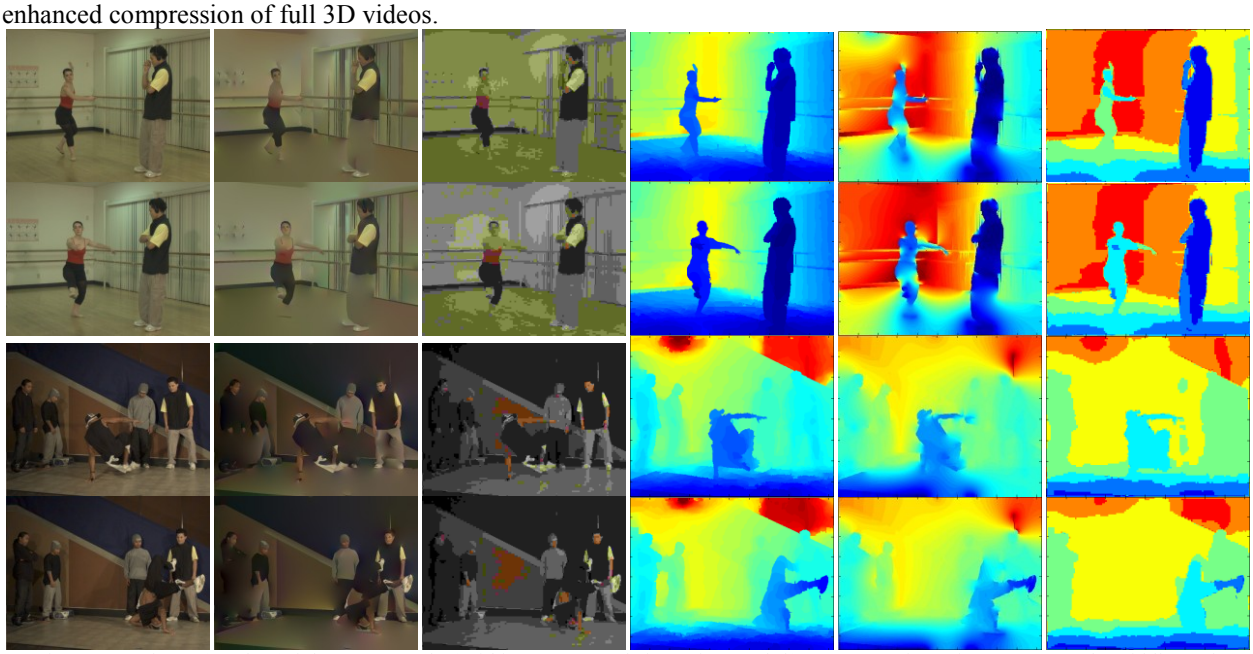

Figure 2. (Left Pane) Color (Right Pane) Depth: (Within pane) Sample input frames (left), reconstruction using our approach (middle) and MJPEG (right) for 12kb combined bit rate. Notice realistic boundaries and color, perceptual fidelity with our approach, whereas results using MJPEG suffer from several artifacts.