3-D VIDEO COMPOSITING : TOWARDS A COMPACT REPRESENTATION FOR VIDEO SEQUENCES

Fernando C. M. Martins and José M. F. Moura

Department of Electrical and Computer Engineering
Carnegie Mellon University
Pittsburgh, PA 15213–3890

ABSTRACT

In order to achieve good quality very low bit rate video coding, new techniques leading to highly compact representations for video sequences must be investigated. We present a novel video codec framework, where video sequences are represented in terms of stochastic non-parametric 3-D object models and motion script estimates. Multi-property object models, carrying both shape and color information, are incrementally built from video and range sequences. Motion estimates are obtained by depth map registration. We refer to this framework as 3-D Video Compositing, or 3DVC for short. In this paper, we will describe 3DVC in detail, and present experimental results where interframe compression ratios in the range of 10^2 to 10^3 have been achieved.

Keywords: Video sequence representation, Model based coding, VLBR video coding, non-parametric 3-D object modeling, range and image sequence processing.

1. INTRODUCTION

Digital video handling entails dealing with massive amounts of data. Without compression, a bandwidth of more than 9 $Mbit/s$ is required to deliver color stamp-sized QCIF images at 30 frames per second. For NTSC quality video, the bandwith requirements increase to more than 160 $Mbit/s$. Cost effective video transmission over low cost 8 to 64 $Kbit/s$ channels demand coding techniques capable of producing compression ratios in the range of 10^3 to 10^4 [1].

Techniques for VLBR video coding can be classified into two major groups: waveform based coding and model based coding. In waveform based coding, the video sequence is considered a multidimensional signal, while in model based coding the images that compose the video sequence are seen as 2-D projections of a 3-D scene [2].

In this paper, we present 3DVC – a model based video coding technique that relies on incremental stochastic non-parametric 3-D object modeling.

During encoding, 3-D stochastic object models are incrementally built without user interference. This is accomplished by the integration of depth and intensity information in a Bayesian framework. Redundant information is used to reduce model entropy, before being discarded by the transmitting end.

Motion is estimated by depth map registration, and is sequentially stored in a motion script for the given object. The final set of constructed object models and motion scripts provides a compact representation for the original video sequence.

To reconstruct each video frame at the receiving end, the 3-D scene is recreated by positioning the constructed 3-D object models in space according to the corresponding motion scripts. The 2-D frames are then reconstructed by a first opacity raycasting volume renderer.

To be able to recreate each frame at the receiving end, 3DVC requires the transmission of object model updates and current motion estimates. As frames are processed and model entropy converges, fewer model updates are transmitted per frame. If motion ceases, no data transmission is required. The variable bit rate characteristic of 3DVC can be explored in packet video based on ATM networks [3].

The stochastic nature of the object models enables robust operation when dealing with inconsistent and noisy measurements. It also enables 3DVC to support both active exploration and passive integration of sensory data. Model entropy is used to guide exploration if active sensors are available.

The non-parametric 3DVC object models are compact uniform tessellations $\Gamma = \{C_i\}$ of a 3-D space, where each cell C_i represents multiple properties in a probabilistic way.

The work of the first author is partially supported by the National Council for Scientific and Technological Development CNPq-Brazil.

0-8186-7310-9/95 $4.00

This non-parametric volumetric description is suitable for model based coding because free-formed objects are supported with selectable spatial resolution, frame rendering performance is independent of object and scene complexity, and the parallel nature of raycasting algorithms can be explored [?].

Besides being compact, 3DVC enables content based video handling and editing. Distinct video sequences can be created at the receiving end by altering the motion scripts and/or object models generated at the transmitting end. Variable focus of attention, cast selection, content based search, and insertion/deletion of virtual and real entities are some of the explorable features of 3DVC. These issues are explored elsewhere.

In this paper, we present the 3DVC framework and demonstrate its suitability for VLBR model based coding. In Section 2, we present the 3DVC codec framework in detail. In Section 3, the experimental results that illustrate the framework potential for VLBR compression are shown. Section 4 concludes the paper.

2. 3DVC FRAMEWORK

2.1. Object Model Structure

A 3DVC object model $\Gamma = \{C_i\}$ is a uniform tessellation of a compact volume. Each cell C_i of this 3-D regular grid has multiple properties, as occupancy $O(C_i)$ and color $T(C_i)$. The properties may assume discrete values from a finite set, i.e., $O(C_i) \in \{occupied, empty\}$. Each cell C_i holds a probability distribution function for every property. For instance, the function $p(O(C_i))$ is stored for occupancy.

Holding distributions instead of current estimates is what makes Γ a useful representation for integration of multiple measurements. Initial lack of knowledge is expressed by assigning equiprobable probability density functions to all properties.

Earlier work on sensor fusion for robot navigation and object modeling for robotic manipulation have successfully explored this stochastic model structure [5].

2.2. Motion Estimation

We perform motion estimation by depth map registration using a variant of the ICP algorithm capable of dealing with incomplete models [6]. The depth measurement associated with the k^{th} frame is registered with respect to the model Γ being built, and an object position and orientation estimate $\vec{q_k}$ is obtained.

For the first frame, the model Γ holds no information to allow registration. For this frame the current object position and orientation is assumed to be the canonical position $\vec{q_0}$.

2.3. Incremental Object Model Construction

The integration of a set of multiple measurements $R_k = \{r_0, r_1, \ldots, r_k\}$ into a single model Γ_k can be treated as the classical random parameter estimation problem in a Bayesian framework. If we select a uniform cost function, the optimal estimator for a given property z accessible through the measurements R_k is the MAP estimator $\hat{z}_k$ [7]. We introduce the notation:

$$p(z)_k = p(z|R_k) \quad (1)$$

$$\hat{z}_k = \arg\max p(z)_k \quad (2)$$

where $p(z|R_k)$ is the the conditional probability distribution of the property z given the set of measurements R_k.

Considering that the measurements R_{k+1} are independent, and applying Bayes' theorem, the incremental update rule follows:

$$p(z)_{k+1} = \frac{p(r_{k+1}|z)p(z)_k}{p(r_{k+1})} \quad (3)$$

The conditional probability distribution $p(r|z)$ of the observation r given property z, is known as sensor model. It is specified by prior knowledge and assumptions regarding the sensors. The estimate $\hat{z}_k$ is computed by equation (2), and is recursively updated in time through (3).

We consider the estimation of shape $\hat{O}(C_i)$ from a sequence of depth measurements R_k. Occupancy $O(C_i)$ is a property that is not directly measurable, but can be estimated by:

$$\hat{O}(C_i) = \arg\max P(O(C_i)|r) \quad (4)$$

$$P(O(C_i)|r) = \frac{p(r|O(C_i))P(O(C_i))}{P(r)} \quad (5)$$

$$p(r|O(C_i)) = \sum_{G \in \{G(C_i)\}} p(r|O(C_i), G)P(G) \quad (6)$$

$$p(r|G) = p(r|z_{min}) \quad (7)$$

In equations (4-7), we assume that the set of measurements R_k correspond to the same physical property, taken with respect to the same reference, and by the same device.

If distinct sensors are available, we generalize equations (4-7) to include the corresponding sensor models $p^{\alpha}(r|z)$. If measurements are taken with respect to distinct references, we pre-process and transform the measurements and the sensor model to a canonical reference before integration. It is important to notice that the subset of cells of Γ_{k-1} that are affected during the integration of measurement r_k depend both on the sensor model $p(r|z)$ and on the reference system used while

measuring r_k. This is the case when there is relative motion between object and sensor. The transformation required to have the current measurement and sensor model refer to a canonical reference is obtained from the canonical and k^{th} frame position and orientation estimates $\vec{q_0}$ and $\vec{q_k}$ [6].

3. EXPERIMENT

We produced a synthetic video and range sequence of 150 frames, 100 × 100 pixels per frame, 8 bits per pixel. The scene is composed of a single rigid object performing 6DOF 3-D motion in front of a static background. In Figs. 1 and 2 sample frames from the described input range and video sequences are shown.

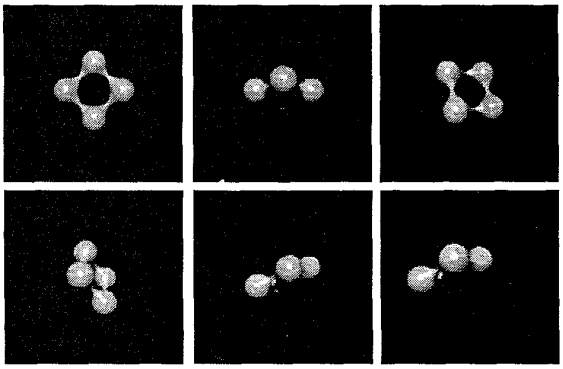


Fig. 1: Samples from input video sequence.

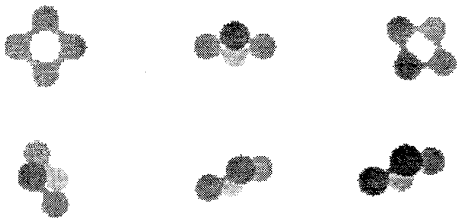
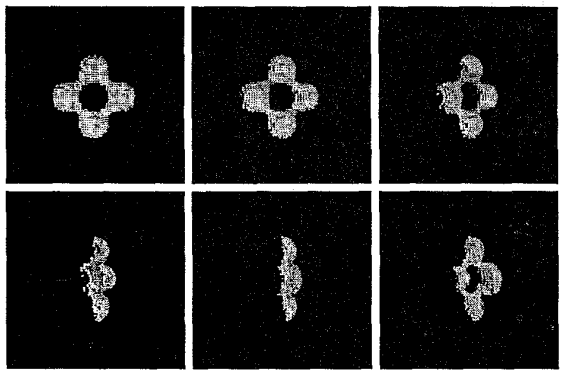


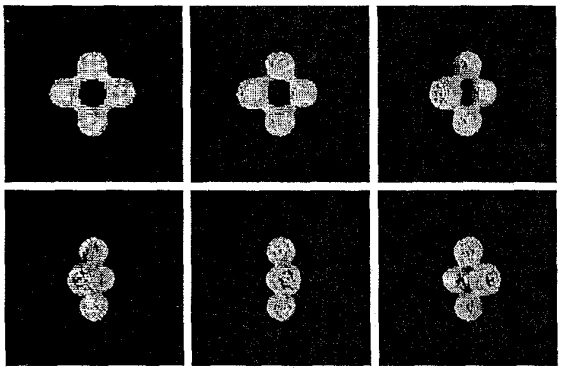
Fig. 2: Samples from input range data sequence.

In this experiment, the depth sensor is considered ideal, segmentation is obtained by thresholding depth information, and the object position and orientation $\vec{q_k}$ is known for all frames. Using these assumptions, the stochastic model Γ_k is then incrementally built according to Section 2.3.

In order to provide means for visual inspection of the constructed 3-D object model, we create a visual representation in terms of constructive solid geometry (CSG) primitives. For each cell C_i of the tessellation Γ, we create a small sphere positioned in 3-D space according to the position of the cell, with radius proportional to the occupancy probability $P(O(C_i) = occupied)$. In Fig. 3, several rendered views of the CSG representation of the 3-D model are presented for two distinct time instants Γ_2 and Γ_{150}. In Fig. 3(a), two frames have been processed. Eventhough the model Γ_2 is clearly incomplete, it is sufficient for the reconstruction of the 2 frames already processed. In Fig. 3(b), 150 frames have been processed achieving a higher level of model completeness.



(a)



(b)

Fig. 3: Views of 3-D object model in distinct time instants (a) Γ_2 and (b) Γ_{150}.

3DVC provides a compact representation for video. For each frame, 3DVC requires the transmission of object pose and model updates. The transmission of a 6DOF pose estimate, without source coding, requires 24 bytes. Each occupied cell C_i requires 4 bytes to carry position and color information. The total amount of data required by 3DVC is given by equation (8). Compression ratio is evaluated through equation (10).

$$S_{3DVC} = 4M + 24F_m \quad (8)$$

$$S_{raw} = 100^2 F \quad (9)$$

$$\mathcal{C} = \frac{S_{3DVC}}{S_{raw}} \quad (10)$$

In equations (8-10), M is the number of occupied cells in the object model, F_m is the number of frames requiring transmission of motion updates, and F is the number of frames in the sequence.

The worst case scenario for compression is when pose estimates are required for all frames and a complete model is required due to complex motion patterns. This case is represented in this experiment by model Γ_{150}, which has $M = 3962$ occupied cells, and by setting $F_m = F$. This leads to $\mathcal{C}_{short} = 1 : 77$ for a sequence of 5 seconds, and $\mathcal{C}_{long} = 1 : 304$ for a sequence of 1 minute.

In a more realistic scenario, pose estimates are required for a fraction of the frames, and an incomplete model suffices. We assume that 30% of the frames require pose estimates, i.e. $F_m = 3F/10$, and that the incomplete model Γ_2 with $M = 1056$ occupied cells suffices. The compression ratios obtained are $\mathcal{C}_{short} = 1 : 277$ for a sequence of 5 seconds, and $\mathcal{C}_{long} = 1 : 967$ for a longer 1 minute sequence.

It is important to notice that these compression ratios are achieved without further source coding, and refer mainly to interframe coding. Additional compression results by intraframe coding, e.g. transform based coding, or by choosing a tessellation Γ with lower spatial resolution.

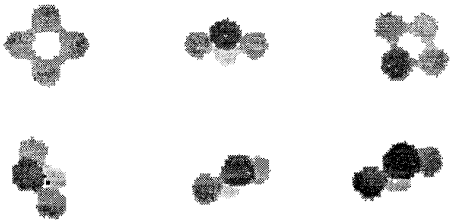


Fig. 4: Samples from reconstructed range sequence.

3DVC provides selectable compression rate and reconstruction quality. The number of cells in Γ per unit of volume defines the spatial resolution of the object model. High resolution models provide high quality reconstruction, but the size of the model M is large. Theoretically, lossless compression is achievable. Lowering the model's resolution leads to higher compression, but it also introduces subsampling artifacts on the reconstructed frames.

Figure 4 presents six samples from the reconstructed range sequence. We can see that even after three orders of magnitude compression, the frames in Fig. 4 are a very good reproduction of the original presented in Fig. 2, with practically no noticeable artifacts.

4. SUMMARY

We introduced 3DVC - a video codec technique that provides a compact representation for video sequences. By representing video sequences with perceptually significant 3-D elementary building blocks, constructed object models and motion scripts, 3DVC furnishes additional functionality to video handling and eliminates interframe redundancy.

In the experiment described, we showed that the interframe compression ratio grows linearly with video sequence length, and also depends on the complexity of scene dynamics. High quality results with interframe compression ratios in the range of 100 to 1000 have been achieved in the absence of further applicable intraframe coding. This illustrates the compactness of 3DVC, and its potential applications to VLBR video coding.

5. REFERENCES

[1] C. Chen, "Video compression: Standards and applications," *Journal of Visual Communication and Image Representation*, vol. 4, pp. 103–111, June 1993.

[2] H. Li, A. Lundmark, and R. Forchheimer, "Image sequence coding at very low bitrates: A review," *IEEE Transactions on Image Processing*, vol. 3, pp. 589–609, September 1994.

[3] F. Kishino, K. Manabe, Y. Hayashi, and H. Yasuda, "Variable bit-rate coding of video signals for ATM networks," *IEEE Journal on Selected Areas in Communications*, vol. 7, pp. 801–6, June 1989.

[4] A. Kaufman, *Volume Visualization*. Los Alamitos, CA: IEEE Computer Society Press, 1991.

[5] A. Elfes, "Occupancy grids: A stochastic spatial representation for active robot perception," in *Proceedings of the Sixth Conference on Uncertainty and AI*, (Cambridge, MA), AAAI, July 1990.

[6] P. Besl and H. Mckay, "A method for registration of 3-D shapes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 14, pp. 239–56, February 1992.

[7] H. V. Trees, *Detection, Estimation, and Modulation Theory - vol 1*. John Wiley and Sons, 1968.