TIME-EVOLVING 3D MODEL REPRESENTATION FOR SCALABLE VIDEO CODING

Raphaèle Balter [1,2], *Patrick Gioia* [1], *Luce Morin* [2]

[1] FRANCE TELECOM R & D, 4 rue du Clos Courtel, 35512 Cesson-Sevigne, France
[2] IRISA-INRIA, Campus de Beaulieu, avenue du General Leclerc, 35042 Rennes, France
email: {raphaele.balter,patrick.gioia}@rd.francetelecom.com and {rbalter, lmorin}@irisa.fr

ABSTRACT

This paper presents an efficient and scalable coding scheme for transmitting a stream of 3D models extracted from a video of a static scene. As in classical model-based video coding, the geometry, connectivity, and texture of the 3D models have to be transmitted, as well as the camera position for each frame in the original video. The proposed method is based on exploiting the interrelations existing between each type of information, instead of coding them independently, allowing a better prediction between the media and of the next information in the stream. Scalability is achieved through the use of wavelet-based representations for both texture and geometry of the models.

A consistent connectivity is built for all 3D models extracted from the video sequence in order to get a consistent representation of the sequence possibly evolving in time. This allows a more compact representation and straightforward geometric morphing between successive time representations of the model. Furthermore this leads to a consistent wavelet decomposition for 3D models in the stream. Targeted applications include distant visualization of the original video at very low bitrate and interactive navigation in the extracted 3D scene on heterogeneous terminals.

1. INTRODUCTION

Video compression for broadcasting of audiovisual content is one of the most challenging issues in signal coding; as coding methods become more and more performing, acceptable bit-rates are stand still and even decrease with the recent need for video over cell phones or PDAs. State-of-the art video coding relies on pixel-based prediction / correction paradigms. While these approaches have globally proven to be the most efficient, research showed that exploitation of particularities of the encoded content can dramatically reduce the size of its encoding by using specific video coders. Typically, 3D model-based video coding aims to reconstruct the 3D environment captured and send it as such. The advantages of this approach are many; the 3D is likely to be more compact than the pixel based representation, and once the scene transmitted, the 3D helps to perform free navigation or add virtual objects for augmented-reality applications.

The difficulty of recovering the 3D information from an image sequence is mainly due to image matching and calibration stages. Some methods exploit some particular feature of the scene or require manual intervention or a special capturing system [MB95]. Finally, some completely automatic methods are based on robust estimation of the camera parameters [Pol02, ZFC99, GM02]. Our work is situated in this context to meet video coding constraint i.e automatic reconstruction without a priori knowledge on camera parameters, scene contents nor video length.

Galpin [GM02] has proposed a coding scheme for a stream of 3D models, using the EBCOT encoder for coding texture and depth information. Each 3D model is valid for a portion of the original sequence called *GOF (Group of Frames)* , and is obtained by uniform triangulation and elevation of a dense depth map. However, this systematic definition of mesh vertices provides no vertex to vertex correspondence between two successive 3D models and produces visual artifacts at GOF transitions, due to geometry and connectivity jumps. Post-treatments based on 3D morphing have been proposed [GBMD04] but they are based on approximative correspondence and they add high computational cost at the decoding stage, unsuitable for our target applications. Moreover, this representation is not entirely scalable, since the mesh has fixed size and no multi-resolution structure.

While 3D model based coding has been widely studied in terms of modeling and acquisition, only few works have focused on compressing and streaming the resulting representation of the scene [MG00] [GBMD04]. Indeed, this aspect of the coding is crucial for broadcasting video or proposing virtual walk through, and should be done in a fully scalable way for displaying the content on arbitrary device, from the PC screen to the cell phone matrix.

In this paper we describe a novel algorithmic framework for compactly streaming image sequences based on 3D reconstruction. Our approach relies on an *evolving model* that is a compromise between the global consistency of a unique 3D model and the robustness for long sequences given by a 3D model stream. This evolving model is based on a consistent connectivity of the scene associated to a time-evolving geometry.

In the next section we expose an overview of our method and list the various blocks composing it. These particular techniques are detailed in Sections 3 to 5. Results on actual scene data are shown in Section 6 and future work is mentioned in Section 7.

2. OVERVIEW OF THE METHOD

We propose to code a video sequence as an evolving model given by a set of consistent 3D models but to use second generation wavelets to achieve efficient compression and scalability in the spatial and temporal domains. Moreover, we use meshes based on non uniform triangulation in order to ensure global consistency and smooth transition between models.

Our representation is based on 3D information extracted with the Galpin reconstruction algorithm using shape from motion methods [GM02]. The sequence is divided into GOFs of content adaptive sizes. For each GOF it provides a dense depth map and camera positions for each frame in the GOF. The first image of a GOF k is called a "key image" denoted K_k. It is used as texture image for the 3D model of the GOF. Our models are given by elevation

of non-uniform triangulations on those dense depth maps. The successive textured 3D models represent overlapping parts of the scene with a consistent connectivity and a time-evolving geometry.

Though theoretically elegant and commonly used for generic 3D models, applying multi-resolution wavelets schemes to the setting we deal with is not straightforward. Our idea is the derivation of a single connectivity model dissociated from geometry (which is represented by the wavelet signal). This model gathers the connectivities of every GOF model.

The first step is to transform each depth map into a hierarchical 3D triangular mesh. We define the "base mesh", denoted BM_k as the mesh at coarser level and the "fine mesh", denoted FM_k as the dense mesh at finer level. The refinement from coarser to finer level is then expressed as wavelets coefficients ($r_i{}^j$ on Fig. 1) using a second generation wavelet transform. Furthermore, successive 3D models in the stream are also encoded differentially with wavelet coefficients ($d_i{}^j$ on Fig. 1).

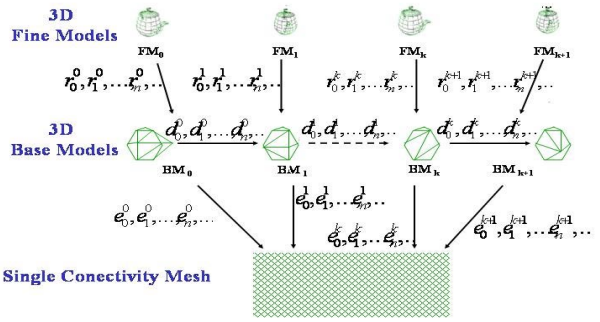

Fig. 1. Proposed representation based on a 3D model stream and second generation wavelets. The fine models FM_i are represented by the base meshes BM_i and wavelet coefficients r_i^j. The scale coefficients e_i^j expressing the geometry of the base meshes are gathered and indexed by the single connectivity mesh. The meshes are encoded incrementally in time by the wavelet coefficients d_i^j.

This representation induces several media streams, such as the consistent connectivity, geometry (wavelet coefficients and incremental model representation), and texture. These streams are closely interrelated and they are multiplexed in order to produce a single streamable format.

In the following sections we further describe the main components and the stream types they generate. We now explain how the hierarchical 3D triangular mesh is constructed from the dense depth map and then compressed.

3. EVOLVING MODEL CONSTRUCTION

The proposed evolving model is based on a 3D model stream with a consistent connectivity and a time evolving geometry. We build a consistent sampling for all 3D models with vertices corresponding to identical physical points. This is done by separating connectivity and geometry; a planar graph, denoted as *single connectivity mesh (SCM)*, gathers the connectivity information of every base mesh in the sequence, regardless of their geometry. This mesh evolves during time in order to take into account outgoing and incoming points. The SCM is computed starting from the connectivity information of the first base mesh, and updated with the connectivity information associated with new points appearing from one base mesh to another. The SCM computation and update is based on the base meshes construction described in Section 4. A global indexing system provides a unique index for each vertex in the SCM, thus implicitly defining matching between base meshes vertices. The SCM is described as a list of triangles expressed in the new global indexing system.

Wavelet decomposition based on the SCM is consistent for all models and leads to compact coding. Moreover, smooth swapping at visualization stage can then be achieved by direct morphing between vertices without ghost effect caused by fading [GM02] nor morphing additional computing cost [GBMD04]. Thanks to the consistent connectivity of all base meshes smooth transition between models M_k and M_{k+1} can be achieved at a given level i by linear interpolation between corresponding vertices:

$$M_c = \alpha * M^i{}_k + (1-\alpha) * M^i{}_{k+1} \text{ with } \alpha = \frac{\| t_{t_{k+1}} - t_{t_c} \|}{\| t_{t_{k+1}} - t_{t_k} \|},$$

where M_c denotes the interpolated model for current time t_c and t_{t_c}, t_{t_k} and $t_{t_{k+1}}$ denote translation vectors defining camera position for the current frame, keyframe K_k and K_{k+1} respectively.

4. BASE MESH CONSTRUCTION AND TRANSMISSION

In order to meet the SCM construction constraint every base mesh has to include the corresponding vertices of the same physical points of the 3D scene.

In the first GOF, the adaptive triangular mesh is based on interest points computed on the first frame in the GOF (key image), using the Harris corner detector [HS88]. A 2D Delaunay triangulation of these points provides the base mesh. The 3D model is then obtained by elevation of this 2D mesh with 3D information given by the depth map.

For the next GOFs, the base mesh is constrained to contain the correspondents of the previous base mesh vertices, if they are still visible in the GOF. We denote CM_k the mesh of the corresponding vertices and the associated faces: $CM_k \subseteq BM_k$. Correspondences are given by a dense motion field. When adding vertices on the border of the model, the new triangulation has to preserve the connectivity derived from the preceding GOF without edge crossing. This is achieved by a 2D Delaunay triangulation constrained by images borders and correspondent mesh CM_k borders.

Thanks to the 3D representation, 2D texture coordinates of the model vertices do not have to be transmitted, as it is usually the case when using a non uniform mesh. Indeed, as we know camera parameters for each frame, any 3D vertex M can be projected onto the GOF key image viewpoint. As the key image is also the texture image, the coordinates of the obtained projection m are the texture coordinates for vertex M. Thus we just have to encode three parameters instead of five for each vertex. Those parameters can be 3D coordinates of the point or 2D coordinates and the associated depth. We choose to encode 3D coordinates of the vertices. 3D coordinates for model BM_k can be differentially encoded with respect to the position of their correspondent in the previous base mesh BM_{k-1}, as coefficients $d_i{}^0$ on Fig. 1.

The base meshes are encoded using the TS (Topological Surgery) encoder [TR98] for geometry and connectivity. We can rapidly identify vertices having a correspondent in the next model by reprojecting vertices of the current model on the key image of the next

GOF. In this way, we retrieve the common information between two models at the decoding stage without transmitting additional information. The global indexing system introduced in Section 3 helps to implicitly encode correspondences between successive base meshes. In order to avoid numerical errors a robust selection of base mesh vertices is added to Harris corner selection.

Since 3D models represent overlapping parts of the scene, the related textures include redundant information. This information is exploited using a classical IPP scheme where the first image is in Intra mode and the others are in Predicted mode. Predicted image $\tilde{I}_{k+1}$ are obtained by the reprojection of the precedent textured model on the current key position. Padding is used in areas where prediction does not apply.

At this stage, we have obtained a set of coarse meshes based on non-uniform triangulation, with corresponding vertices. This representation has several advantages: vertex positions can be adapted to scene contents; vertex to vertex correspondence between successive models is implicitly provided by the mesh structure and thus need not be transmitted or estimated at the decoder; smooth transitions between 3D models through implicit morphing can be achieved at low computing cost, using a simple linear interpolation between vertices; vertex consistency allows a more efficient coding of 3D information through differential encoding.

In the next section, we describe the wavelet analysis scheme applied on the base meshes in order to provide a multi-resolution scalable representation for each 3D model.

5. WAVELET DECOMPOSITION

In both image coding and synthetic 3D models coding, wavelets [Dau92] have been effectively used to achieve scalability in an elegant and efficient way. Second generation wavelets [LDW97] provide hierarchical representations for arbitrary sampled data and they are the current most effective tool for scalable representation of 3D models [KSS00], and are already parts of the new MPEG-4 Animation Framework eXtension (AFX).

In our case, second generation wavelets are defined in order to describe the shape of the recovered depth maps. Since we describe geometric deformations, first generation wavelets do not apply. Indeed, these parameterizations are defined over topogical spaces (typically base meshes BM_k of Fig. 1) which are not linear spaces. In the context of Subdivision Surfaces [LDW97], wavelets can be defined starting from a low pass reconstruction filter P^j. This filter operates over a global topological subdivision consisting in facets quadrisections, similarly as interval dichotomies in the classical wavelet setting. Filter P^j transforms coefficients at level $j-1$ into a prediction at level j: $c^{j+1} = P^j c^j$.

The resulting coefficients are an approximation, without adding any information, which coincides with the refinement operator in the case of Subdivision Surfaces. The wavelet setting can be seen as "completing" the representation by adding details through a high pass reconstruction filter Q^j. We use lazy wavelet where P^j is an averaging filter and Q^j the identity. We then process the wavelet transform by applying $(P^j Q^j)^{-1}$ for every j:

$$\begin{pmatrix} c^j \\ d^j \end{pmatrix} = (P^j \; Q^j)^{-1} c^{j+1}. \qquad (1)$$

This wavelet representation is then further compressed with an adaptation [KSS00] of the classic SPIHT zero-tree encoder, adding bitplane scalability to spatial and frequency scalability provided by the wavelets themselves. Since zero-trees are indexed by edges of the base meshes, the planar graph SCM gives a standalone indexing reference for the wavelet decomposition of the 3D meshes composing the evolving model. For a more compact coding wavelet coefficients can also be encoded in a differential way as represented by coefficients ${d_i}^j$ on Fig. 1.

6. RESULTS AND DISCUSSION

We show results on two sequences, illustrating the compression rates reached by comparison with Galpin and H264 encoders [SW02] at low bitrates, on both constrained and free navigation.

While PSNR is appropriate for measuring block based errors, it has little meaning when it comes to geometric distortion or for free viewpoint images. We thus rather use visual quality to evaluate the quality of the reconstruction.

To smooth transition between models Galpin proposed a 3D fading which gives good results . But as soon as we take an image situated on the middle of the GOF or corresponding to a virtual viewpoint outside the original path, ghost effect appears because of geometric jump between GOFs as shown on Fig. 4.

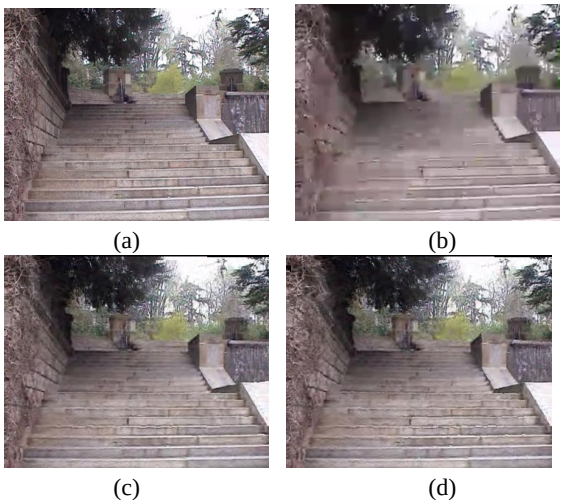

Fig. 2. "Thabor" sequence: Image 125 from original sequence (CIF, 25Hz) (a) and reconstructed images at 125kb/s with naive coder (b), with Galpin coder (c) and with our coder (d)

First of all Fig. 2 shows a stair sequence at low bitrate. The expression of the temporal increments in a wavelet basis does not really bring an impressive compression gain since the parts of the scene that appear in the following model has to be expressed in intra mode. Nevertheless our coder gives better results than H264 and Galpin scheme. With an H264 coder we can observe classical block effects due to wavelets block transforms. The use of non-uniform mesh, but also of an implicit morphing strongly contributes to the visual quality of the scene, avoiding block effects or ghost effects and jumps in connectivity and geometry showed by previous methods.

Our coder exploits temporal redundancy and inter-relations between camera positions, texture and 3D to get a IPP scheme [BGMG04]. In order to show the interest of the proposed predictive coding scheme we compare the results of our coder with a *naive encoder* where the single stream is obtained by concataining

camera positions, texture and 3D streams in Intra mode. Compression results for encoding the same information are given on Fig. 3 showing the profits for the bitrate by using our coder against Galpin and naive coders. Since texture information is prohominent over geometry the profits allowed by wavelets is particularly interesting for very low bitrates.

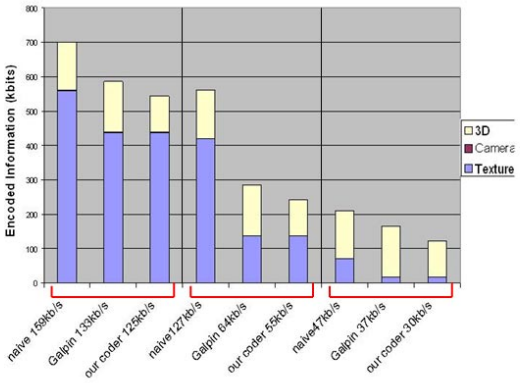

Fig. 3. Compression results for the same encoded information for Thabor sequence

Then Fig. 4 shows results during free navigation, i.e. when the viewer is not restricted to the camera path defined during capture. As previously, our coder shows significantly better visual quality thanks to the 3D morphing avoiding fading ghost effects.

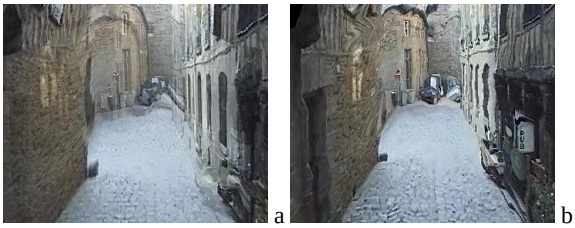

Fig. 4. "Street" sequence: Reconstruction on virtual path. With Galpin method (uniform mesh and fading) (a), with proposed method (non uniform mesh and implicit morphing) (b).

Finally Fig. 5 gives some examples of augmented reality applications.

7. CONCLUSIONS AND FUTURE WORK

We have presented a new format for model-based video coding with fine-grain scalability, allowing content adaptation over a wide spectrum of terminals and networks. In particular, 3D can be streamed adaptively in applications of free navigation over networks. The coder, showing better compression results and more scalability than previous schemes, exploits the power of second generation wavelets thanks to the design of an evolving model with a single connectivity mesh gathering every GOFs information.

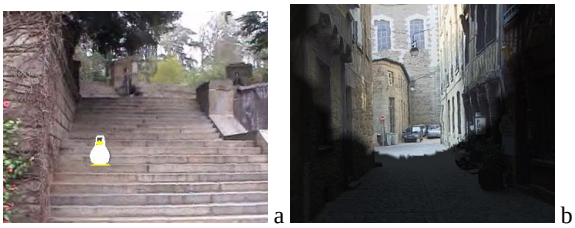

Fig. 5. Augmented reality images, addition of a virtual object (a) and change of illumination (b).

In despite of the good results already reached, some improvements still have to be performed, such as the derivation of a suitable error metric taking into account the geometric distortion.

8. REFERENCES

[BGMG04] R. Balter, P. Gioia, L. Morin, and F. Galpin. Scalable and efficient coding of 3d model extracted from a video. In *3DPTV, september*, 2004.

[Dau92] I. Daubechies. *Ten lectures on wavelets*. CBMS-NSF regional conf. series in appl. math., 1992.

[GBMD04] F. Galpin, R. Balter, L. Morin, and K. Deguchi. 3d models coding and morphing for efficient video compression. In *CVPR*, 2004.

[GM02] F. Galpin and L. Morin. Sliding adjustment for 3d video representation. *Eurasip Journal ASP*, 2002.

[HS88] C. Harris and M. Stephens. A combined corner and edge detector. In *Proc. 4th Alvey Vision Conf.*, 1988.

[KSS00] A. Khodakovsky, P. Schroder, and W. Sweldens. Progressive geometry compression. In *SIGGRAPH 2000 Conference Proceedings*, 2000.

[LDW97] M. Lounsbery, T. D. DeRose, and J. Warren. Multiresolution analysis for surfaces of arbitrary topological type. *ACM Transactions on Graphics*, 1997.

[MB95] L. McMillan and G. Bishop. Plenoptic modeling: an image based rendering system. In *Computer Graphics, Annual Conferences Series*, 1995.

[MG00] M. Magnor and B. Girod. Model-based coding of multi-viewpoint imagery. In *VCIP*, June 2000.

[Pol02] M. Pollefeys. Obtaining 3d models with a hand-held camera. In *Siggraph 2002, San Antonio*, July 2002.

[SW02] H. Schwarz and T. Wiegrand. The emerging jvt/h.26l video coding standard. In *IBC,*, 2002.

[TR98] Gabriel Taubin and Jarek Rossignac. Geometric compression through topological surgery. *ACM Trans. Graph.*, 17(2):84–115, 1998.

[ZFC99] Andrew Zisserman, Andrew Fitzgibbon, and Geoff Cross. Vhs to vrml: 3d graphical models from video sequences. In *Conference on Multimedia Computing and System*, 1999.