GREEDY NON-LINEAR APPROXIMATION OF THE PLENOPTIC FUNCTION FOR INTERACTIVE TRANSMISSION OF 3D SCENES

*Pietro Zanuttigh**, *Nicola Brusco**, *David Taubman*** *and Guido Cortelazzo**

* University of Padova, Italy
** The University of New South Wales, Sydney, Australia

ABSTRACT

We consider an interactive browsing environment, with greedy optimization of a current view, conditioned on the availability of previously transmitted information for other (possibly nearby) views, and subject to a transmission budget constraint. Texture information is available at a server in the form of scalably compressed images, corresponding to a multitude of original image views. Surface geometry is also represented at the server in a scalable fashion. At any point in the interactive browsing experience, the server must decide how to allocate transmission resources between the delivery of new elements from the various original view bit-streams and new elements from the geometry bit-stream. The proposed framework may be interpreted as a greedy strategy for non-linear approximation of the plenoptic function, since it considers both view sampling and rate-distortion criteria. We particularly elaborate upon a novel geometry- and distortion-sensitive strategy for blending the information available from different views at the client.

1. INTRODUCTION AND FRAMEWORK

This paper is concerned with the problem of efficient interactive retrieval and rendering of 3D scene information. We envisage a server and a client, connected via a bandlimited channel. At the client side, a user interactively determines the particular view of interest. An important feature of such applications is that the user can be expected to navigate between a variety of different views, although we do not know ahead of time which views will be of interest. We also do not know in advance how much time (transmission resources) the user will choose to devote to any particular view.

At one extreme, the user's interest may remain focused on a single view for a considerable period of time, waiting until very high quality imagery has been recovered before moving on. At this extreme, the interactive retrieval problem is tantamount to that of interactive image browsing, which is addressed most elegantly by progressive transmission of a single scalably compressed image, formed at the server. One way to achieve this is to combine the JPIP interactive imaging protocol with a JPEG2000 compressed representation of the view in question [1].

At the opposite extreme, the interactive user may select many different views in rapid succession, with the aim of understanding the scene's geometry. This phase might itself be a precursor to later detailed inspection of some particular view of interest. Since successive views are closely related, one natural way to improve the efficiency of the browsing experience is to predict each new view from the views which have already been transmitted, forming the same prediction at the server and client so that only the prediction residual need be transmitted. Explorations along this direction may be found in, e.g., [2, 3]. The predictive approach, however, suffers from a number of drawbacks. Firstly, the server must precisely replicate the steps used by the client to render each new view from existing previous views. Secondly, the server must compress the residual images corresponding to each change of view by the client. Perhaps most importantly, the predictive approach delivers a distinct approximate representation for each view requested by the interactive user, no matter how close those views may be to each other. It is difficult, if not impossible, to combine the information from several similar yet different views to synthesize a new, higher quality image at a later time. This limits the extent to which previously transmitted data can be leveraged in the future.

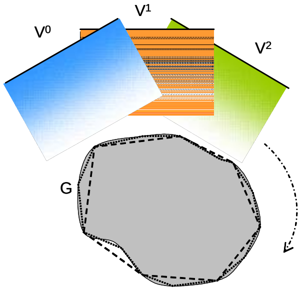

Fig. 1. *Overview of the browsing environment. The server has scalably compressed representations for a fixed set of "original view images,"* V^i *and the surface geometry,* G.

In view of the above arguments, we propose a framework for interactive scene browsing, in which the server delivers incremental contributions from two types of pre-existing data: 1) scalably compressed images of the scene from a collection of pre-defined views, V^i; and 2) a scalably compressed representation of the scene surface geometry, G. These elements are depicted in Fig. 1. We use the term "original view images" to distinguish the compressed server images V^i from new views rendered by the client. The server does not generate new views or compress differential imagery. Instead, it determines and sends appropriate elements from a fixed set of scalable compressed bit-streams, so as to provide its clients with the most appropriate data from which to render their desired views.

The proposed framework is particularly appropriate in view of the fact that 3D scene representations are usually generated from a collection of original 2D images; these are natural candidates for V^i. If the client happens to request one of the original view images, it can be incrementally served directly from its scalably compressed representation. Interestingly, though, this might not always be the best policy. If the client has already received suffi-

This work was supported in part by the FIRB-PRIMO project.

cient elements (sufficient quality) from one or more nearby original view images, V^{k_1}, V^{k_2}, ..., it may be more efficient to send only the geometric information required for the client to synthesize the requested view, using the resulting bandwidth savings to further augment the quality of these nearby original view images. It follows that even if the server has a huge number of original view images, an efficient service policy would effectively subsample them based on the interactive user's navigation patterns. More generally, the server may choose to send some elements from V^i, while expecting the client to derive other aspects of the view from the previously delivered, but less closely aligned original view images, V^{k_n}.

The proposed framework may thus be interpreted as fostering a greedy strategy for non-linear approximation of the plenoptic function, since it considers both view sub-sampling and rate-distortion criteria. The fact that efficient service policies can be expected to sub-sample the existing content automatically, brings the proposed approach into contrast with the predictive approach mentioned previously, where imagery is delivered for every view requested by the user. The system outlined above gives rise to the following interesting questions:

1) How should the client combine information from available original view images into a new view of interest, using an available description of the surface geometry?

2) How should the server distribute available transmission resources amongst the various original view images and the geometry information which the client may need to render a new view? Included in this question is that of whether the server should transmit elements from an entirely new original view image which is more closely aligned with the requested view, rather than refining nearby original view images for which the client already has more data.

Within the scope of this present paper, it is not possible to explore both of these questions in detail. Instead, we focus our attention on the first, since answers to the second question depend on how the server expects the client to use the information which it has. Section 2 develops our proposed approach to distortion-sensitive rendering at the client, while Section 3 provides some preliminary experimental evidence to validate this approach. As for the complete system, Section 4 outlines some directions which we are currently investigating and identifies related existing methodologies.

2. DISTORTION-SENSITIVE VIEW RENDERING

2.1. Rendering from a Single View

Let V^* denote a desired view. In this section, we briefly discuss the process of rendering V^* from a single original view image, V^i. We assume a triangular mesh representation for the surface geometry G, projecting its nodes onto the image planes corresponding to V^* and V^i. Isometric or perspective projections might be employed, for example. Let Δ_n^* and Δ_n^i denote corresponding triangular patches of the two projected meshes, as illustrated in Fig. 2. Of course some of the projected triangles may be hidden in one image, but not the other. If Δ_n^* is hidden, then Δ_n^i is not involved in rendering, while if Δ_n^i is hidden, Δ_n^* is a "hole" in V^*, which cannot be rendered from V^i. We can avoid the possibility of partially hidden triangles by suitable remeshing of the available geometry in the vicinity of hidden surfaces.

Apart from the holes, each exposed triangle Δ_n^* is rendered by affine warping of Δ_n^i. We write this as $V^* = \mathcal{W}_i\left(V_i\right)$ or, over the domain of each triangle, as $\Delta_n^* = \mathcal{W}_i\left(\Delta_n^i\right)$. Admittedly, affine warping does not exactly extend the behaviour of a perspective imaging model into the interior of the projected surface triangles. However, this error can be rendered arbitrarily small by reducing the size of the surface mesh elements.

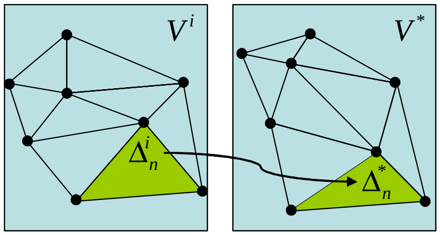

Fig. 2. *Surface mesh projected onto desired view, V^*, and an original view image, V^i.*

2.2. Combining Multiple Views

One way to combine the information from multiple original view images, V^{i_0}, V^{i_1}, ..., is to simply average the results obtained by mapping each of them onto the desired view. Unfortunately, any imperfections in the surface geometry description produce misalignment amongst the separate renderings, $\mathcal{W}_{i_k}\left(\Delta_n^{i_k}\right)$, so that averaging tends to blur high frequency spatial features. Also, the simple average shows no preference for one possible rendering over another.

An alternative strategy is to select a single "most appropriate" original view image from which to render each triangle. We refer to this as "stitching" and write $\Delta_n^* = \mathcal{W}_{i_{k'}}\left(\Delta_n^{i_{k'}}\right)$, where $i_{k'}$ is the "best stitching source" for the n^{th} triangle. The problem, of course, is identifying this best stitching source; this is the subject of the next two sub-sections. Although stitching avoids the blurring problem, it tends to produce visible discontinuities at the boundaries between adjacent triangles which are rendered from different original source views. This is because the surface geometry will inevitably contain modeling errors, and the rendering process described here does not account for illuminant-dependent shading effects.

To hide these artifacts, and also open the door to resolution-dependent stitching algorithms, we prefer to form V^* in the DWT (discrete wavelet transform) domain. Each possible rendering, $\mathcal{W}_{i_k}\left(V^{i_k}\right)$, is generated in the image domain and separately subjected to DWT analysis. This produces a collection of subbands, $\text{LL}_D^{i_k}$ and $\text{LH}_d^{i_k}$, $\text{HL}_d^{i_k}$, $\text{LH}_d^{i_k}$ for $d = 1, 2, \ldots, D$, where D is the number of DWT decomposition levels. Stitching is carried out within the individual subbands to produce LL_D^* and LH_d^*, HL_d^*, LH_d^*, from which V_* is recovered by DWT synthesis. Section 3 provides visible evidence for the benefits of this DWT-based stitching approach.

2.3. Incorporating Distortion Information

In this section, we describe a method for selecting the best stitching source $i_{k'}$, for each triangle Δ_n^*, based solely on the amount of quantization error power which the selection will incur. We thus ignore the limitations of geometric modeling, which are the subject of the next section. We assume that the original view images were compressed in the DWT domain (e.g., using JPEG2000). The quantization error associated with a single sample in subband b of V^i, falling within the scope of triangle Δ_n^i, finds its way into subband b^* of $\mathcal{W}_i\left(V^i\right)$ through DWT synthesis, warping and further DWT analysis. Let $\mathbf{s}_b$ denote the relevant subband synthesis vector, and let $\mathcal{A}_{b^*}$ denote the subband analysis operator which produces subband b^* from an input image. Then the squared error in our original subband sample should be scaled by $W^{i,n}_{b\to b^*} = \|\mathcal{A}_{b^*}(\mathcal{T}_{i,n}(\mathbf{s}_b))\|^2$ to determine its contribution to the total squared error in subband b^* of $\mathcal{W}_i\left(V^i\right)$. Here, $\mathcal{T}_{i,n}$ is the affine operator associated with $\mathcal{W}_i$ within triangle Δ^i_n. Assuming uncorrelated quantization errors, or orthogonal basis vectors, the total quantization distortion appearing within the region defined by Δ^*_n in subband b^* of $\mathcal{W}_i\left(V^i\right)$ can be approximated[1] by

$$D^*_{i,n,b^*} = \sum_b W^{i,n}_{b\to b^*} D_{i,n,b} \qquad (1)$$

where $D_{i,n,b}$ is the total squared quantization error appearing within the region defined by Δ^i_n in subband b of the original view image, V^i. When considering quantization distortion alone, the best stitching source i_k, is the one for which $D^*_{i_k,n,b^*}$ is smallest; this best stitching source could potentially differ from subband to subband.

It is worth noting that the affine operator $\mathcal{T}_{i,n}$ serves to stretch the synthesis basis function $\mathbf{s}_b$ by an amount $|\Delta^*_n| / \left|\Delta^i_n\right|$, amplifying its energy by roughly the same amount. Assuming an orthonormal transform[2], we can say that $\|\mathbf{s}_b\| = 1$, $\|\mathcal{T}_{i,n}(\mathbf{s}_b)\|^2 = |\Delta^*_n| / \left|\Delta^i_n\right|$ and hence $\sum_{b^*} W^{i,n}_{b\to b^*} = |\Delta^*_n| / \left|\Delta^i_n\right|$, so that

$$\sum_{b^*} D^*_{i,n,b^*} = |\Delta^*_n| / \left|\Delta^i_n\right| \cdot \sum_b D_{i,n,b} \qquad (2)$$

At first glance, this would appear to suggest that the total distortion in the warped triangle (left hand side) should be roughly independent of the affine operator $\mathcal{T}_{i,n}$, since the total distortion in the source triangle $\sum_b D_{i,n,b}$, should be roughly proportional to its area, $\left|\Delta^i_n\right|$. However, two things are missing from this picture. Firstly, $\mathcal{T}_{i,n}$ must be a bandlimited warping operator, so that $\|\mathcal{T}_{i,n}(\mathbf{s}_b)\|^2 = |\Delta^*_n| / \left|\Delta^i_n\right| \cdot F\left(\mathcal{T}_{i,n}(\mathbf{s}_b)\right)$, where $F\left(\mathcal{T}_{i,n}(\mathbf{s}_b)\right)$ is the fraction of the energy in $\mathcal{T}_{i,n}(\mathbf{s}_b)$ which falls within the Nyquist sampling limit. Second, expansive operators $\mathcal{T}_{i,n}$ cannot predict the highest frequency details of V^* at all. Both of these effects can be taken into account by extending the sums in (1) and (2) to include subbands from a set of hypothetical resolutions above those of the original images. The source distortions $D_{i,n,b}$ for these missing subbands are equal to their energies $E_{i,n,b}$, which we estimate by projecting each source image onto the other in turn and taking the maximum of the energy produced by such projections. The target distortions D_{i,n,b^*} of hypothetical subbands b^* represent an unfelt contribution to the left hand side of (2), which grows with $\left|\Delta^i_n\right| / |\Delta^*_n|$. As a result, we expect to find that the best stitching source is that for which $\mathcal{T}_{i,n}$ is most contractive (i.e., $\left|\Delta^i_n\right| / |\Delta^*_n|$ is maximum), all other things being equal.

A direct application of (1) would require the client to know the values, $D_{i,n,b}$; this is not generally possible, since it has access only to the compressed imagery. Suppose, however, that JPEG2000 has been used to compress the original view images. Then the client has received some number of quality layers from the embedded bit-stream(s) associated with the code-block(s) containing triangle Δ^i_n in subband b of V^i. It has available the compressed bit-rates, the effective quantization step sizes and, with a small amount of global side information, the rate-distortion slope thresholds associated with each quality layer. The client may fit this information to a parametric model of the subband statistics in order to form reasonable estimates for the distortion values, $D_{i,n,b}$. Unfortunately, the distortion cannot be estimated in this way within code-blocks for which no bits have yet been delivered. In this case, $D_{i,n,b}$ should be taken as $E_{i,n,b}$, where the energy $E_{i,n,b}$ is estimated in the manner described above.

2.4. Accounting for Geometric Modeling Errors

As noted above, if quantization error alone is used to determine the best stitching source, the source selected for the n^{th} triangle will tend to be that for which $\left|\Delta^i_n\right|$ is largest. This is the original view image whose focal plane is most parallel to the corresponding 3D surface triangle. While this policy makes intuitive sense, if the geometric model were highly unreliable, we would expect to do better by selecting the original view image which is most closely aligned with the desired view; this is the one which for which the rendering process is least dependent on accurate knowledge of the geometry.

We identify here two aspects of modeling error which are worth capturing. Firstly, uncertainty in the surface geometry translates into uncertainty in the parameters of the affine transformations, $\mathcal{T}_{i,n}$. This, in turn, represents a translational uncertainty, which has been studied previously in [4]. Its effect may be modeled by augmenting each term $D_{i,n,b}$ in (1) by a second contribution of the form $\sigma^2_G \phi^2_{i,n} |\omega_b|^2 E_{i,n,b}$. Here, ω_b is "representative" of the spatial frequencies belonging to subband b, $E_{i,n,b}$ is the estimated subband energy defined above, σ^2_G reflects uncertainty (MSE) in the surface node positions, and $\phi_{i,n}$ represents the sensitivity of 2D mesh node positions to displacements in the original 3D surface nodes.

Since our surface model does not account for the illuminant-dependent effects of shading and reflection, we can expect a second distortion contribution which grows with the deviation between the orientation of views V^* and V^i. Ignoring specularity, we expect this distortion term to be proportional to the signal power, suggesting the following augmented version of equation (1).

$$D^*_{i,n,b^*}=\sum_b W^{i,n}_{b\to b^*}\Big(D_{i,n,b}+\Big[\sigma^2_G\phi^2_{i,n}|\omega_b|^2+g(\langle\mathbf{n}^i,\mathbf{n}^*\rangle)\Big]\,E_{i,n,b}\Big)$$

Here, $\mathbf{n}^i$ and $\mathbf{n}^*$ are the surface normals and, in the absence of careful modeling, $g\left(x\right)$ is set to $\gamma \tan\left(\cos^{-1} x\right)$, where γ determines the value we place on illumination fidelity. We expect the server to provide an indicative value for the model uncertainty, σ^2_G.

3. RENDERING EXPERIMENTS

In this section we provide some preliminary experimental results to justify the methodology presented above. We construct a complete 3D surface model from 60 original view images, selecting V^1 and V^6 to reconstruct V^*, which happens to align with V^3. Fig. 3 displays the effect of three merging strategies. As expected, simple averaging blurs the features. DWT-based stitching preserves the details, while hiding the discontinuities which appear when stitching in the image domain. In this case the best stitching source (upper left in figure) is taken as the one which minimizes the ratio $|\Delta^*_n| / \left|\Delta^i_n\right|$.

Fig. 4 demonstrates the benefits of our proposed distortion-based stitching procedure. In this case, V^1 is compressed more heavily than V^6, which changes the best stitching source (upper

[1] This is an approximation only because we assign all of the incurred distortion to Δ^i_n ignoring the fact that the error signal is smeered out by the overlapping DWT bassis vectors.

[2] The 9/7 biorthogonal wavelet transform used for our experiments is very nearly orthonormal, subject to appropriate normalization of the subband samples.

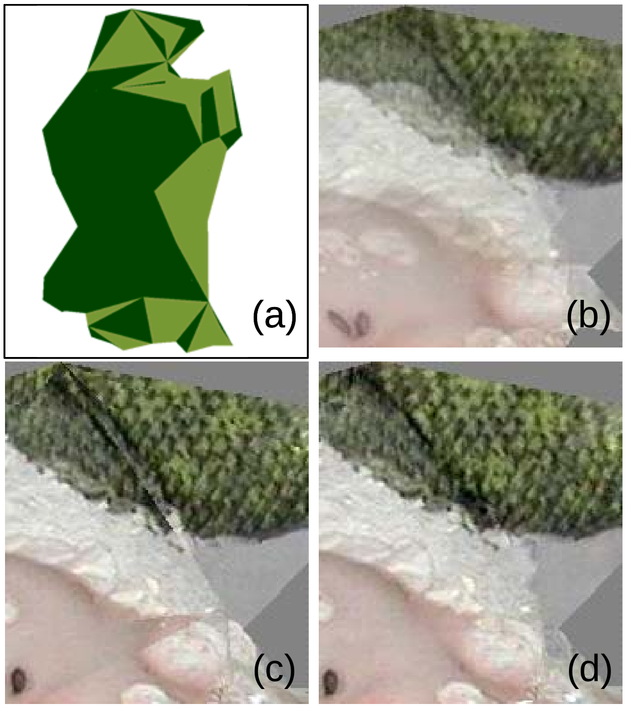

Fig. 3. *Merging uncompressed views: a) best stitching map; b) naive averaging; c) image domain stitching; d) DWT stitching.*

left in figure). Triangles from V^1 are still selected, but less often. From the figure, we see that this stitching map yields better results than that obtained with the distortion insensitive map of Fig. 3 or that obtained using V^6 alone. In these experiments, distortion alone is used to determine the best stitching source. We have yet to include the effects of geometry mismatch from Section 2.4.

4. TOWARD OPTIMAL SERVICE POLICIES

In this section, we return briefly to the high level perspective from which we started in Section 1, to consider other aspects of the interactive browsing framework developed there. For scalable coding of the surface mesh G, a variety of solutions have already been proposed, e.g. [5], while JPEG2000 represents an excellent choice for the scalable compression of each V^i. The recent JPIP standard [1] provides a vehicle for efficient dissemination of incremental contributions from the scalable compressed bit-streams, allowing for server control of the transmission sequence; it can be extended to incorporate the elements from scalable geometry models.

The key missing ingredient is a rate-distortion based framework within which the server can decide which new compressed texture and geometry elements should be transmitted in order to maximize the client's rendered imagery, subject to transmission constraints. Many of the elements required to realize an appropriate service policy are already present in the distortion formulations of Section 2, which can be replicated or approximated by the server. In particular, this equation tells us which image the client will select as its stitching source for each triangle, identifying the impact of texture distortion $D_{i,n,b}$ and geometry distortion σ_G^2 on the quality of the reconstructed view. Service policies themselves, however, are beyond the scope of this present paper.

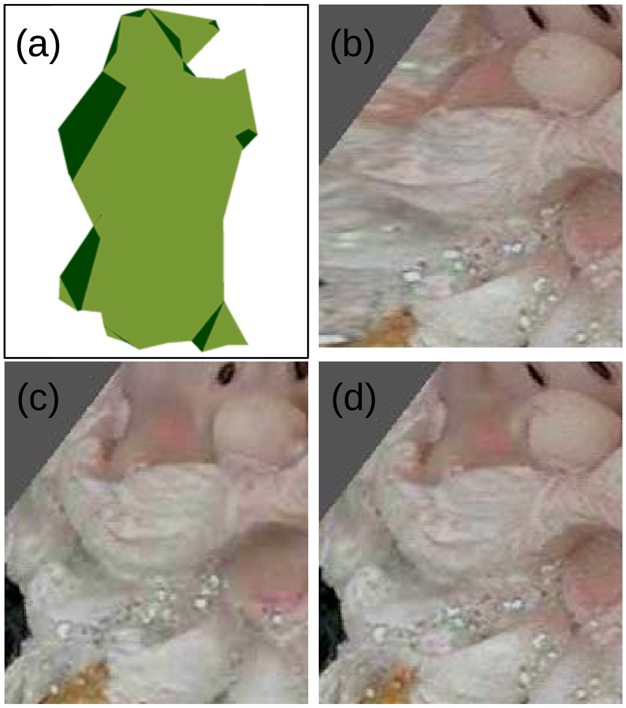

Fig. 4. *Merging compressed views: a) distortion sensitive stitching map; b) single (least distorted) view reconstruction; c) distortion insensitive DWT stitching; d) DWT stitching based on map in (a).*

5. CONCLUSIONS

This paper represents a first step toward a completely novel approach to the interactive dissemination of compressed 3D scenes. The framework presented here also draws attention to a variety of important problems such as the optimal distribution of compressed bits between texture and geometry information, and non-linear approximation (not just sub-sampling) of the plenoptic function. As a convincing start in this direction, we have described mechanisms for estimating the distortion associated with rendering an intended view from a variety of compressed images with uncertain geometry. We have also experimentally validated a client-side rendering algorithm which aims to minimize this distortion.

6. REFERENCES

[1] D. Taubman and R. Prandolini, "Architecture, philosophy and performance of jpip: internet protocol standard for JPEG 2000," *Int. Symp. Visual Comm. and Image Proc.*, vol. 5150, pp. 649–663, July 2003.

[2] M. Levoy, "Polygon-assisted JPEG and MPEG compression of synthetic images," in *Proc. SIGGRAPH*, vol. 3, pp. 21–28, Aug 1995.

[3] D. Cohen-Or, "Model-based view-extrapolation for interactive VR web systems," in *Proc. Computer Graphics International*, pp. 104–112, Jun 1997.

[4] A. Secker and D. Taubman, "Highly scalable video compression with scalable motion coding," *IEEE Trans. Image Proc.*, (to appear) 2004.

[5] A. Khodakovsky, P. Schroder, and W. Sweldens, "Progressive geometry compression," in *Proc. SIGGRAPH*, pp. 271–278, 2000.