A Model-Based Approach for Coding 3D Head Sequences
L. Granai, T. Vlachos, J. R. Tena, T. Davies

I. INTRODUCTION

Recently, the interest in 3D imaging has increased tremendously and alongside 3DTV has developed quickly both from a research and commercial point of view. The 3D video processing chain includes acquisition, representation, compression and display. We focus our attention on the compression step, introducing a novel technique for model-based 3D video coding. The method presented here is not an extension of a 2D video coding scheme to match the higher dimensionality of the data, but a fully three-dimensional process which uses 3D models in order to understand the actual geometry of a scene and to provide efficient representation and compression. This work aims at the compression of 3D head and head-and-shoulders sequences, based on a 3D model. Figure 1 shows the generic coding scheme adopted to represent and compress a 3D frame of a 3D video sequence. This can be divided into two sequential stages: first the extraction of 3D information, then compression.

In the first phase, the \( n \)-th frame is analyzed and registered using a 3D face model, which is fixed and known also at the decoder side. This is done in the “Registration” block, whose output is a set of parameters which describe the shape of the face with respect to the model and an image which contains the texture information. The parameters “Param” are in fact a list of \( N \) 3D points that control the model. “Texture” is an image containing the flattened texture of the face. In the second phase the output of the “Registration” block is compressed. Shape parameters are differentiated with respect to the parameters relative to the previous frame (“Param \( n-1 \)”), transformed, quantized and coded. Texture images are also compressed reducing their spatial and temporal redundancy.

Fig. 1. Generic scheme of the proposed coding technique.

II. REGISTRATION

The technique we use here to register each 3D frame is based on the face dense registration algorithm proposed in [1] and motivated by the approach of Mao et al. [2]. It establishes dense correspondences between a generic face model (MG) and a 3D image containing a face (referred to as “MD”). The algorithm has three distinct stages: i) global mapping, ii) local matching and iii) energy minimisation. A set of 3D landmarks is required as initialisation. During the global mapping stage, MG and MD landmarks are brought into exact alignment using the thin-plate spline interpolation technique, which smoothly deforms MG minimising the bending energy. In the second stage, for each vertex on MG the closest vertex on MD, within an adaptive search radius, is found. In the final stage MG is conformed to MD by minimizing a weighted sum of internal and external energy. The fitting error can be further reduced by following an iterative coarse-to-fine approach during the local matching and energy minimisation stages. Further details and evaluation results can be found in [1].

The registration algorithm also produces a texture image which is generated by using the correspondence between shape and texture in the input 3D face scan. As this texture corresponds to the generic model, the texture map is the same for all input faces.

L. Granai, T. Vlachos and J. R. Tena are with the Centre for Vision, Speech and Signal Processing, University of Surrey, UK
T. Davies is with BBC Research, UK
E-mail: L.Granai@surrey.ac.uk
III. COMPRESSION

Once the 3D information has been extracted from the input frame, one has to reduce the temporal and spatial redundancy of both shape and texture data. This is achieved by taking into account the previous or neighbouring frames (see Fig. 1).

The parameters describing the face shape are a set of $N = 845$ ordered points in the 3D space. The order in which the points are visited is fixed by the model. In order to enhance data predictability, we have introduced a permutation so that the global length of the line that joins all the points is minimised. Basically it is equivalent to solving an open traveling salesman problem (TSP) in 3D, (i.e. a TSP without coming back to the starting point). Note that the permutation is fixed and based on the model only and therefore can be assumed to be known to the decoder side, together with the model. Otherwise it has to be sent only once and this has a negligible impact on the global bit-load.

After this permutation, the coordinates of the points are differentiated with respect to the parameters relative to the previous frame. The differential values are then transformed using a Discrete Wavelet Transform. Each axis in the 3D space is treated independently, originating three 1D series of coefficients. The low-pass sub-bands are coded in a differential way (DPCM) and uniformly quantized. Every other sub-band of every axis is quantized using a uniform dead-zone quantization. Adaptive Arithmetic Coding is used to reduce the entropy of quantized wavelet coefficients.

The “Registration” block in Fig. 1 produces a colour image that contains the flattened texture of the face. Obviously, all the texture images extracted from a sequence present a high temporal redundancy. A first idea is to transmit only the image relative to the I-frame and use it also for all the others. This brings an enormous advantage in terms of bit-rate, but it can cause visually very annoying artifacts in the area of the face where the registration is not perfect. More in general, one can think about refreshing the texture image whenever necessary, or coding the sequence of textures with a standard technique designed for classical 2D video.

IV. RESULTS AND CONCLUSIONS

We tested our algorithm on sequences of a person is speaking and slightly moving the head at a frame-rate of 25 frames/sec. The error for the shape is measured by considering the mean geometric distance between original and reconstructed points. The shape information can be compressed at a high visual quality with an average bit-rate about 9 Kbit/sec. The texture information is coded by means of the Dirac video codec (see [3]). The compression strongly benefits form the very high temporal redundancy and a PSNR of around 40 dB can be achieved at 130 Kbit/sec. Further results can be found in [4] together with a more detailed description of the algorithm.

Finally we would like to emphasise that the proposed scheme would also work with other 3D models. This means that a better 3D face model can be used to perform the registration or that one can code other sequences by using a different model, e.g. upper body.

REFERENCES