ON HUMAN TIME-VARYING MESH COMPRESSION EXPLOITING ACTIVITY-RELATED CHARACTERISTICS

Alexandros Doumanoglou, Dimitrios Alexiadis, Stylianos Asteriadis, Dimitrios Zarpalas, Petros Daras, Senior Member, IEEE

Information Technologies Institute, Centre for Research and Technology Hellas
6th km Charilaou - Thermi, GR-57001, Thessaloniki, Greece

ABSTRACT

In this work, we explore the potential of exploiting activity-related global features in order to improve the performance of an existing human Time-Varying Mesh (TVM) compression scheme. The TVM compression scheme used, employs two kinds of frames, namely Intra(I)-Fames and Enhanced Predicted(EP) Frames. In this scheme, I-Frames are used as a reference to encode EP-Frames. The paper introduces a strategy for selecting the most appropriate I-Frame that will serve as a reference frame for the encoding of EP-Frames, exploiting activity-related characteristics. Two different strategies are presented, using a skeleton-matching criterion and a periodicity measurement metric based on human skeleton. Evaluation is conducted on two sequences of the MPEG-3DGC database [1]. Results show that the concept is sound, but they also reveal the sensitivity of the proposed methods to the skeleton quality, thus the need for more robust skeleton tracking techniques.

Index Terms— time-varying mesh, compression, skeleton matching, periodicity estimation

1. INTRODUCTION

Real-time transmission of full 3D realistic representations of moving humans is a challenging issue, due to the vast amount of information they entail. Although technology allows for increasingly large bandwidths, still, transmitting 3D meshes in real-time necessitates robust compression algorithms. When those meshes originate from sensing real environments, such as when using a low-cost depth camera like Microsoft Kinect, they constitute Time Varying Meshes (TVMs) with variable geometry, as well as connectivity across frames. In most research works, motion features are usually taken into account on a local, frame-based level, thus ignoring the global character of human expressivity. The proposed work examines the role of activity-related features in human TVM compression. More specifically, the role of key-frame extraction with subsequent pose-dependent I-frame matching, along with that of a qualitative expressivity feature (here, periodicity) are examined.

The most notable work in TVM compression can be represented by [2], [3] and [4]. In [2], Han et al. extend the concept of Block Matching, known from traditional 2D Video, in 3D Space for motion compensation. In [3], the same authors used coarse and fine levels of quantization to compress TVMs and were able to note an improvement. Finally, in [4] Yamasaki et al, proposed a patch-based compression of TVMs showing a further improvement over the previous attempts.

Methods using human motion qualitative features as context-related knowledge for compression constitute an area not yet widely explored by the research community. When considering 2D video, robust extraction of human activity-related features leading to effective compression involves a series of steps (detection, temporal and spatial segmentation, tracking), prone to error and noise accumulation. In a typical example of such work [5], a face tracker is employed for the creation of a basis sequence consisting of face-centered images, with incoming frames compressed through mapping on this sequence.

With the advent of depth sensors, however, robust, three-dimensional data processing provides mechanisms for easy human silhouette segmentation and further use in tele immersive environments [6]. Activity-related knowledge and compression can be achieved through a single modality (point cloud), while novel compression schemes can take advantage of human silhouette segmentation. Chen et al. [7] propose an activity-aware 3D mesh compression architecture, using Microsoft Kinect depth sensors. However, activity is retrieved with the help of hand-held sensors, which inform the producer regarding a tolerable number of frames to be discarded. In the hereby proposed scheme, Microsoft Kinect depth sensor is used for 3D data acquisition and activity recognition. The potentiality of extracting activity-related, global features is examined on two use-cases (skiing and jogging) with highly expressive characteristics. The aim of the paper is to explore the potential of exploiting those activity-related global features in order to improve the performance of human TVM compression.

This work was supported by the EU funded project 3DLIVE, GA 318483. http://3dliveproject.eu/

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

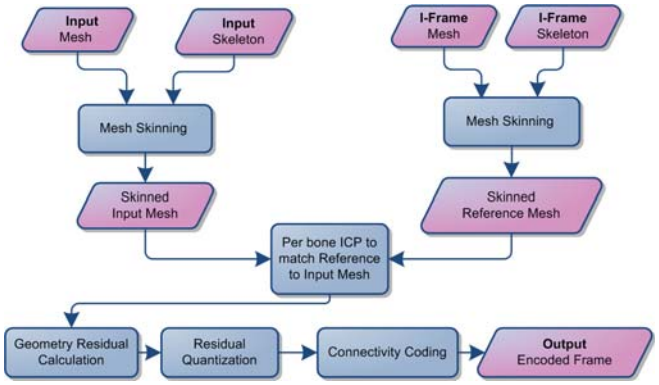

Fig. 1. Encoding of EP-Frames

2. TVM ENCODER ARCHITECTURE

In this section an overview of the TVM encoder architecture, that this work was based on, is given. Similar to many video coders, the hereby TVM encoding scheme considers two kinds of frames, Intra(I)-Frames and Enhanced Predicted(EP) frames. The encoding of I-Frames is made using an existing state-of-the-art static mesh coder, like [8] and are decoded independently of any previous frames. I-Frames are generated at fixed, predefined intervals. On the other hand, EP-Frames are encoded with respect to some previously encoded I-Frame that is used as a reference. In Fig 1, the encoding process of EP-Frames is depicted. In order to encode EP-Frames, the proposed encoder makes use of the human skeleton, obtained using recent advances in skeleton tracking technology, like [9]. An automated skinning process is used to assign mesh vertices to skeleton bones for both the input mesh as well as the reference mesh obtained from a previous I-Frame. Then, per bone ICP (Iterative Closest Points) is utilized to align the reference mesh to the input mesh and thus minimizing the geometry (vertex positions) prediction errors. The geometry prediction residuals are then quantized (using the Entropy-Constrained Vector Quantization method) and entropy coded (using an Arithmetic Coder), along with the mesh connectivity information.

In this work, we aim to introduce a strategy in selecting the most appropriate I-Frame that will serve as a reference frame for the encoding of EP-Frames. The potential of using activity-related information to develop this strategy is explored.

3. SELECTION OF REFERENCE I-FRAME

The most straightforward approach to encode an EP-Frame, is to use the last encoded I-Frame of the sequence as a reference. However, in a periodic human motion scenario, it is possible that a previous I-Frame matches better with the current frame, depending on the periodicity of the motion.

Therefore, based on the above facts, a buffer in the codec is used to hold the last N_I I-Frames in memory. The encoder has to decide which of these is more appropriate to be used as a reference, with the potential to minimize the prediction residuals. We experimented with two different approaches for detecting the most appropriate reference I-Frame, a skeleton matching-based and a periodicity detection based, as described below.

3.1. Based on a skeleton matching score

In this approach, the encoder searches for the best I-Frame based on the skeleton "similarity" (skeleton similarity was chosen here, as an alternative to slow comparisons among bulky 3D meshes) between the current frame t and the candidate I-Frames. Therefore, we define a skeleton matching score, as follows.

Skeleton matching score: Let $c_{\mathrm{p}}^{\mathcal{I}} \in [0,1]$ denote the tracking confidence of the bone's root joint position for the candidate reference I-Frame and c_{p}^{t} the corresponding tracking confidence for frame t. The orientation tracking confidences are similarly denoted as c_{o}^{I} and c_{o}^{t}, respectively. Then, the confidences for the I-t pair are defined as:

$$c_{\mathrm{p}} = c_{\mathrm{p}}^{\mathcal{I}} c_{\mathrm{p}}^{t}, \quad c_{\mathrm{o}} = c_{\mathrm{o}}^{\mathcal{I}} c_{\mathrm{o}}^{t}. \tag{1}$$

Let also d_{p} denote the distance (measured in mm) between the I-Frame bone's center and the t-frame bone's center. Similarly, d_{o} denotes the "angular" distance (measured in degrees) between the bone's orientations in the two frames.

Now, we introduce a position-based (per-bone) dissimilarity metric, which takes into account both the position distance $d_{\mathrm{p}}(i)$ (i indexes the bones), as well as the position tracking confidences:

$$D_{\mathrm{p}}(i) = (1 - W_{\mathrm{c}}) d_{\mathrm{p}}(i) + W_{\mathrm{c}} \Big[c_{\mathrm{p}}(i) d_{\mathrm{p}}(i) + \big(1 - c_{\mathrm{p}}(i)\big) K_{\mathrm{p}} \Big], \tag{2}$$

where K_{p} is a penalty constant to be applied when skeleton confidence is low, while $W_{\mathrm{c}} \in [0,1]$ is a weight to balance our trust to the skeleton confidence output of skeleton tracking. Notice that for $W_{\mathrm{c}} = 0$, only the measured distance $d_{\mathrm{p}}(i)$ is taken into account in the dissimilarity metric. On the other hand, for $W_{\mathrm{c}} = 1$ the dissimilarity metric depends on the tracking confidence: when tracking is confident ($c_{\mathrm{p}} = 1$) the dissimilarity metric becomes again equal to the measured distance $d_{\mathrm{p}}(i)$. On the other hand, unconfident tracking results into assigning a dissimilarity value equal to the penalty constant K_{p}.

An orientation-based (per-bone) dissimilarity metric is also introduced:

$$D_{\mathrm{o}}(i) = (1 - W_{\mathrm{c}}) d_{\mathrm{o}}(i) + W_{\mathrm{c}} \Big[c_{\mathrm{o}}(i) d_{\mathrm{o}}(i) + \big(1 - c_{\mathrm{o}}(i)\big) K_{\mathrm{o}} \Big], \tag{3}$$

where K_{o} is penalty constant similar to K_{p}. The corresponding whole-skeleton position-based metrics are defined as:

$$\overline{D_{\mathrm{p}}} = \frac{1}{BK_{\mathrm{p}}} \sum_i D_{\mathrm{p}}(i), \quad \overline{D_{\mathrm{o}}} = \frac{1}{BK_{\mathrm{o}}} \sum_i D_{\mathrm{o}}(i), \quad (4)$$

with B denoting the number of bones in the skeletons. Finally, the skeleton matching score is given by:

$$S = W_{\mathrm{p}} \max\left((1 - \overline{D_{\mathrm{p}}}), 0\right) + W_{\mathrm{o}} \max\left((1 - \overline{D_{\mathrm{o}}}), 0\right), \quad (5)$$

where the two weights W_{p} and W_{o} sum up to unity. The defined matching score lies in the interval $[0, 1]$.

At this point it is important to note that there can be two different strategies to apply the presented approach on skeleton matching. The first one, is to use the final skeleton matching score in order to select one I-Frame as a reference for the EP-Frame at t. The second, is to apply a per bone match score in order to select multiple I-Frames, one for each bone. Then, when encoding an EP-Frame, the mesh at t is encoded with respect to multiple I-Frames, one for each bone. Similar to the above discussion, the per-bone score is defined as: $S(i) = W_{\mathrm{p}} \max\left((1 - \overline{D_{\mathrm{p}}(i)}), 0\right) + W_{\mathrm{o}} \max\left((1 - \overline{D_{\mathrm{o}}(i)}), 0\right)$, where $\overline{D_{\mathrm{p}}(i)} = \frac{1}{K_{\mathrm{p}}} D_{\mathrm{p}}(i)$ and $\overline{D_{\mathrm{o}}(i)} = \frac{1}{K_{\mathrm{o}}} D_{\mathrm{o}}(i)$.

3.2. Based on periodicity

Additionally to the last N_I I-Frames, the encoder keeps in a memory buffer the skeleton information for the last N frames. Let $\mathbf{p}(i;t)$ denote the center of the i-th bone, in the time sample t (with respect to $t = 0$ in the buffer). Then, the L2-norm $r(i;t) = ||\mathbf{p}(i;t)||_2$ (distance from the global coordinates center) is considered, in order to opine about the periodicity of the bone's motion in the last N frames. This is achieved by taking the Discrete Fourier Transform (DFT) of $r(i;t)$. Let this be denoted as $R(i;k)$, where k stands for the k-th frequency-sample. Finding $k^{\mathrm{m}} = \mathrm{argmax}\{|R(i;k)|\}$, provides the dominant period of the bone's motion. Given that the samples of DFT are N, the period is given by $\hat{T} = N/k^{\mathrm{m}}$.

Notice that we ignore the mean value (DC component) of $r(i;t)$ ($k = 0$ in $R(i;k)$) and thus the method is invariant to the selected global world center. Additionally, it is rotation invariant, since we use the L2-norm of the bone's position. Finally, it is scale-invariant, since the position k^{m} of the maximum DFT modulus is independent to any scaling of $r(i;t)$.

In case the estimated period is $\hat{T}$, then the selected reference I-frame is the one closest to $N - \hat{T}$. It should be highlighted that if the energy is concentrated in very low frequency components ($k \in [0, N/30]$), this means that the motion is not periodic (the period is very large) and, therefore, the last I-frame is selected as reference. Moreover, this strategy works on a per-bone basis, conceptually similar to the discussion in 3.1, by selecting multiple reference I-Frames when encoding an EP-Frame, one for each bone.

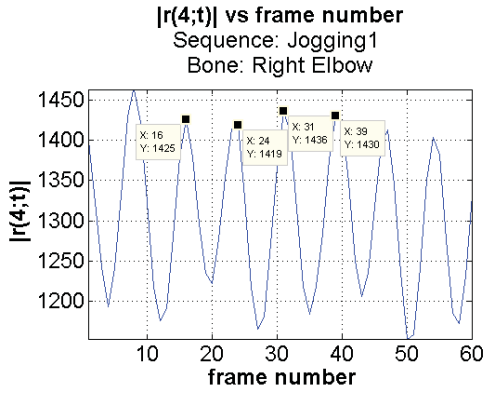

Fig. 2. Periodicity Metric in "Jogging" sequence for the elbow bone.

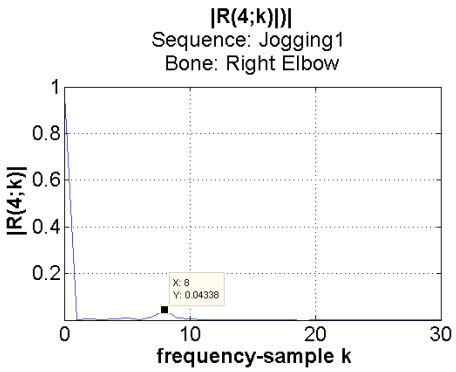

Fig. 3. Fourier Transform of the Periodicity Metric in "Jogging" sequence for the elbow bone.

4. EXPERIMENTAL RESULTS

In this section we provide experimental results of the presented algorithms. The two sequences used for evaluation are human TVMs (publicly available at `http://vcl.iti.gr/reconstruction`) that are considered part of the official MPEG-3DGC database [1]. The two TVM sequences, namely "Skiing3-PoissonLow" and "Jogging1-PoissonLow" ("Skiing" and "Jogging" for short), were captured using our multi-sensor Microsoft Kinect setup [6] and were reconstructed using the Poisson method [10]. The "Skiing" sequence consists of 189 frames with 50000 vertices per frame on average, while the "Jogging" sequence consists of 230 frames of the same average vertex count per frame. The skeleton tracking algorithm used during the evaluation was PrimeSense's NITE [11].

4.1. Evaluating the Periodicity Metric

In Fig. 2 the metric to estimate the periodicity of the right elbow (hand) bone is depicted when considering the "Jogging" activity. As expected, in a jogging scenario, both hands do a

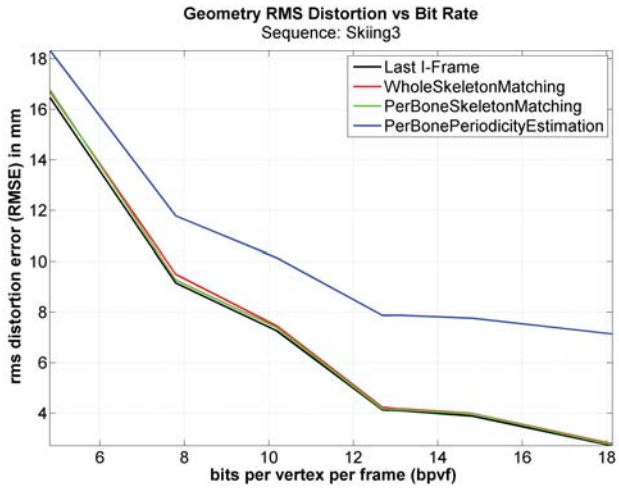

Fig. 4. Bit-rate vs Distortion for "Skiing" sequence

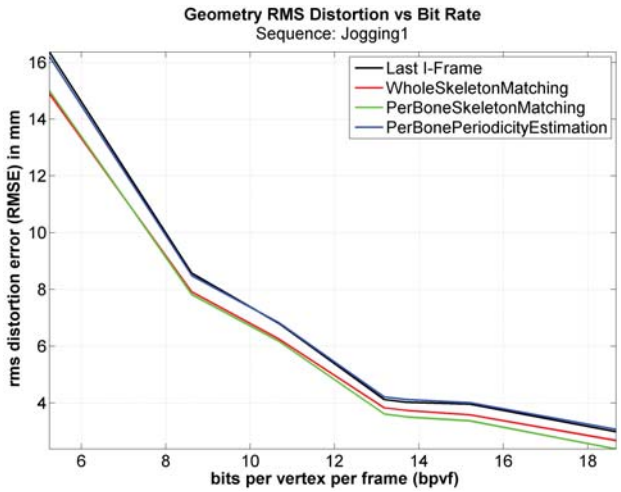

Fig. 5. Bit-rate vs Distortion for "Jogging" sequence

periodic movement, which can easily be captured by the proposed metric, as both Fig. 2 and Fig. 3 show.

4.2. Distortion Metric

Before evaluating all the presented methods in rate-distortion terms, we define the distortion metric that is going to be used. Let the original TVM sequence be denoted as **SM** and the compressed one as $\hat{\mathbf{SM}}$. Let also N be the total number of frames and K_t denote the number of mesh vertices in frame t. Then, the Root Mean Squared (RMS) Distance of $\hat{\mathbf{SM}}$ from **SM** is given from:

$$\mathrm{d}_{\mathrm{rms}}(\mathbf{SM}, \hat{\mathbf{SM}}) := \sqrt{\frac{1}{\sum_{t=0}^{N-1} K_t} \sum_{t=0}^{N-1} \sum_{k=0}^{K_t-1} ||\mathbf{v}_t^k - \mathbf{H}(\mathbf{v}_t^k)||_2^2}, \tag{6}$$

where $\mathbf{v}_t^k$ denotes the k-th vertex of the original mesh in frame t and $\mathbf{H}(\mathbf{v}_t^k)$ stands for its nearest vertex in the corresponding compressed mesh. The Root Mean Squared (RMS) Distance of **SM** from $\hat{\mathbf{SM}}$, let $\mathrm{d}_{\mathrm{rms}}(\hat{\mathbf{SM}}, \mathbf{SM})$, is defined equivalently. Then, the distortion introduced by compression is expressed by the metric:

$$\mathrm{RMSE} := \max(\mathrm{d}_{\mathrm{rms}}(\mathbf{SM}, \hat{\mathbf{SM}}), \mathrm{d}_{\mathrm{rms}}(\hat{\mathbf{SM}}, \mathbf{SM})). \tag{7}$$

4.3. Results

During the experimental results, for the skeleton matching approach as well as for the periodicity estimation approach, a frame buffer of 30 I-Frames was used. Moreover, for periodicity estimation, a running window of 60 frames was employed. I-Frames where generated every 5-th frame. The rest of the parameters were chosen as follows: $K_\mathrm{p} = 80$, $K_\mathrm{o} = 10$, $W_\mathrm{p} = 0.9$, $W_\mathrm{o} = 0.1$, $W_\mathrm{c} = 0.05$. In Fig.4 the results of the presented methods are depicted for "Skiing". We notice a similar performance when encoding always with respect to the last encoded I-Frame or when using both of the skeleton matching methods (per bone, or as a whole). On the other hand, periodicity fails terribly. These results, despite being unexpected, can be very well explained when considering the underlying skeleton quality for this sequence. The skeleton quality for the "Skiing" sequence, when judged visually, can be easily classified as bad. In contrast, in "Jogging" sequence (Fig. 5), where the skeleton quality is significantly better but still not perfect, more interesting results are obtained. The skeleton-matching approach in a per bone basis along with skeleton matching as a whole show a performance advantage over the rest of the methods. Contrariwise, using periodicity estimation or encoding with respect to the last I-Frame have a similar worse performance. The results in the "Jogging" sequence, better reflect our intuition for the expected outcome.

5. CONCLUSIONS

In this paper, a strategy for selecting the most appropriate I-Frame that will serve as a reference frame for the encoding of EP-Frames in TVM compression, exploiting activity-related characteristics, was introduced. Two different strategies were presented, using a skeleton-matching criterion and a periodicity measurement metric based on skeleton data. Evaluation results show that the concept is sound, but also they reveal the sensitivity of the proposed methods to the skeleton quality, thus the need for more robust skeleton tracking techniques. Future research could also explore different strategies to generate I-Frames depending on the activity scenario. Dynamically generating I-Frames and EP-Frames in continuous predefined intervals and thus, potentially further exploiting the activity characteristics, will constitute a future direction.

6. REFERENCES

[1] [Online]. Available: https://www.gti.ssr.upm.es/~mpeg/3dgc/3Dmodels/

[2] S.-R. Han, T. Yamasaki, and K. Aizawa, "Time-varying mesh compression using an extended block matching algorithm," *IEEE Trans. Circuits Syst. Video Techn.*, vol. 17, no. 11, pp. 1506–1518, 2007.

[3] S.-R. Han, T. Yamasaki, and K. Aizawa, "Geometry compression for time-varying meshes using coarse and fine levels of quantization and run-length encoding," in *ICIP*, 2008, pp. 1045–1048.

[4] T. Yamasaki and K. Aizawa, "Patch-based compression for time-varying meshes," in *ICIP*, 2010, pp. 3433–3436.

[5] W. E. Vieux, K. Schwerdt, and J. L. Crowley, "Face-tracking and coding for video compression." in *ICVS*, ser. Lecture Notes in Computer Science, H. I. Christensen, Ed., vol. 1542. Springer, 1999, pp. 151–160.

[6] D. S. Alexiadis, D. Zarpalas, and P. Daras, "Real-time, full 3-d reconstruction of moving foreground objects from multiple consumer depth cameras," *IEEE Transactions on Multimedia*, vol. 15, no. 2, pp. 339–358, 2013.

[7] S. Chen, P. Xia, and K. Nahrstedt, "Activity-aware adaptive compression: a morphing-based frame synthesis application in 3dti," in *ACM Multimedia*, 2013, pp. 349–352.

[8] K. Mamou, T. B. Zaharia, and F. J. Prêteux, "Tfan: A low complexity 3d mesh compression algorithm," *Journal of Visualization and Computer Animation*, vol. 20, no. 2-3, pp. 343–354, 2009.

[9] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake, "Real-time human pose recognition in parts from single depth images," in *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition*, ser. CVPR '11. Washington, DC, USA: IEEE Computer Society, 2011, pp. 1297–1304.

[10] M. Kazhdan, M. Bolitho, and H. Hoppe, "Poisson surface reconstruction," in *Proceedings of the fourth Eurographics symposium on Geometry processing*, ser. SGP '06. Aire-la-Ville, Switzerland, Switzerland: Eurographics Association, 2006, pp. 61–70.

[11] [Online]. Available: http://www.openni.org/files/nite/