Metadata-Based Activity Analysis in 3D Tele-Immersion

Shannon Chen*, Aadhar Jain, Zhenhuan Gao*, Klara Nahrstedt*, Ahsan Arefin, and Raoul Rivas
University of Illinois at Urbana-Champaign*  BigO Technologies  Google  Intel
{cchen116, zgaoo11, klara}@illinois.edu  aadharjain@thebigo.co.in  aarefin@google.com  rarivas@intel.com

Abstract—In view of the necessity of activity analysis on Gbps-scale 3D Tele-Immersive (3DTI) content bundle and petabyte-scale 3DTI recordings, this work proposes and verifies the feasibility of replacing high-latency intrusive analysis with light-weighted metadata-based analysis. For real-time in-session use case, result shows that metadata-based analysis module, when personalized with user’s body index, can achieve accuracies in 90 percentile on classifying various activity classes. For offline cross-session use case, we propose a hybrid analysis scheme which combines the advantages of intrusive analysis (i.e., conventional activity analysis on content level) and metadata-based analysis and achieves high accuracy with low computation latency.

Keywords – 3DTI; Metadata; Activity analysis

I. INTRODUCTION

3D Tele-immersion (3DTI) technology allows full-body, free-viewpoint visual content delivery among geographically dispersed users. In a 3DTI physical site (Fig. 1), user’s 3D model is captured by multiple RGB-D (color plus depth) cameras. This enables omnidirectional activity recording and free-viewpoint online interaction or offline playback. While 3DTI opens up opportunities for a wide variety of cyber-collaborations including remote healthcare [1][2], art performance [3][4], and critical scene investigation [5][6], the large size of 3DTI content incurred by multi-view capturing makes analysis of 3DTI activities inevitably computation-expensive.

The necessity of activity analysis comes in two major parts of the delivery chain of 3DTI applications: 1) online quality adaptation and 2) offline record processing. 3DTI system is characterized by the diversity and contingency of human activities [7]. Regarding quality adaptation, this contingency is critical as the nature of user activity drives the type and the number of devices to be activated and their sampling quality during 3DTI session runtime. Furthermore, the collaborative nature of 3DTI activities imposes different QoS requirements in terms of bandwidth, delay, jitter and skew across activities [7][8]. For example, 3DTI conversation requires low skew and moderate video quality, while exergaming requires low delay and high video frame rate to ensure high interactivity. Therefore, detection and recognition of human activity is essential to provide efficient QoS provisioning and device management in 3DTI space.

On offline record processing, anomalous activity detection across recorded 3DTI sessions provide important hint on review recommendation and summarization in many asynchronous applications [14][15]. For example, in asynchronous physical rehabilitation, patients will be offline-supervised by uploading their recorded rehabilitation sessions at home; while professional therapist follows up on patients’ progress by reviewing the recordings at clinic. In this scenario, anomaly detection pinpoints the time frames of patients fell or injured across all stored recordings to help therapists prioritize their reviews. In the rest of this paper, we refer to this activity analysis for offline record processing as cross-session analysis as oppose to in-session analysis which serves the aforementioned online quality adaptation purpose.

Both types of activity analysis are challenged by the large data size of 3DTI’s media-enriched content. For in-session analysis, major challenge comes from the high bitrate of the media stream bundle of 3DTI. Encompassing multiple RGB-D video streams, audio stream, and other application-specific sensory streams, 3DTI’s raw media bundle consumes approximately 1,100 Mbps bandwidth [9][10]. With data rate this high, real-time in-session analysis becomes a substantial computation burden for general-purpose home PC [9], let alone portable devices [17].

As for cross-session analysis, both quantity and size of 3DTI recordings become obstacles for efficient offline anomaly detection. Take physical healthcare as an example, for a patient to submit 30-minute rehabilitation exercise daily and for each recording to be preserved for 30 days, the storage entity is expecting 0.7 TB (compressed [10]) data uploaded by each patient [14]. As modern electronic health record databases are designed to serve a number of patients in hundred-scale, the total size of stored 3DTI recordings to be analyzed can easily approach petabyte magnitude.

In addition to bitrate, data size, and quantity issues, a more fundamental computational barrier which both in-session and cross-session analysis have to overcome is the 3D nature of its content. If the analysis is intrusive (i.e.,
direct analysis on the content level such as computer-vision-based analysis, then the extra dimension would impose higher complexity than conventional 2D content. Although there exist a large body of researches and tools regarding fast activity analysis on 2D videos, content analysis as simple as feature detection and feature tracking for 3D content still remain in second-scale per-frame complexity [18][19] (without hardware acceleration as Kinect does), which is far from feasible for big data processing.

In view of the problems, we propose to approach activity analysis in 3DTI with a different set of inputs: instead of analyzing the multimedia content itself, we analyze the application generated metadata and related system metadata. From previous works [20][21], we know that 3DTI system, like any other distributed interactive multimedia environments, has important characteristics of dependency constraints among content, application, and underlying systems components over time. For example, a high-mobility activity incurs smaller key-frame intervals and hence requires more CPU resource on decoding. Therefore, by monitoring the generated frame type and CPU usage on-the-fly, we may be able to backtrack the motion level of the user activity. Comparing to intrusive analysis on the content level, monitoring and analysis of metadata on application and system levels are computationally light-weighted. Thus, we argue that metadata-based activity analysis is more suitable for big multimedia data such as 3DTI.

To verify the practicality of metadata-based activity analysis in 3DTI, this paper answers the following questions:

1. Where and how can metadata analysis help a 3DTI system throughout the service delivery?
2. What is the performance of metadata analysis in terms of activity recognition/detection accuracy?
3. What is our gain from adopting metadata analysis in terms of computation latency?
4. Can metadata analysis fully replace intrusive content analysis on activity recognition? If not, then:
   a. What are the constraining factors?
   b. How can metadata analysis aid intrusive analysis in a hybrid solution?

The rest of the paper is organized as follows. In Section II, we review some related works. In Section III, we present the system model. In Section VI and V, we introduce the metadata analysis modules we devise for in-session analysis and cross-session analysis, respectively. In Section VI, we verify the performance of our designs and answer the raised questions with experiment results. Finally, in Section VII we conclude the paper.

II. RELATED WORKS

A. How Big is a 3DTI Application Session?

Equipped with 3D camera array, microphone, and other application-specific sensors, a 3DTI site captures the activity of its user and digitizes it to enable full-body, free-viewpoint experience. However, as pointed out by several previous works [9][10][11] on 3DTI capturing interfaces, bitrate of the captured content bundle is oftentimes too high to be analyzed efficiently. An empirical calculation provided by Yang et al. [9] reports a 300 Mbps bitrate for visual streams in 3DTI with low quality setting (320x240 resolution, 10 fps). A more modern hardware setting in interface proposed by Mekuria et al. [10] introduces an even higher 1,032 Mbps bitrate with Kinect camera (assuming 640x480 resolution, 30 fps) without compression.

B. Activity Analysis in 3DTI

Activity analysis in 3DTI serves the important purpose of providing semantic information to quality adaptation and resource allocation modules. Quantitative analysis [7][21][22] on QoS-QoE relationship (quality of service/experience) in various 3DTI applications shows that, properties of different human activities (e.g., spatial coverage and motion level) decide which of the QoS factors (e.g., frame rate, resolution) dominant the resulting QoE. Thus, in [23], the author purposes the concept of activity semantics in 3DTI to help the underlying system be aware of the characteristics of current user activity. This enables the system to adjust the quality of content capturing accordingly (e.g., to ramp up the frame rate when activity has high motion) and hence to achieve more efficient content delivery. In [24] and [25], user activity identified with motion data collected by on-body sensors are used to configure the compression parameters of 3DTI content in online and offline applications, respectively, to save transmission bandwidth without introducing noticeable quality degradation on the semantic level.

C. The Cost of Intrusive Analysis

Intrusive analysis on video content is known to suffer from high latency. The state-of-the-art computer-vision-based analysis for feature tracking: optical flow and its variants, have per-frame latencies of 10–37s [26][27][28]. In [29], an implementation of optical flow with GPU acceleration is proposed, which achieves a 5s per-frame latency. Although optical flow is originally designed to analysis 2D content, extension has been made in [19] and [30] to make possible the analysis of 3D videos. The latencies reported in these works are 3–5 min per frame. These reported latencies show that intrusive analysis is not sufficient for providing real-time information for quality adaptation, or offline analysis of big 3DTI recordings in petabyte scale. Yet, in return for the high computation cost, the accuracy of activity analysis based on these techniques is higher than 97% [31].

III. SYSTEM MODEL

The system model of 3DTI is illustrated in Fig. 1. Depending on properties of the application and the role each user plays, a 3DTI user site can be classified as immersive or non-immersive. An immersive site is equipped with a camera array comprise of multiple RGB-D cameras, an output screen, and other application-specific sensors (e.g., medical sensors or game consoles). User’s activity in an immersive site will be captured and rendered by a gateway machine.
through these I/O devices. The gateways of multiple immersive sites are connected together through Internet2 so that immersive users can interact with each other in a synchronized virtual space created by the system.

As the interaction happens, copies of streams from all participating immersive sites are sent to a storage entity. A non-immersive site can stream a recorded session from the storage entity in an on-demand manner.

Activity analysis for quality adaptation, i.e., in-session analysis, is carried out by the gateway machines of immersive sites. Since in-session analysis have to provide on-the-fly feedbacks to configure capturing and rendering details (e.g., resolution and frame rate), the analysis model has to reside in the same machine with these components. On the other hand, activity analysis for record processing, i.e., cross-session analysis, is carried out by the storage entity. Since this part of the analysis focuses on anomalous activity detection, the analysis model has to have global knowledge across previous recorded sessions. In addition, as we mentioned before, cross-session analysis is challenged by computation in petabyte scale. Therefore, we design it to be carried out by the most powerful entity in our model.

IV. IN-SESSION ANALYSIS

In this section, we present our design for in-session analysis. First, we introduce the set of metadata we monitored during the session runtime and the extraction of features. Second, we introduce our analysis module trained by the extracted metadata. Last, we discuss the cyber-physical factors that may affect the performance of our analysis.

A. Metadata Extraction

At each participating site, we monitor application and system metadata corresponding to application, I/O devices, and underlying system resources. A list of metadata we focus on in our analysis is provided in Table I.

<table>
<thead>
<tr>
<th>Application metadata</th>
<th>System metadata</th>
</tr>
</thead>
<tbody>
<tr>
<td>Camera frame rate</td>
<td>CPU usage</td>
</tr>
<tr>
<td>Video frame size</td>
<td>Memory usage</td>
</tr>
<tr>
<td>3D reconstruction time</td>
<td>Camera latency</td>
</tr>
<tr>
<td>Rendering time</td>
<td>Gateway latency</td>
</tr>
</tbody>
</table>

These metadata are sampled periodically by our metadata monitoring framework (DIAMOND [12]) and the values are stored in a timestamp-indexed database.

Feature extraction transforms recorded application and system metadata into a reduced representation, i.e., a feature vector. Our feature vector consists of metadata corresponding to T consecutive time units. We consider three kinds of features extracted from the time-series metadata: 1) absolute value, 2) time difference value, which computes the variation of metadata value with respect to the previous time epoch, and 3) device difference value, which computes the variation of the same type of metadata value generated from correlated input devices (e.g., framerate of all local cameras) at the same time epoch.

B. Activity Analysis Module

The architecture of our activity analysis model is shown in Fig. 2. It takes a supervised machine learning approach to classify human activity using application and system metadata. The module contains three components: 1) Feature Extractor that constructs feature vectors introduced in the previous section, 2) Model Trainer that uses these feature vectors and input activity labels to train a classification model based on SVM [13], and 3) Activity Classifier that uses both the feature vectors and the model feeds to finally classify real-time human activity during 3DTI session runtime.

C. Cyber-Physical Factors

Since we classify human activities based on time-series metadata, the classification process we propose is expected to be fast and unobtrusive. However, since our analysis module does not have knowledge about the streamed content on pixel level, several cyber-physical factors can influence its performance.

Cyber-physical dimensions such as visual color (e.g., clothing of users and lighting of the physical user space) and space volume of the physical contents (e.g., body index of users) can impact the application and system metadata (e.g., the reconstruction time and frame size). Intrusive computer-vision-based analysis does not have this problem because it neutralize the cyber-physical effects by focusing on extracted feature points (e.g., skeleton joints of users).

To address this issue, in the verification of our module (Section VI) we will evaluate the impacts of color and volume factors on the performance of our in-session analysis.

V. CROSS-SESSION ANALYSIS

In this section, we present our design for cross-session analysis. First, we introduce the storage (compression) scheme used in the storage entity. Second, we introduce the activity analysis module and the metadata it considers. Last, we discuss the likelihood of false-negative detections and propose a hybrid analysis scheme that combines the advantage of metadata and intrusive analysis.
The total bitrate of a raw 3DTI content bundle is in gigabit magnitude [9][10]. This implies that, without compression, a 30 min recording will take up more than 200 GB space in the storage entity. Therefore, an effective compression scheme is essential. Since RGB-D streams contribute most of the bitrate [14], the compression scheme for visual content is customized for asynchronous applications to aim for higher compression ratio.

From our previous works [14][16], we observe that visual content of common asynchronous applications (e.g., physical training, rehabilitation) share the following properties:

1. Users will be repeating the same set of movements in different recordings (e.g., rehabilitation).
2. Users will move their bodies in consistent pace and motion range in every recording since they are following the demonstration video provide by the reviewer.
3. Background can be removed with depth information since it is not the focus of reviewer.

Combining these observations, we know that visual contents recorded in consecutive submissions will be very similar to each other. We devise a compression scheme which compresses multiple consecutive recordings together in order to exploit their similarity.

The idea is realized by imposing inter-frame coding not only on adjacent frames in the same video, but also on videos recorded in adjacent submissions. An illustration of our approach is provided in Fig. 3. In Fig. 3a, we see the recorded raw video frames of each submission can be arranged in a 2D array. Each row of frames represents frames in the same video, and adjacent rows are videos in adjacent submissions. By applying MPEG encoding on color and depth frames separately for each video, frames will be encoded into I-, P-, and B-frames as illustrated in Fig. 3b. While conventional codecs stop at this point, we propose to further encode each column of I- and P-frames together (Fig. 3c) to exploit the inter-video similarity. Fig. 3d shows the final coded frames after column-wise inter-frame coding. A large number of key frames are replaced by predicted frames. In result, our evaluation in [14] shows substantial compression gain (x1.73) from conventional video compression.

**Activity Analysis Module**

As we mentioned earlier, the focus of cross-session analysis is to detect the anomalous activities across all stored recordings. To this effect, our activity analysis model examines the size of predicted frames (P- and B-frames) in the column-wise inter-frame coding (Fig. 3c). By the design of video frame types (I, P, and B), a predicted frame contains only the difference between itself and its reference frame. Thus, a larger difference in the contents contributes to a larger size.

The result of a preliminary experiment to show the correlation between predicted frame size and the difference between a predicted frame and its reference frame is provided in Fig. 4. In the experiment, we take a 3DTI video and encode it with different “frame elapses”. The frame elapse determines the number of frames between a predicted frame and its reference frame in the encoded video. When the elapse is set high, the contents carried by a predicted frame and its reference would be more different. From the plot we see a positive correlation between the predicted frame size and contents difference (elapse). The result of ANOVA test also shows a positive correlation with statistical significance (p<0.001) on the compiled data.

Knowing that the predicted frame size being a good indicator for the differences between two video submissions, we devise our analysis module to examine the size of the column-wise encoded predicted frames (Fig. 3c) to detect the anomalous activities. In this case, since the relationship between the chosen metadata (i.e., frame size) and the content differences is clear, the predicted model adopted by our analysis module is a simple decision tree [13]. The objective of the training phase is to learn a set of numerical thresholds that can filter out the negative instances at each node in the tree.

**Hybrid Analysis Scheme**

Anomaly detection is not equivalent to detection of content differences. Although the latter is a good indicator for the former in common cases, when it comes to applications with more stringent accuracy requirements (e.g., healthcare), the false-negative rate of our analysis module may not be qualified.

Thus, rather than fully replacing intrusive analysis by metadata-based analysis, we propose a hybrid analysis
scheme that runs the metadata-based analysis first to provide preliminary hints which can be used to speed up the following intrusive analysis. For example, assume we have a high latency intrusive analysis module with 100% precision and recall; and a fast metadata analysis module with 50% precision and recall. To perform anomaly detection on a dataset of size 1,000 with 100 evenly scattered anomalies, intrusive analysis alone need to go through 50% of the dataset to locate 50 of the anomalies. However, with pre-processing by metadata-based analysis, the following intrusive analysis module only need to go through 10% of the dataset, suggested by metadata-based analysis, to locate 50 anomalies. While the reviewer is reviewing the 50 anomalies, the rest of the dataset will be processed by the intrusive analysis to capture the remaining anomalous activities.

VI. VERIFICATION

A. Latency of Metadata Analysis

1) Experiment Settings

Since in-session and cross-session analyses are targeting on different use cases, we setup two separate hardware environments to evaluate the latencies introduced by their metadata-based activity analyses. For in-session analysis, we run the proposed module on a desktop with Intel Core i5 CPU (4 cores, 3.50 GHz). The setting resembles the capability of a common gateway machine in a 3DTI site. For cross-session analysis, we run the analysis module on a HPC server with 24 Intel Xeon CPUs (each has 6 cores, 2.50 GHz). The HPC server resembles the capability of the storage entity in our system model.

2) Latency

In Table II we show the latency of activity analysis in in-session and cross-session use cases. As references, we also list the latency of various intrusive analysis reported in previous works based on optical flow [26][29][30].

<table>
<thead>
<tr>
<th>Metadata analysis</th>
<th>In-Session Analysis</th>
<th>Cross-Session Analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Activity analysis on 3D content [30]</td>
<td>300 s</td>
<td>1027 years</td>
</tr>
<tr>
<td>Activity analysis on 2D content [26]</td>
<td>10 s</td>
<td>34 years</td>
</tr>
<tr>
<td>Activity analysis on 2D content with GPU [29]</td>
<td>5 s</td>
<td>17 years</td>
</tr>
</tbody>
</table>

As we can see from the table, the latency of in-session analysis is two magnitudes higher than cross-session analysis. The obvious reason behind this is the computing power difference between their hardware environments. However, even if we set the cross-session analysis module to be run by only one of the 24 processors on the server, the resulting latency (0.37 ms) is still much smaller than in-session analysis. This is because both collection and analysis of metadata in the two use cases have different complexities. In the collection phase, the metadata set collected for in-session analysis (CPU usage, rendering time, etc.) is not only much larger than cross-session analysis (frame size), but it also spans across system and application domains and involves monitoring of multiple modules. In the analysis phase, the SVM-based module of in-session analysis has a higher computational complexity than decision-tree-based module of cross-session analysis. A notable property is that, even with a higher complexity, the latency of in-session analysis is still eligible of processing more than 30 frames per second. This means that the proposed in-session analysis module does not violate the real-time constraint and is capable of analyzing user activities on-the-fly.

B. In-Session Analysis

1) Experiment Settings

We setup a 3DTI site with a gateway with two (upper and lower body) RGB-D cameras and one screen. We run the proposed in-session analysis module and attempt to classify among a set of six activity classes shown in Fig. 5. For each of the six activity classes, 4 minutes of metadata is captured during the activity session with 8 different participants. The data is collected under white lighting condition. Participants involved in the experiments are given basic instructions on how to perform the respective activities, without any further information on how the application would classify their activities.

2) Accuracy

To evaluate the accuracy, we adopt k-fold cross-validation [13] to guarantee that the trained model is tested on random activity sessions’ metadata that is not seen during the training phase. In Table III we show the overall accuracy of our module, where the rows are the ground truth and the columns are the predicted activity class. The labels of the activity classes comply with Section VI.B.1 (Fig. 5). To analyze the results in detail, we group the accuracies on recognizing different activity classes by their motion characteristics including position, posture, speed, and range (Fig. 5). The grouped accuracies are show in Table IV. We can see that almost all of the classifications reach accuracies in 90 percentile. The only exception is the recognition of activity speed, which only reaches an overall accuracy of

![Fig. 5. Activity Classes.](image-url)
74%. This shows that the set of input metadata considered by our in-session analysis module does not capture the speed of user’s movement as well as the other motion characteristics.

<table>
<thead>
<tr>
<th>TABLE III.</th>
<th>RESULT OF ACTIVITY CLASS RECOGNITION (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Act 1</td>
</tr>
<tr>
<td>Act 1</td>
<td>96.30</td>
</tr>
<tr>
<td>Act 2</td>
<td>0.00</td>
</tr>
<tr>
<td>Act 3</td>
<td>0.00</td>
</tr>
<tr>
<td>Act 4</td>
<td>2.50</td>
</tr>
<tr>
<td>Act 5</td>
<td>0.00</td>
</tr>
<tr>
<td>Act 6</td>
<td>0.00</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>TABLE IV.</th>
<th>RESULT OF ACTIVITY CLASS RECOGNITION (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Position</td>
<td>still</td>
</tr>
<tr>
<td>still</td>
<td>96.07</td>
</tr>
<tr>
<td>moving</td>
<td>5.54</td>
</tr>
<tr>
<td>Speed</td>
<td>fast</td>
</tr>
<tr>
<td>fast</td>
<td>75.92</td>
</tr>
<tr>
<td>slow</td>
<td>28.06</td>
</tr>
</tbody>
</table>

3) Impact of Cyber-Physical Factors

We realize that human activity classification accuracy is influenced by certain cyber-physical factors. We consider two cyber-physical factors in our study: 1) visual color of the cyber-physical content, and 2) volume of the cyber-physical content.

To address the color factor, we ask the participants to put on t-shirts of red, green, blue, and white and record the metadata when they are standing, sitting, and moving around the user space. After the data collection, we run the experiment in four phases, each focusing on one color. In each phase, the metadata corresponding to the focused color becomes the training data, and the rest becomes the test data. Due to space limit, in Fig. 6a we only show the result of the red phase. Across colors and activities, our analysis module achieves >85% accuracies. As shown in the figure, the results of each color being similar to each other indicates limited impact from the color factor on metadata-based analysis.

To address the volume factor, we collect the body indices (weight $\times$ height$^2$) from participants who are willing to reveal. Since the main contents of most 3DTI applications are users’ bodies, the volume of the human body is the dominant contributor to the volume factor. Body indices of each participant are listed in Table V. The experiment again is broken down into several phases. In each phase, the metadata corresponding to one participant becomes the training data, and the rest becomes the test data. In Fig. 6b, we plot the result of using metadata of participant #3 (body index = 176) to train the model. As shown in the plot, we see a gradual dropping of accuracy as the difference between body indices of the training data and the testing data increases. This indicates that the impact of volume factor is significant enough to affect the metadata, and hence the accuracy becomes defected as the volumes of training and testing data are different. The same phenomenon is observed when taking metadata corresponding to other participant as training data.

<table>
<thead>
<tr>
<th>TABLE V.</th>
<th>BODY INDICES OF PARTICIPANTS</th>
</tr>
</thead>
<tbody>
<tr>
<td>ID</td>
<td>Height (m)</td>
</tr>
<tr>
<td>1</td>
<td>1.62</td>
</tr>
<tr>
<td>2</td>
<td>1.70</td>
</tr>
<tr>
<td>3</td>
<td>1.73</td>
</tr>
<tr>
<td>4</td>
<td>1.66</td>
</tr>
<tr>
<td>5</td>
<td>1.80</td>
</tr>
<tr>
<td>6</td>
<td>1.77</td>
</tr>
<tr>
<td>7</td>
<td>1.77</td>
</tr>
</tbody>
</table>

The observed impacts from cyber-physical factors imply that, in practical in-session use case, metadata-based analysis modules have to be trained and used by the same user to preserve the same body index in training and testing data. In other words, metadata-based in-session analysis models should be personalized with each user’s body index to get the best activity recognition accuracy. In some applications like healthcare, this restriction would not be a problem. However, applications like gaming or conferencing could involve 3DTI sites for public use. In such cases, the analysis module will have to be equipped with multiple SVM models trained by metadata corresponding to different body indices. Each users’ body index can be included in their online account, so that when a user logs on to a public 3DTI site, the analysis module will know which model to load according to the user’s body index.

C. Cross-Session Analysis

1) Experiment Settings

To evaluate the performance of cross-session analysis, we record a series of simulated rehabilitation sessions of a 3DTI physiotherapy application. The recordings are taken in a studio setting with four RBG-D cameras surrounding the user space. The dataset consists twelve recordings, simulating rehabilitation sessions a patient would conduct in twelve consecutive submissions. To imitate recovery progress of patient throughout the twelve sessions. We strap
different amount of weights on actor’s body (Fig. 7). For instance, to imitate shoulder injury, we strap 13 to 3 lb. on actor’s left arm to imitate recovery progress from the first to the last submissions.

To simulate anomalies, we extract from each of the twelve recordings a 30 seconds long exercise that only involves shoulder rehabilitation (Fig. 7a). Then, in this set of shoulder exercise recordings, we randomly replace 36 one-second chunks by irrelevant exercises (Fig. 7bc) to simulate anomalous activities.

2) Accuracy

In Fig. 8 we plot the frame sizes of predicted frames in the encoded set of recordings. Each column of dots in the plot represent frames being column-wise inter-frame coded together. The blue round dots are original frames capturing shoulder exercise. The orange triangle dots are frames capturing anomalies. In Table VI, we show the accuracy of anomaly detection of our decision tree-based analysis module. The overall accuracy is 85%.

<table>
<thead>
<tr>
<th>TABLE VI. RESULT OF ANOMALY DETECTION (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Detect True</td>
</tr>
<tr>
<td>True Anomaly</td>
</tr>
<tr>
<td>False Anomaly</td>
</tr>
</tbody>
</table>

3) Hybrid Analysis Scheme

Although 85% accuracy of the proposed module seems promising, as we mentioned earlier, applications like healthcare may have more stringent requirements. Especially, as shown in Table VI, the accuracy of true anomaly detection is less than 70%. This indicates that metadata-based cross-session analysis alone is insufficient. Intrusive analysis on 3D content for anomaly detection, on the other hand, can reach >97% accuracy according to results reported in [31], which is more sufficient for stringent applications.

The goal of hybrid analysis scheme is to preserve the high accuracy of intrusive analysis and the low latency of metadata-based analysis at the same time. In Table VII, we shown the overall accuracy and latency of intrusive, metadata-based, and hybrid analyses.

<table>
<thead>
<tr>
<th>TABLE VII. OVERALL PERFORMANCE OF HYBRID ANALYSIS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time required to detect 70% anomalies</td>
</tr>
<tr>
<td>Metadata analysis</td>
</tr>
<tr>
<td>Intrusive analysis [31]</td>
</tr>
<tr>
<td>Hybrid analysis</td>
</tr>
</tbody>
</table>

We can see that hybrid analysis scheme saves a substantial amount of time comparing to intrusive analysis. This is due to the preprocessing of metadata analysis module. From the evaluation experiment in the previous section (Table VI), we know that the preprocessing of metadata-based analysis can locate 70% of the total anomalies (recall) in its detected instances, while the size of the detected instances account for 25% of the input contents. This implies that, the proceeding intrusive analysis only need to go through 25% of the input frames to locate 70% of the anomalous instances. Therefore, the required time of hybrid analysis is approximately ¼ of the time of intrusive analysis alone. Moreover, since the final detections of hybrid analysis are verified by intrusive analysis module, it directly inherits the accuracy and hence achieves the same >97% accuracy.

VII. CONCLUSION

In view of the necessity of activity analysis on Gbps-scale 3DTI content bundle and petabyte-scale 3DTI recordings, we proposed to partially replace high-latency intrusive analysis with light-weighted metadata-based analysis. In this work, we propose to adopt metadata-based analysis in both the 3DTI sites and the storage entity for in-session and cross-session analysis, respectively.

For in-session analysis, result shows that metadata-based analysis module, when personalized with user’s body index, can achieve accuracies in 90 percentile on classification of various activity characteristics. For cross-session analysis, we propose a hybrid analysis scheme which combines the advantages of both intrusive and metadata-based analysis and achieves high accuracy with low latency.
REFERENCES


