3D Teleimmersive Activity Classification Based on Application-System Metadata

Aadhar Jain Ahsan Arefin Raoul Rivas Chien-nan Chen Klara Nahrstedt
University of Illinois at Urbana-Champaign
{jain37,marefin2,trivas,cchen116,klara}@illinois.edu

ABSTRACT

Being able to detect and recognize human activities is essential for 3D collaborative applications for efficient quality of service provisioning and device management. A broad range of research has been devoted to analyze media data to identify human activity, which requires the knowledge of data format, application-specific coding technique and computationally expensive image analysis. In this paper, we propose a human activity detection technique based on application generated metadata and related system metadata. Our approach does not depend on specific data format or coding technique. We evaluate our algorithm with different cyber-physical setups, and show that we can achieve very high accuracy (above 97%) by using a good learning model.

Categories and Subject Descriptors

H.4.3 [**Information Systems Applications**]: Communications Applications: Computer conferencing, teleconferencing, and videoconferencing; C.2.2 [**Multimedia Information Systems**]: Video

General Terms

Algorithm, Performance, Design

Keywords

3D Tele-immersion, Activity, Classification

1. INTRODUCTION

3D teleimmersion (3DTI) technology has expanded the horizon of telepresence conferencing applications by supporting full-body interaction of physical activities in virtual reality environments. Applications have been found in cyber-archeology, collaborative dancing, video conferencing, and online gaming (e.g., [6], [5], [11]).

In 3DTI, participants' 3D images are captured in real-time and exchanged among multiple participants to allow shared virtual collaborations. The system is characterized by the diversity and contingency of human activities[1]. This contingency of activities in 3DTI is critical as the nature of user-activity drives the type and the number of devices to be activated during 3DTI session run-time. Furthermore, the collaborative nature of 3DTI activities imposes different QoS requirements in terms of bandwidth, delay, jitter and skew across activities [1, 2]. For example, 3DTI conversation requires low skew and moderate video quality, while exer-gaming 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. The primary goal of this paper is to identify *fine-grained* human activities (e.g., walking, sitting, running), which can be combined together to detect *coarse-grained* activities (e.g., video conferencing, exer-gaming).

A broad range of research has been devoted to analyze media data to identify human activity [7, 10], which requires the knowledge of application data format, data coding technique and computationally expensive image analysis. Here we propose a supervised learning-based *Activity Classification System (ACS)* that considers application generated metadata and related system metadata (*application-system* metadata) instead of application media data. A supervised learning algorithm trains our classification model for 3DTI setup without the knowledge of media data format or coding complexity. To our knowledge, our research is the very first attempt to use time-series application-system metadata in 3DTI activity detection.

Since we classify human activity based on time-series metadata, the classification process we propose here is fast (less than 4ms) and unobtrusive. However, the classification is influenced by several cyber-physical dimensions such as visual color and space volume of the physical contents (e.g., participants) which have impact on the application and system metadata (e.g., the reconstruction time and frame size). We quantify the extent of cyber-physical impacts on our activity classification model, and apply the experimental inferences to classify human activity in a real 3DTI setup efficiently. Using a real 3DTI setup, our solution achieves more than 97% accuracy in human activity classification.

2. SYSTEM MODEL

2.1 3DTI Application Model

3D teleimmersive system is a distributed platform that connects multiple remote participants into one virtual shared

ACM MM '13 Barcelona, Spain

space for 3D collaborative activities. A large number of input and output devices (e.g., 3D cameras, microphones and displays) are connected at each participating location (called site) and they are accessed via a single entry point (site gateway). Though multiple parties are involved, we only focus on classifying fine-grained human activities independently at each site. Our algorithm can be easily used for multiple sites by running activity detection at each individual site.

2.2 Application-System Metadata Model

At each participating site, we monitor application-system metadata information corresponding to application I/O devices and underlying system resources. Application metadata includes camera frame rate, audio bit rate, 3D reconstruction time, rendering time, audio and video frame size. The system metadata includes CPU usage, memory usage and processing time of the participating hosts (such as camera node and gateway node). Each host generates a set of metadata values periodically at each epoch or interval (e.g., every 50ms). The combined set of application and system metadata will hereafter be referred to as 'metadata' in this paper. Metadata values are stored in a timestamp indexed database and are used for activity classification.

2.3 Activity Model

We attempt to achieve classification for fine-grained human activities including sitting, standing still and different walk movements performed by the 3DTI participants. We also analyze the impact of cyber-physical dimensions on the classification accuracy. As we mentioned before, this classification model can be extended to classify coarse-grained 3DTI activities by combining the contingency and extent of fine-grained activities.

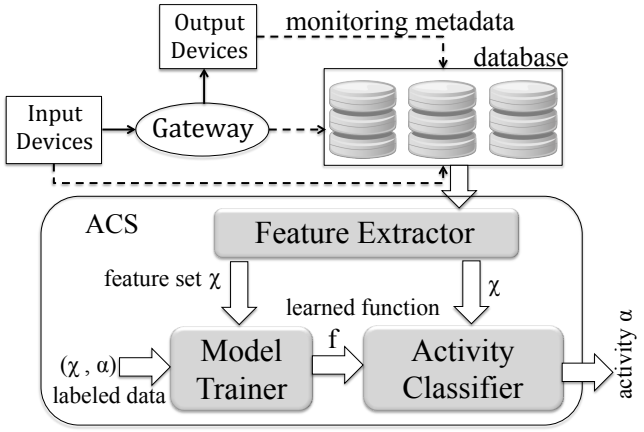

Figure 1: Architecture of activity classification system.

3. SYSTEM ARCHITECTURE

The system architecture of ACS is shown in Figure 1. It takes a supervised machine learning approach to classify human activity using application-system metadata. The time-series metadata values are stored in a local database attached to each site. ACS contains three components: 1) *Feature Extractor* that reads metadata from the database and extracts features to construct feature vectors interpretable by a machine learning model, 2) *Model Trainer* that uses these feature vectors and input activity labels to train a classification model, and 3) *Activity Classifier* that uses both the feature vectors and the model feeds to finally classify real-time human activity during 3DTI session run-time.

3.1 Feature Extraction

Feature extraction transforms recorded application and system metadata into a reduced representation of a set of features (also known as a feature vector). A feature vector is the smallest unit of training and testing. It 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 w.r.t. the previous time epoch, and 3) *device difference value*, which computes the variation of the metadata value generated from the correlated input devices (e.g., all local cameras) at the same time epoch. If there are N metadata parameters and x_i^t is the i^{th} metadata value collected at time t, the feature vector (χ) contains: $\cup_{t=1}^{T} x_i^t$, $\forall_{1\leq i\leq N}$ (absolute values), $\cup_{t=2}^{T} x_i^t - x_i^{t-1}$, $\forall_{1\leq i\leq N}$ (time difference values), and $\cup_{t=1}^{T}(x_i^t - x_j^t)$ (device difference values), where i^{th} and j^{th} metadata values are generated from correlated devices and $1 \leq i, j \leq N$.

3.2 Model Training

Our objective is to train a classifier using a supervised machine-learning model to classify human activity based on 3DTI metadata. To this effect, we use the implementation of SVM [3] by Learning Based Java (LBJ) [9], which is a special purpose programming language based on Java for machine learning. Given labeled m samples containing feature vectors and associated labels $\{(\chi_1, \alpha_1), (\chi_2, \alpha_2), ..., (\chi_m, \alpha_m)\}$ where α_i is the class label for the sample feature vector χ_i, the SVM algorithm learns $f{:}\chi{\rightarrow}\{-1, +1\}$, where the function f maps samples χ to a class $\alpha \in \{+1, -1\}$ and is represented by $sgn(w^T\chi - \Theta)$ where $w \in R^n$ (weight vector) and $\theta \in R$ (constant threshold).

Since we consider multiple activities, for multi class classification, the LBJ library implementation of SVM employs the one vs. all strategy. This strategy learns independent binary classifiers for each class label α_i treating a sample (χ_i, α_i) as a positive training sample $(+1)$ only for class label α_i and negative (-1) for all other class labels.

3.3 Activity Classification

Once the model has been trained, it predicts the human activity given a feature vector. Having learnt independent classifiers (weight vectors w_j) for each class label α_j, the model performs the following calculation: $f(\chi){=}argmax_j$ $w_j^T(\chi)$, where χ is the to-be-classified feature vector, w_j is the weight vector learnt for the j^{th} activity class and $f(\chi)$ is the predicted class label.

4. EXPERIMENTS

Using ACS framework, we conduct a series of experiments to 1) study the impact of certain cyber physical parameters on our activity classification model, and 2) employ an efficiently trained model to classify activities.

4.1 Effect of Cyber Physical Parameters

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

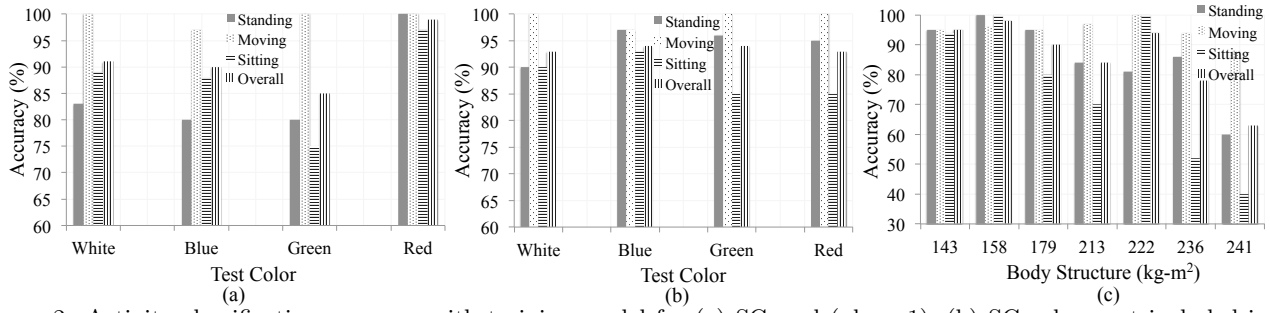

Figure 2: Activity classification accuracy with training model for (a) SC=red (phase-1), (b) SC values not included in the test (phase-2), and c) participant having BS=176 (phase-1).

usually subtracted, the visual color and physical texture of human body impact frame size metadata in mesh-based 3D reconstruction of fore-ground image. Likewise, the volume of the fore-ground content impacts frame size value of the constructed image. Change in frame size metadata changes 3D reconstruction time, data transmission time, network bandwidth and hence frame rate.

The visual color dimension is dependent upon physical lighting condition, shirt color as well as skin color of the participants. The volume dimension is dependent upon the participants' weights and heights. To quantify the impact of these two cyber-physical dimensions, we consider two metrics: 1) *Shirt Color (SC)*, and 2) *Body Structure (BS)*. We consider four different shirt colors: white, blue, green, and red. The body structure (BS) is derived from the height (H) and weight (W) of the participant as follows: $BS = H(m)^2 * W(kgs)$. Following the definition of *BMI (body mass index)*, we include the height parameter as squared in the BS computation formula.

4.1.1 Experimental Setup

We setup a 3DTI site with one gateway, two (upper and lower body) 3D bumblebee cameras (constructing mesh based 3D video streams via connected computers), one renderer and one database node. We run ACS on the local database machine and attempt to classify between a basic activity set $S = \{stand, sit, move\}$. For each activity session, 4 minutes of metadata is recorded. The data is collected under constant white lighting conditions. People 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 activity. Each classification unit is constructed with $T = 50$ and $N = 20$, which are the customizable parameters in 3.1.

To understand the impact of cyber-physical dimensions on human activity classification, we run experiments in two phases. In **phase-1**, we train the model on one random cyber-physical dimension value from the dimension spectrum (e.g., SC=blue) and record the classification accuracies given by the model on testing with other values (e.g., SC=red). In **phase-2**, we train the model on a more representative set (i.e., with multiple values) from the cyber-physical dimension spectrum, and record the classification accuracies on a random value (not included in the training).

4.1.2 Results and Discussion

Impact of Shirt Color (SC). To understand the impact of the SC dimension, we minimize the impact of content volume (i.e., BS) by recording application-system metadata for the same person under the same white light condition. We show the result for phase-1 in Figure 2(a). A high level of accuracy (> 85% overall) is achieved for activity classification where training is done based on activity (moving, standing and sitting) metadata collected only for SC=red. Changing the SC value (i.e., visual shirt color) of the participants shows limited impact on classification accuracy. We can conclude that the skin color of the participants and the lighting condition will not have high impact on activity classification accuracy either, since the learning is done based on the time-series patterns (which are similar across different SC values) of metadata.

In phase-2, the model is trained on data corresponding to all SC values except the SC value included in the test. The model achieves an average accuracy of about 94% as shown in Figure 2(b). This implies that when the ML algorithm is exposed to data corresponding to a wide range of SC values during training, the impact that the change in color has on the features helps it achieve higher accuracies for other values in the spectrum (even though the particular SC value is not seen during training).

Table 1: User data of body structure.

ID	Height (m)	Weight (kgs)	BS (kg-m^2)
1	1.73	59.0	176
2	1.70	54.5	158
3	1.80	65.7	213
4	1.75	72.5	222
5	1.77	77.0	241
6	1.66	65.0	179
7	1.62	54.5	143
8	1.77	75.0	236

Impact of Body Structure (BS). To study the effect of body structure, activity data is collected for 8 different people, all wearing a red shirt to minimize the impact of SC dimension. Table 1 shows the collected BS data for experiments' participants. The BS values range from 143 to 241. In phase-1, where the training of BS metadata is taken from the lower end of the spectrum (BS=176), we observe a general trend of dropping classification accuracies as we move towards the higher end of the spectrum (BS=236), with the overall (for all fine-grained activities) accuracy dropping as low as 65% (Figure 2(c)). This is because different cameras (upper and lower body cameras) get different spatial coverage of the participants (sometimes no inclusion of the participants' image) across the training and testing phase in ACS. This implies that BS shows high impact on the metadata and activity classification. This is especially true for the sitting activity, where the upper body and lower body camera captures different volume dimensions if we abruptly

change BS value from the training to the testing phase. In this case, the classification accuracy falls to about 40%.

However, in phase-2, our classification model achieves an overall average accuracy of 97% whereby the training set included activity data corresponding to all BS values except the test BS value. This implies that if the machine learning model is trained with a representative set of data distributed uniformly over the whole BS spectrum, it is able to accurately classify activities for any data points throughout the spectrum.

4.2 Activity Classification

4.2.1 Experimental Setup

The experimental setup was the same as 4.1.1, with the exception that the model is trained to classify between a basic activity set S = {stand, sit, move00, move01, move10, move11}, where the definitions of the movement activities (moveXY) are given in Table 2.

Table 2: Movement Activity Definition

	Coverage = Low	**Coverage = High**
Speed = Low	move00	move01
Speed = High	move10	move11

The $Coverage \in \{Low, High\}$ refers to the spatial coverage of the participant in the physical space and the $Speed \in \{Low, High\}$ refers to the speed of movement and limb actions [1]. For each activity listed, 4 minutes of metadata was recorded for 5 people across the BS spectrum wearing the same color shirt (SC=RED), since earlier experiments in 4.1.2 have shown that body structure has a significant impact on activity classification accuracy.

4.2.2 Results and Discussion

The trained model is tested on random activity sessions' metadata, that is not seen during training. An (i, j) entry in the Table 3 denotes the percentage of activity samples corresponding to activity i that are classified as activity j. Based on the results, the following observations are made: 1) our model performs very well in classifying Stand and Sit, 2) classification between different movements is harder due to the minor variation in the impact of the different types of movement on metadata. However, we can still achieve 88% accuracy for move01 and 78% accuracy for move10, and 3) the major mis-classification occurs due to the variation in *Speed*; move00 is frequently mis-classified as move10, and move11 is frequently mis-classified as move01. In all cases, the *Coverage* dimension is classified correctly.

Table 3: Activity Classification Distribution (%)

	move00	*move01*	*move10*	*move11*	*stand*	*sit*
move00	47.50	0.00	32.50	0.00	17.50	2.50
move01	2.17	78.26	8.70	8.70	2.17	0.00
move10	7.14	0.00	88.10	4.76	0.00	0.00
move11	2.56	38.46	5.13	53.85	0.00	0.00
stand	0.00	0.00	4.17	0.00	95.83	0.00
sit	0.00	0.00	0.00	3.70	0.00	96.30

5. RELATED WORK

Niu et al in [7] attempted to detect human activity in a global 3D estimate by fusing estimates from multiple 2D sensors and observing people trajectories. However, this approach was designed towards detecting questionable activities for video surveillance as opposed to identifying particular activities in our case. Many others have applied machine learning models for activity classification, e.g., employment of hierarchical maximum entropy markov model to classify human activity in unstructured environments from Kinect Sensor data [10], unsupervised learning to detect unusual activity in video [12], combining supervised and unsupervised learning to model human behavior from mobile sensor [8] and [4]. However, no approaches derive features and information from application-system metadata to classify human activity, which is agnostic to application logic.

6. CONCLUSION

We propose a system that addresses the activity classification problem in 3DTI systems from an unique perspective, using time-series application-system metadata. With a few exceptions, our system is able to achieve high accuracy for classifying a basic set of human activities in a running 3DTI session. The detection process is fast (4 ms). Moreover, the system provides a base for constructing a classification system for coarse-grained activity detection. Due to the decoupling of ACS from the application-specific media analyses, our solution is generalized and can be adopted as a middleware to many 3DTI setup.

7. REFERENCES

[1] A. Arefin, Z. Huang, R. Rivas, S. Shi, P. Xie, K. Nahrstedt, W. Wu, G. Kurillo, and R. Bajcsy. Classification and analysis of 3D tele-immersive activities. In *IEEE Multimedia*, 2013.

[2] A. Arefin, R. Rivas, and K. Nahrstedt. Prioritized evolutionary optimization in open session management for 3D tele-immersion. In *Proc. of MMSys*, 2013.

[3] B. Boser, I. M. Guyon, and V. N. Vapnik. A training algorithm for optimal margin classifiers. In *Proc. of COLT*, 1992.

[4] C. Chen and K. Grauman. Efficient activity detection with max-subgraph search. In *Proc. of CVPR*, 2012.

[5] G. Kurillo, M. Forte, and R. Bajcsy. Tele-immersive 3D collaborative environment for cyberarchaeology. In *Proc. of CVPR*, 2010.

[6] A. Maimone and H. Fuchs. Real-time volumetric 3D capture of room-sized scenes for telepresence. In *Proc. of 3DTV-CON*, 2012.

[7] W. Niu, J. Long, D. Han, and Y.-F. Wang. Human activity detection and recognition for video surveillance. In *Proc. of ICME*, 2004.

[8] D. Peebles, H. Lu, N. D. Lane, T. Choudhury, and A. T. Campbell. Community-guided learning: Exploiting mobile sensor users to model human behavior. In *Proc. of AAAI*, 2010.

[9] N. Rizzolo and D. Roth. Learning based java for rapid development of NLP systems. http://bit.ly/13J1kuq, 2010.

[10] J. Sung, C. Ponce, B. Selman, and A. Saxena. Human activity detection from RGDB images. In *Proc. of AAAI workshop on PAIR*, 2011.

[11] Z. Zhang, D. Chu, J. Qiu, and T. Moscibroda. Demo: Sword fight with smartphones. In *Proc. of SenSys*, 2011.

[12] H. Zhong, J. Shi, and M. Visontai. Detecting unusual activity in video. In *Proc. of CVPR*, 2004.