HardenVR: Harassment Detection in Social Virtual Reality

Na Wang *
George Mason University

Jin Zhou †
George Mason University

Jie Li‡
Global Research Institute EPAM

Bo Han§
George Mason University

Fei Li¶
George Mason University

Songqing Chen‖
George Mason University

2024 IEEE Conference Virtual Reality and 3D User Interfaces (VR) | 979-8-3503-7402-5/24/$31.00 ©2024 IEEE | DOI: 10.1109/VR58804.2024.00033

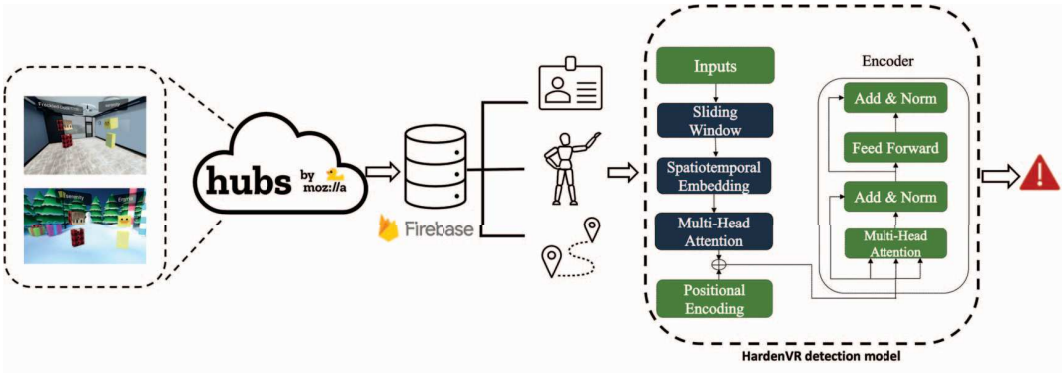


Figure 1: HardenVR: workflow of harassment detection in social VR

Abstract

Social Virtual Reality (VR) is regarded as one of the most popular VR applications since it transcends geographical barriers, allowing users to interact in simulated environments for various purposes. Despite its promising prospects, there is a growing concern about the harassment issue due to the immersive nature of social VR compared to other online social environments. Existing protections against harassment in social VR are highly limited in terms of practical effectiveness. The deficiency of studies toward understanding and preventing harassment in social VR further complicates the regulation and intervention efforts of social VR platforms in such situations. To address these challenges, we, in this paper, quantitatively investigate human interaction behaviors in social VR. More specifically, we first build a customized platform based on Mozilla Hubs, a popular social VR platform, to collect data about users' social interaction behaviors involving harassment instances. A subsequent analysis of the collected dataset SAHARA (Social interAction beHAviors in vR with hArassment) reveals that the task of online harassment detection in social VR is complicated since it depends on not only users' actions but also their spatial and temporal relationships. To accurately discern harassment, we propose a novel framework *HardenVR* (HArassment DEtectioN framework for social VR). As a context-aware harassment detection framework, *HardenVR* employs a transformer-based model to capture relative poses and learn users' hand actions in 6-DOF (Degree-of-Freedom). Meanwhile, multiple mechanisms, including the extra attention mechanism, distance-aware clustering method, and the sliding window, have been introduced into the model to handle challenges of data imbalance, over-fitting, and continuous detection. The design of *HardenVR* aims to achieve the balance between accuracy, efficiency, and cost-effectiveness for the task of harassment detection. As a starting point, *HardenVR* successfully learns pose information as the context to identify harassment and the experiment results show its detection accuracy as high as 98.26%.

Index Terms: Human-centered computing—Human computer interaction (HCI)—HCI design and evaluation methods—User studies; Human-centered computing—Human computer interaction (HCI)—Interaction paradigms—Virtual reality

1 Introduction

Thanks to recent advances in Virtual Reality (VR) technologies, the VR market is expected to reach $51.5 billion by 2030 with the growing adoption in areas such as entertainment, healthcare, automotive, and education [23]. As a type of shared VR experience overcoming geographical barriers, social VR can support users all over the world to meet and interact within a simulated environment for various purposes [15]. For instance, Bigscreen is a well-known social VR platform that allows users to watch movies in virtual theaters or play video games on big screens, alone or with friends [8]. Other popular social VR platforms include Horizon Worlds [36], Mozilla Hubs [43], Rec Room [52], and VRChat [61].

Similar to earlier social network platforms and applications, social VR fosters online harassment and thus inevitably causes a harmful or even detrimental impact on users' health [13, 19, 21]. Furthermore, compared to online harassment in forms of the text, image, or audio, harassment in the nascent digital virtual space demonstrates more unique characteristics due to its physicalized nature. Many reports and interviews show that the high level of immersion and rich embodied interactions, provided exclusively by social VR, induces users to be more sensitive and less tolerant towards inappropriate or unwanted behaviors [5, 6]. Correspondingly, users are more inclined to perceive them as the harassment.

Meanwhile, safety tools against the harassment in social VR are highly limited in terms of types available and usefulness. The most common tools available on well-known social VR platforms include

*e-mail: nwang4@gmu.edu
†e-mail: jzhou23@gmu.edu
‡e-mail: jasminejue@gmail.com
§e-mail: bohan@gmu.edu
¶e-mail: fli4@gmu.edu
‖e-mail: sqchen@gmu.edu

*First two authors contributed equally to this work.

979-8-3503-7402-5/24/$31.00 ©2024 IEEE
DOI 10.1109/VR58804.2024.00033

personal "boundary" or "bubble" which can be used to set up personal space to prevent others from being too close, muting which helps to turn off others' sound, and blocking that empower users to remove other avatars from the space [12, 36, 43, 52]. The mechanisms' result is not satisfactory, however, because the harassment perpetrators can escape from the scene after the harassment via the teleportation mechanism provided by platforms, and thus the victims fail to respond instantly to such behaviors. Besides, as a common practice of current social VR platforms, it is users' responsibility to report such incidents [27]. The outsourcing solution not only lays the extra burden on users but also results in recurring harm to victims [21]. Furthermore, there are very few studies on mitigating harassment in the context of social VR. In fact, by now there is no consensus on the definition of inappropriate or harassment behaviors in social VR [9]. Accordingly, the establishment of general guidelines is still in its infancy. The lack of standards leads to difficulties for social VR platforms and applications to regulate, moderate, and intervene in such situations.

In this paper, we aim to understand and detect online harassment in the context of social VR. The attempt to mitigate harassment in social VR, however, is obstructed by two main challenges. First, there is no such dataset about social interaction among VR users. Existing datasets are often collected for a single user performing specific actions in non-social VR environments. Second, making decisions about the harassment occurrence is highly contextual. Our analysis of the collected data using existing models for human activity recognition (HAR) reveals that the detection results depend on not only the harassment perpetrator's action but also the spatial and temporal relationships among users. Therefore, the detection results can only be co-determined by the action, movement, and positions of multiple avatars in VR.

To address these challenges, we first build a customized Mozilla Hubs instance on Amazon Web Services to collect users' social behavior data via VR hardware including controllers and headsets. During the data collection, we focus on common harassment behaviors in users' social interactions, such as slapping, punching, grabbing, pushing, and pinching. In order to gain a better understanding of harassment characteristics in social VR, we collect data about users' body positions, hand positions, and hand actions in the virtual world. To detect harassment in the collected data, there are multiple challenges, including the great amount of data generated during the continuous detection, harassment involving three or more persons, and datasets with highly skewed class proportions. To tackle these challenges and provide strong protection from harassment for users in social VR, we design and implement a framework, named as *HardenVR*, a *Har*rassement *de*tectio*n* framework for social VR applications. In the design of *HardenVR*, as shown in Figure 1, we propose the sliding window mechanism and a spatiotemporal embedding layer to reduce the amount of data for the subsequent detection and deal with harassment involving multiple persons. Then we employ a transformer-based model to further improve the accuracy and reduce the negative impact from the dataset imbalance. The extensive evaluations demonstrate that the innovative design can help HardenVR effectively learn and process temporal and spatial relationships among users so as to identify and determine harassment behaviors in the context of social VR. The results show HardenVR outperforms state-of-the-art approaches by a large margin, with harassment detection accuracy as high as 98.26%.

In summary, this paper makes the following major contributions:

- Through an IRB-approved user study, we collect a dataset, SAHARA, to facilitate future research on online harassment studies in social VR.
- Through the analysis of the dataset, we show that the harassment detection in VR is complex, depending on the combined effect of temporal and spatial relationships among users and their actions. Existing solutions for HAR produce poor outcomes in the task and thus cannot provide enough protection from the harassment for users.
- We design a harassment detection framework named HardenVR which employs a transformer-based model to determine harassment occurrences, along with three optimization techniques, including the extra attention mechanism, distance-aware clustering method, and the sliding window, to improve detection accuracy and efficiency. Evaluation results confirm HardenVR's great potential to identify harassment accurately.

The rest of the paper is organized as follows. Section 2 introduces the related work. Section 3 describes the data collection methodology, followed by the introduction of the SAHARA dataset in Section 4. We next present the proposed framework HardenVR for social harassment detection in Section 5. The evaluation results and discussion are presented in Section 6 and Section 7. Section 8 concludes the paper.

2 Related Work

Online Harassment. Online harassment generally refers to aggressive behaviors against other users on online social media [55]. In tandem with multiple generations of online social media, there is a rapid shift in techniques employed by harassment perpetrators, from text messages used on social websites such as Facebook and Twitter to visual images on Instagram. Due to its severe negative impacts on victims' physical and/or psychological health [26], online harassment has been studied over the years in multiple areas, such as psychology and social science. In the computer science community, researchers have also been developing various techniques to mitigate the threat. With regard to the case of textual harassment, a range of text features, such as topic similarity, sentiment, token frequencies, and punctuation, are mined for the purpose of automatic detection [17, 18, 41, 66]. Meanwhile, there are multiple state-of-the-art offensive text detectors available, such as Amazon Comprehend and Perspective API by Google [3, 24]. Similarly, several automatic detection approaches targeting image-based harassment have been developed [31, 60]. However, few studies have looked into the threat in social VR to date. The limited studies conduct and analyze users' interviews, focusing on either conceptual discussions or platform governance [9, 21, 54]. There is little research that attempts to focus on the technologies or approaches for the identification and mitigation of such behaviors.

Social VR Platforms. Social VR platforms allow users to create and customize avatars to represent themselves in the virtual world, creating a sense of presence and immersion. Avatars communicate with each other through speech, text, and gestures as people do in the real world. Social VR platforms can be used for various purposes. Correspondingly, there are several popular social VR platforms, each with its own unique features and user communities. For example, as one of the most popular social VR platforms, VRChat has a large user pool and a wide variety of virtual environments and activities to explore [61]. Rec Room offers a range of multiplayer games, such as paintball, quests, and escape rooms [52]. Mozilla Hubs [43], accessible from any device with a web browser, offers a paid option Hubs Cloud so users can run their own hub on private servers set up on AWS or DigitalOcean, for better security, or customize it for academic study [42].

Deep Learning in HAR. Human activity recognition (HAR) refers to the process of classifying human actions in a given period of time based on the collected data. Generally, the input data is organized in two forms: images or discrete measurements [35, 62]. The discrete measurements often include the measurements, made by various sensors equipped in routers, smartphones, or specialized devices, about acceleration, rotation speed, etc [2]. To recognize the human action from the data, early research employed traditional machine

Table 1: Set of Social Interaction Behaviors in Social VR and Corresponding Operations on Controllers

			Operations on controllers					
Behavior	Nature	Label	A	B	Thumbstick	Trigger	Grip	Notes
pinching[1]	Harassment	1	YES				YES	
pinching[2]	Harassment	1		YES			YES	
pinching[3]	Harassment	1			YES		YES	
punching	Harassment	2	YES	YES	YES	YES	YES	
pushing	Harassment	3						YES
slapping	Harassment	4			YES	YES		YES
thumbing down	Harassment	5				YES	YES	
pointing at others	Harassment	6	YES	YES	YES		YES	
grabbing others	Harassment	7				YES	YES	YES
hook finger	Harassment	8	YES				YES	
waving hands	Neural/Friendly	9						(depending on position)
hugging	Neural/Friendly	10						(depending on position or arm action)
shaking hands	Neural/Friendly	11						(depending on position)
"OK" gesture[1]	Neural/Friendly	12	YES			YES		
"OK" gesture[2]	Neural/Friendly	12		YES		YES		

learning methods such as support vector machine(SVM) or random forest to perform on the sensor data [28,29]. To avoid manual feature extraction, the method of deep learning is introduced into the area. The convolutional and Long-Short-Term-Memory (LSTM) layers are often used together since they help to better capture the spatial and temporal relationships in the data [50]. A set of variations on the method are proposed to improve further the performance, such as the introduction of the dilated convolution layer [64], the addition of the attention layer [44], the increase of the number of both layers [65], etc. However, there is disagreement concerning the minimum necessary depth of LSTM layers to reach state-of-the-art recognition performance [11, 14].

3 Harassment Study in Social VR

Several works have introduced datasets about users' behaviors in VR [32,34,39,40,45,47,49,51,58,63]. These studies, however, generally collect users' head, hand, or eye motion data while carrying out pre-assigned tasks or activities such as grabbing, walking, typing, and throwing within the VR environment. The collected data is often used to profile distinctive patterns for the purpose of identification or authentication. As such, these datasets cannot be used in the harassment mitigation study since the behaviors under examination are for individual tasks instead of social interaction. Furthermore, for the highly limited datasets involving users' social interaction in VR [38], there are no harassment instances within them. Thus, we set to first collect such a dataset by building a customized data collector in the open-source Mozilla Hubs VR platform, aiming to quantitatively study users' social interaction behaviors in VR and technically mitigate harassment in social VR. This section briefly describes the methodology of our data collection study of social behaviors, including harassment behaviors under investigation, study design, and participants' demographics.

3.1 Social Behaviors under Examination

To mitigate online harassment in social VR, it is necessary to gather sufficient data regarding social interaction behaviors related to harassment, which should be indicative of actual harassment. To this end, we select a set of typical behaviors for the study. A complete list of selected behaviors is presented in Table 1. The behaviors are labeled to facilitate the following analysis, and we use label 0 denoting users' other activities not listed in the table. The decision is made in accordance with two rules.

First, the behaviors should be supported by available VR devices. Currently, besides VR headsets, the controllers, usually held by users, are equipped with buttons and/or touchpads to allow tracking the movement of hands and fingers and facilitate users' interaction with others and the environment in the virtual world. However, unlike prosthetic hands, the ability of current controllers to simulate human hand gestures or actions is limited. For example, the middle finger, a well-known harassment gesture in Western culture, is not supported by the VR controller. Other examples include the "L" sign standing for the loser insult and the corona hand gesture, which has mixed meanings in different cultures, all of which are not supported. In our study, the VR device used in the experiment is Oculus Quest 2.

Second, to detect harassment from users' interactions, we select those behaviors involving harassment, friendly, and neural meanings in the U.S. culture [16, 21, 60]. To better simulate real-world scenarios, the harassment behaviors examined in the research are based on the extensive literature review, including published papers, laws, and opinions of professionals from diverse areas [7, 12, 48]. The Quest controller used in the experiment is presented in Figure 2. The selected social interaction behaviors and corresponding operations on controllers are demonstrated in Table 1. In this table, if a specific button/thumbstick/trigger is necessary for a behavior of interest, the corresponding value is Yes. It is noted for the thumbstick, Button A, and Button B on both controllers, involuntary touching and voluntary pressing generate the same result. Also, for the behaviors of pinching and "OK" gestures, there are three and two combinations of hands on the controller to perform the behavior, respectively.

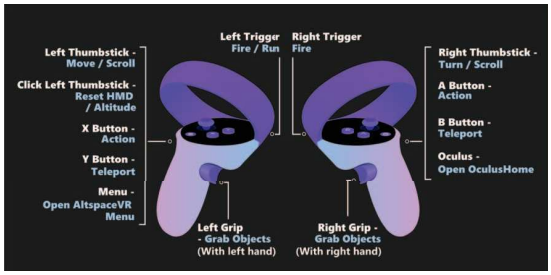


Figure 2: Oculus Quest Controllers with Various Buttons [37]

3.2 Study Design

Our data collection experiment has been approved by the George Mason University IRB (protocol #1881804). In the study, we asked each participant for informed consent at the beginning of the experiment. To better protect participants, they are allowed to exit at any

Table 2: Features Used in Our Study Grouped by Type

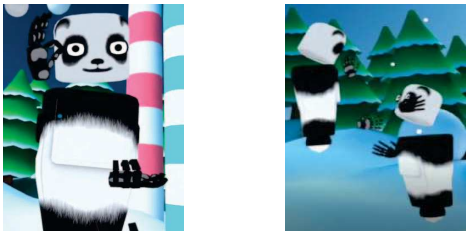
Category	Attribute	Description	Fields
Type I	ID	participant ID	1
	Timestamp	floating-point value	2
Type II	Avatar Position	avatar's position in 6-DOF of VR, in floating-point values	3-8
Type III	Hand Position	positions for both hands of avatar in 6-DOF of VR, in floating-point values	9-18 for left hand 19-28 for right hand
Type IV	Controller Action	avatar's actions on the controller buttons, in boolean values	29-36 for left 37-44 for right

point during experiments. The consent includes activity tracking and recordings in the virtual world and excludes audio recordings for privacy concerns. Besides, participants' usernames in the virtual world are removed in the following analysis.

The study is conducted with 14 groups, with three different group sizes of 2, 3, and 4. The setting of different group sizes is designed to mimic real-world social VR interactions in different scenarios where harassment is not necessarily one-to-one. One participant can be part of different groups for the experiments. Every group interacts on custom Mozilla Hubs via the Oculus Quest 2 bundle including a headset and two controllers.

An experiment starts with a brief introduction by researchers about the workflow. It is followed by the main part, the VR session, in which all group members are required to perform all social behaviors, as listed in Table 1, as many times as they want, in the random order they choose. At other times of the session, participants may communicate and interact as usual. The experiment ends with a demographic questionnaire. Each experiment takes no more than one hour.

The virtual environment used in the experiment is shown in Figure 3. The figure, including two screenshots, also demonstrates two studied social behaviors in our study, "OK" hand gesture as a neural behavior and slapping as a harassment behavior.



(a) OK gesture **(b)** Slapping

Figure 3: Screenshots of two studied social behaviors in a virtual room in social VR

3.3 Participants

A total of 22 participants took part in our data collection study. 64% of them are composed of men and only one person is left-handed. Participants are composed of students, including undergraduates and graduates, and staff members from a university with mixed academic backgrounds and majors. 15 of them wear glasses, 2 wear contact lenses, and 5 have no sightedness restrictions.

Regarding experience in virtual reality on a scale from novice to proficient, participants are intermediate experienced. Half of the participants have never used a head-mounted display (HMD) or only used it once or twice. 4 have experienced VR on the Mozilla Hubs platform while 3 of them have average experiences in VR but never used Mozilla Hubs. The study is conducted with 14 groups, with 9 groups with size 2, 3 groups with size 3, and 2 groups with size 4. Table 3 below summarizes the statistics about study participants.

Table 3: Demographics of Participants

Gender		Age		VR Experience		Group Size	
female:	8	20-25:	12	novice:	5	2:	9
male:	14	25-30:	8	beginner:	13	3:	3
		>30:	2	proficient:	4	4:	2

4 SAHARA Dataset

In this section, we present the collected dataset named SAHARA (Social interAction beHAviors in vR with hArassment), containing the recorded user movement traces and interaction activities in social VR on Mozilla Hubs via Oculus Quest 2. Both harassment behaviors and regular social interaction behaviors may be repeated in the experiment sessions. There are more than 20,000 sample behaviors in the dataset. The ratio between harassment behaviors and regular social interactions is around 0.2, suggesting that harassment data may be scarce compared to the normal user data for the interaction in social VR experiences. Our dataset is available under `https://cs.gmu.edu/~sqchen/open-access/SAHARA.html` for public research studies.

Each record in the SAHARA dataset has the same structure, with 44 fields to quantitatively describe the participant's activity in social VR. For the convenience of the following analysis in this paper, we divide all fields into 4 categories, including general information, avatar's position, hand position of avatars (for both left and right), and avatar's hand action on the controller (for both left and right). For privacy concerns, the sensitive personal information, including age, gender, and identity, is systematically expunged from the dataset. With regards to the future distribution, the anonymized data will be restricted to individuals directly affiliated with the research project and verified qualified researchers, despite its public nature. The detailed information for each category and data field can be found in Table 2.

Lastly, to obtain the labels or ground truth for the collected data to facilitate the subsequent analysis, all activity data is annotated by aligning with the corresponding recorded video to get labels with values 0-12.

5 HardenVR: Design and Implementation

In this section, we present our design of HardenVR, a framework attempting to mitigate harassment in social VR. Figure 1 presents the complete HardenVR workflow. To track and collect users' interactions in social VR, we employ the client-server design where the server is an open-source VR platform we customize and set up on AWS, while the client refers to those VR devices manipulated by participants, including VR headsets and hand controllers, as introduced previously. Then, the collected data, sent to and stored in the cloud database, is delivered to the detection engine for harassment detection. The detection engine employs a deep learning model based on the transformer to determine if a specific action is harassment. If a harassment behavior is detected, the framework will produce an alert. Next, we present details about each component and especially describe the detection model, which is the core of the entire HardenVR framework.

Table 4: Detection Results of Previous HAR Models

Model	Layers	HAR Accuracy	HAR F1 Score	SAHARA Accuracy	SAHARA F1 Score
LSTM Ensemble [25]	2	✗	0.926	62.14	0.6233
Deep-Res-Bidir-LSTM [67]	4	93.57	0.9354	77.89	0.7843
Shallow DeepConvLSTM [11]	1	✗	0.744	83.04	0.8435

5.1 Custom Build of VR Platform

As an open-source VR platform, Mozilla Hubs provides users easy access to set up customized servers and clients for the observation and analysis of social VR experiences for research purposes. For example, with the help of a customized build of Mozilla Hubs to track messages and users in a virtual workshop, researchers find how the virtual space design influences the proxemics and communication in such experiences [63]. Extending from the experiment setup, we design and implement a custom Hubs Cloud client to track users and collect their action and position data related to social interaction in VR experiences.

In our experiments, the server instance is deployed on the Amazon Web Services (AWS) private server to protect participants' privacy. Meanwhile, to reduce work overload on the server, the client is designed to keep real-time logging of users' actions and positions. Similar to the previous study [63], to guarantee the socket performance, we employ the batch collecting ticks and the bulk POST batch for each client. Moreover, the WebXR Device API is introduced into our JavaScript data logging code to track users' activity from multiple hardware sources, including headsets and controllers. Lastly, the collected data logs are recorded into a secure Firebase database in real-time via the encrypted channel.

5.2 Our Proposed Harassment Detection Model

In this section, we discuss the deep learning model we propose in HardenVR to detect the occurrence of harassment in social VR. Since the performance of the detection model is the key to the success of HardenVR, we propose our own model based on the transformer, motivated by performance comparison results performed on the SAHARA dataset from existing models.

5.2.1 Preliminary Analysis and Motivation

There is plenty of research on human activity recognition (HAR) [11,25,44,64,65,67]. Thus, we first investigate if the existing models could be used to recognize harassment behavior in social VR. As such, we perform a series of experiments to examine the performance of existing HAR models to identify and detect harassment in social VR. Table 4 lists the models used in the comparison experiments along with their key features and reports both their performance in HAR and harassment detection performance on the SAHARA dataset. It should be noted that all listed models have been trained on the same SAHARA dataset in our experiments to guarantee fairness in comparison.

The first model, LSTM (Long Short-Term Memory) Ensemble combines multiple LSTM learners into ensemble classifiers to tackle the challenge of problematic data [25]. The second model, Deep-Res-Bidir-LSTM employs both bidirectional and residual connection within the network architecture to enhance the recognition chance [67]. The third model, shallow DeepConvLSTM demonstrates that decreasing the layers of the LSTM is helpful to improve both the efficiency and performance [11].

In Table 4, for each row, the reported best performance of the three models above running on their existing HAR datasets, including accuracy and F1 score, is listed in the third and fourth columns, named "HAR Accuracy" and "HAR F1 Score", respectively. Three HAR datasets originally used include Skoda [56] for LSTM Ensemble, UCI [4] for deep Deep-Res-Bidir-LSTM, and RWHAR [57] for the shallow version of DeepConvLSTM, respectively. In contrast, their harassment detection performance on the SAHARA dataset is reported in the fifth and sixth columns, named "SAHARA Accuracy" and "SAHARA F1 Score", respectively.

From Table 4, we can observe the subpar performance of three HAR models to detect harassment in social VR. For the first and second models, their performance in harassment detection is worse than that in the original HAR case, showing a decrease of 30% and 15% in F1 scores, respectively. The second model Deep-Res-Bidir-LSTM reports the best performance in HAR with a 93.54% F1 score. When used in harassment detection, however, it demonstrates a remarkable decline in performance, producing a 78.43% F1 score only. The decrease may be attributed to their deficiency in taking into account social factors in social VR, such as the relative states of participants at the specific time, including users' hand actions and the spatial and temporal relationships. Furthermore, the best performance for harassment detection, achieved by the third model, is an 84.35% F1 score. We will demonstrate in the following that the performance is not good enough for the task of harassment detection since the accuracy performance of the detection model is the key to the success of HardenVR. We also test with traditional classifiers on SAHARA, such as random forest and Naive Bayes. Similar to previous studies, we retrieve regular data features, such as the average, median, and range, to train the classifiers with our dataset [30]. The best accuracy of the traditional classifiers is no higher than 60%, failing to identify harassment behaviors properly.

5.2.2 Model Design

The results of our preliminary analysis demonstrate that state-of-the-art learning models for HAR are not effective for harassment detection. Therefore, we aim to build a more effective model, as shown in Figure 1, to solve the problem. Next, we dissect the problem, discuss related challenges, and present our proposed solutions. Figure 4 sketches the architecture of our proposed model.

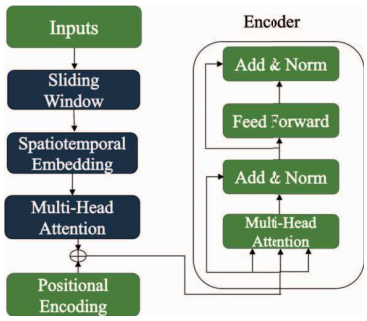


Figure 4: The architecture of the learning model for HardenVR (newly added parts in dark blue based on the design by [59])

Problem formulation and notation. In our sensor-based harassment detection, the input is a sequence of the concatenation of the last three types of sensor data, as shown in Table 2, including avatar position $\mathbf{p} \in \mathbb{R}^6$, avatar's hand position $\mathbf{h} \in \mathbb{R}^{20}$, and controller action $\mathbf{f} \in \mathbb{R}^{16}$. Without loss of generality, it is assumed that all sensors have the same sampling rate. Thus, a data sample would be $\mathbf{S} \in \mathbb{R}^{\mathbf{T} \times 42}$, representing a sequence of recorded measurements in the window of size $\mathbf{T}$. The corresponding label is defined as in

Table 1. In this way, the problem of harassment detection from sensor data can be formulated as a sequence-to-one problem via learning from the input as a sequential embedding of the raw sensor measurements while minimizing the empirical error.

Challenge 1: Data stream of large amount for continuous detection. Harassment detection must be done continuously. Continuous harassment detection involves measurement data accumulated over time, resulting in the problem of determining which data to choose for the following learning process. The length of the chosen data sequence can be regarded as the window size **T**. Generally, if the window size **T** is too small, the data sample $\mathbf{S} \in \mathbb{R}^{\mathbf{T} \times 42}$ may lack sufficient information for behavior analysis and harassment detection, producing results with low accuracy. Moreover, the high-granularity analysis may consume a great amount of computing resources and bring on significant degradation in terms of efficiency. On the other hand, a longer time duration for the window involves the possibility of multiple-behavior aggregation, resulting in erroneous results.

Proposed solution. To strike a balance between accuracy, efficiency, and computing cost, we propose the sliding window mechanism to be integrated into our model to perform on the raw data inputs. Clearly, the trade-off can be balanced by tuning the window size **T**, as we shall test with variable window sizes in training for better performance in the evaluation.

Challenge 2: Harassment in a group of three or more people. In social VR experiences, harassment does not always take place between two persons in a one-on-one setting. They can also occur in the one-to-many or many-to-many format. To tentatively study the scenarios, we collect movement and activity data of participants in such scenarios involving three and four persons, as shown in Table 3. Considering the fact that for a specific social VR session, not everyone in the group would be participating in the harassment, either as perpetrators or victims. Therefore, it would be more efficient to separate avatars involved in the harassment from others before feeding the collected data to the model. Furthermore, as shown in Table 2, each data entry has a high dimension of 44. The great amount of irrelevant data about participants' movement and activity constitutes noisy data, which not only results in over-fitting but also makes the subsequent training phase extremely time-consuming.

Proposed solution. To increase the training efficiency and accuracy for the output, we introduce the spatiotemporal embedding after the sliding window. The core of the component is a clustering process based on density-based spatial clustering of applications with noise (DBSCAN) [20], which helps to regroup the avatars in the specific session based on the position similarity. If the group size is greater than or equal to three, the approach significantly decreases the computation overhead by eliminating computing and comparing distances for all pairs of points on movement traces for all group members. In this way, the mechanism successfully accelerates the training process. In practice, the time complexity is reduced from $O\left(n^2\right)$ to $O\left(n \log n\right)$ in most cases.

Challenge 3: Negative impact from over-fitting. In the pilot study, we observe the negative impact of the over-fitting problem is not negligible on the performance of HardenVR. The problem can be attributed to the nature of the dataset for harassment detection. More specifically, our measurement study, introduced in Section 3, is designed to imitate real-world communication involving harassment in the context of social VR. Therefore, the occurrence of harassment behaviors is rare compared to normal social interaction in a specific session, which means the collected dataset is naturally imbalanced. In our collected SAHARA dataset, the ratio between harassment behaviors and regular social interactions is around 1:5. Besides, the high dimension of data entry, as presented in Section 4, contributes to the problem of over-fitting to some extent. To summarize, the imbalance nature and high-dimension of data lead to the over-fitting problem, which results in the incompetence of the model to detect harassment in real-world problems.

Proposed solution. To address this problem, we add a multi-head attention (MHA) block to the detection engine and also apply the regularization technique. The first mechanism, MHA block [59], with multiple attention functions in parallel within it, allows the model to jointly consider and learn information from data in a more comprehensive way to prevent over-fitting. The second mechanism, regularization, as a common countermeasure to prevent over-fitting [22, 33], is introduced to eliminate possible factors with less impact on the detection outcomes while emphasizing other features to increase the accuracy. Furthermore, L1-regularization is employed here because of its better performance for multiple features.

Challenge 4: Accuracy required by real-world applications. Protection from harassment in the context of social VR has long been demanded in the nascent digital space. The necessity is frequently validated by users and researchers who report or examine the negative or even destructive effects of online harassment on victims' psychological health [5,6,9,10,21]. In turn, the request has a demand for the accurate detection of harassment in social VR. The preliminary analysis results shown in Table 4 demonstrate LSTM-based models obtain mediocre results in the task. Existing works have tried to improve the performance of the LSTM model by combining it with a Convolutional Neural Network (CNN) so as to extract features more efficiently [46, 65]. However, this revision would result in a significant increase in model size.

Proposed solution. To adapt to the wider range of real-world scenarios for harassment detection, such as web browser plug-ins or end-user devices, while maintaining high performance, our choice of the deep learning model, as the main structure of HardenVR detection engine, is revised and enhanced on the basis of the transformer design by Vaswani et al. [59]. More specifically, as discussed previously, since harassment behaviors are rare compared to normal ones in social VR experiences [21], the collected dataset is naturally imbalanced. To accommodate the feature, we adopt the encoder-only in our model instead of the encoder-decoder combination in the original design, since we empirically validate the modification is able to extract features more efficiently for the imbalanced data. Also, compared to an LSTM model, the encoder better captures the interaction relationships between users and their actions. On the other hand, the encode-only design is more compact to be implemented and deployed in mobile applications for social VR.

5.2.3 Implementation

The core of HardenVR is implemented in PyTorch [1]. The MHA block proceeding the encoder employs twelve heads (parallel attention layers). The transformer encoder consists of one layer which contains another twelve-head MHA block within it. We train the model using standard cross-entropy loss and ADAM optimizer, with a batch size of 64, on the computing machines with NVIDIA A100-SXM GPU.

6 Evaluation

In this section, we conduct a series of experiments to evaluate the harassment-detection performance of HardenVR. The experiments are performed for both the entire SAHARA dataset and participating groups of different sizes. For the group results, we use **GroupS-N** to represent the N^{th} group of size S in the remainder of the paper. In particular, we aim to answer the following research questions (RQs):

- **RQ1:** How does HardenVR perform compared with current state-of-the-art approaches?
- **RQ2:** How does different types of sensor data in SAHARA complement each other in detecting harassment?
- **RQ3:** How is the parameter sensitivity of the HardenVR model?

Table 5: Performance comparison for harassment detection

Model	Accuracy	Precision	Recall	F1 Score	AUROC
LSTM Ensemble [25]	0.6214	0.6217	0.6247	0.6233	0.6906
Deep-Res-Bidir-LSTM [67]	0.7789	0.7894	0.7792	0.7843	0.8514
Shallow DeepConvLSTM [11]	0.8304	0.8395	0.8475	0.8435	0.9002
LSTM-ST	0.8969	0.8975	0.8972	0.8973	0.9745
HardenVR	**0.9826**	**0.9828**	**0.9828**	**0.9828**	**0.9933**

Table 6: Performance comparison of the models for both the general and all participating groups of different sizes *(GroupS-N: the N^{th} group of size S)*

	HardenVR					LSTM-ST				
	Accuracy	Precision	Recall	F1 Score	AUROC	Accuracy	Precision	Recall	F1 score	AUROC
General	**0.9826**	0.9828	0.9828	0.9828	0.9933	**0.8969**	0.8975	0.8972	0.8973	0.9745
Group2-1	0.9810	0.9812	0.9813	0.9812	0.9912	**0.9046**	0.9054	0.9022	0.9038	0.9872
Group2-2	0.9805	0.9804	0.9807	0.9805	0.9909	0.8924	0.9083	0.9074	0.9078	0.9713
Group2-3	0.9829	0.9828	0.9829	0.9828	0.9936	0.8978	**0.9241**	**0.9216**	**0.9228**	0.9782
Group2-4	0.9827	0.9828	0.9830	0.9829	0.9935	0.9005	0.9020	0.9004	0.9012	0.9804
Group2-5	0.9828	0.9830	0.9829	0.9829	0.9935	0.8948	0.8911	0.8977	0.8944	0.9726
Group2-6	0.9833	0.9832	0.9834	0.9833	0.9939	0.8990	0.9238	0.9034	0.9135	**0.9875**
Group2-7	0.9831	0.9834	0.9833	0.9833	0.9937	0.9026	0.9059	0.9031	0.9045	0.9802
Group2-8	0.9822	0.9824	0.9826	0.9825	0.9929	0.8994	0.9002	0.8995	0.8998	0.9731
Group2-9	**0.9842**	**0.9849**	**0.9844**	**0.9846**	**0.9953**	0.8965	0.8898	0.8902	0.8900	0.9736
Group3-1	**0.9838**	**0.9839**	**0.9837**	**0.9838**	**0.9946**	0.8973	0.9033	0.9014	0.9023	0.9788
Group3-2	0.9821	0.9822	0.9824	0.9823	0.9926	**0.9004**	**0.9086**	**0.9060**	**0.9073**	**0.9871**
Group3-3	0.9825	0.9827	0.9824	0.9825	0.9931	0.8946	0.8988	0.8942	0.8965	0.9719
Group4-1	0.9823	0.9824	0.9823	0.9823	0.9930	**0.8998**	0.9016	0.9003	0.9009	0.9737
Group4-2	**0.9835**	**0.9839**	**0.9837**	**0.9838**	**0.9942**	0.8980	**0.9031**	**0.9026**	**0.9028**	**0.9742**

- **RQ4:** How effective is our innovative design for improving the detection performance?

6.1 Experimental Settings

Data pre-processing. For the data pre-processing, we divide each group data into multiple subgroups based on the participant ID, as introduced in Section 4, so that each subgroup contains the complete activity information for a participant in the session. Meanwhile, we also remove data noise generated by involuntary teleportation, a feature that is provided in social VR. According to the post-experiment survey, participants may activate the feature unintentionally because of their unfamiliarity with social VR platforms, disrupting ongoing interaction with others.

Evaluation metrics. Following the common practices, we adopt multiple metrics to evaluate the performance of HardenVR, including accuracy, precision, recall, F1 score, and the area under the receiver operating characteristic curve (AUROC).

The precision metric measures what proportion of harassment identifications are true harassment behaviors while recall measures the ratio of actual harassment behaviors being correctly detected as harassment. Depending on the diverse preferences of harassment detection services in practice, the evaluation result may vary. Given our goal of protecting users from harassment in social VR, we consider a model with high recall to be more desirable than one with high precision.

Baselines approaches. We compare HardenVR with four state-of-the-art approaches: LSTM Ensemble [25], Deep-Res-Bidir-LSTM [67], Shallow DeepConvLSTM [11], and our own LSTM model incorporating spatiotemporal embedding layer as shown in Figure 4, named LSTM-ST.

6.2 Comparisons with Baseline Methods (RQ1)

Table 5 shows the performance of our model compared to the baselines. Our proposed HardenVR outperforms all baselines in all five metrics, demonstrating the effectiveness of our framework in capturing the complex spatial and temporal relationships among users in social VR. Specifically, on the SAHARA dataset, the gaps between HardenVR and LSTM Ensemble, Deep-Res-Bidir-LSTM, Shallow DeepConvLSTM, and LSTM-ST are extended to 36.12%, 20.37%, 15.22% and 8.57%, respectively. In addition to accuracy, it achieves the best precision and recall, 98.28% and 98.28%, respectively, implying the framework is able to protect users from potential harassment without minimally interfering with user experiences due to incorrect alerts.

With regards to the detection efficiency, in our experimental setup with NVIDIA A100 GPU and 16 GB memory, the computing speed is not as good as originally expected. For example, when the sliding window size is set to 3 seconds, the inference time for HardenVR is about 4 seconds. Improving the inference speed may involve a better trade-off between the window size, optimized attention mechanism, and better hardware, which we plan to explore further in future work.

In addition, we train personalized models for each participating group to compare the performance of HardenVR with LSTM-ST which achieves the second-best performance on the entire SAHARA dataset. The results of each model for both the general and all personalized models are presented in Table 6.

According to the table, HardenVR performs significantly better than LSTM-ST for all groups of different sizes for all metrics, which fully demonstrates the transformer-based model learns from users' movement and action data better than LSTM-based models. On the other hand, the results show, for both approaches, the personalized models outperform the general model in most cases. This implies that the variations between groups in collected data, including users' movement and hand action patterns, might be large enough so that the information learned from the entire dataset is not sufficient to identify target behaviors for new data.

6.3 Sensor Importance Comparison (RQ2)

In this subsection, we investigate which aspects of the users' activity data contribute the most to harassment detection in social VR. As presented in Section 4, our study collects three types of sensor data to describe users' interaction, including Type II (Avatar Position), Type III (Hand Position), and Type IV (Controller Action). Figure 5 presents the detection accuracy of our proposed HardenVR for

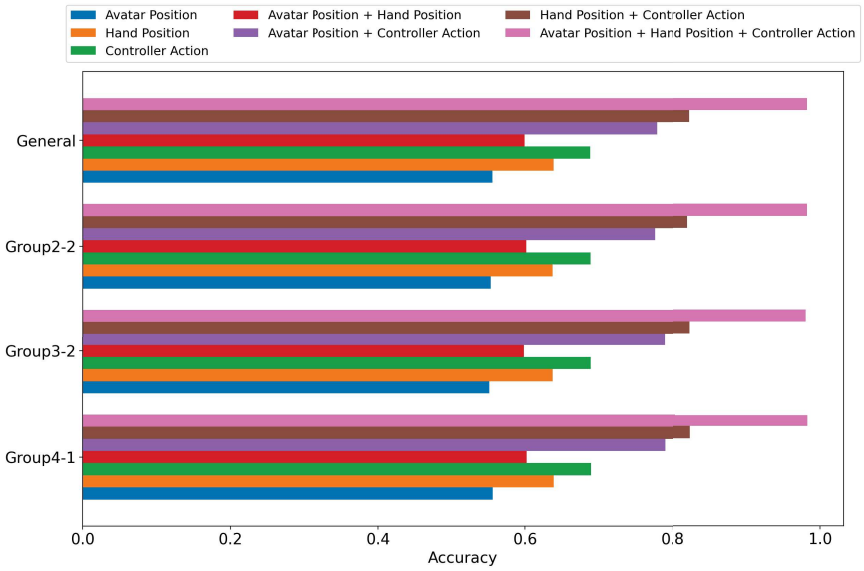


Figure 5: Accuracy comparison of different combinations of sensor data types *(GroupS-N: the Nth group of size S)*

all possible combinations of sensor data types, for both the general model and group models of different sizes. Moreover, we choose Group2-2, Group3-2, Group4-1 to demonstrate the result for different group sizes (GroupS-N: the Nth group of size S). Similar observations can be drawn in other groups.

The result shows that the model achieves the best accuracy when all three types of information are exploited for detection. This observation holds true for both the general model and all group models. Moreover, Controller Action data contributes more than the other two types of sensor data, Avatar Position and Hand Position, to the harassment detection accuracy. Specifically, among all seven combinations of data inputs, all four combinations containing the controller action lead to better outcomes than the remaining three combinations. On the other hand, no single type of data dominates other types in the harassment detection task, because the model produces the best results by a wide margin when all three types of information are utilized.

The results are consistent with the previous analysis in Section 5, suggesting that harassment detection in social VR depends on not only the actions of the users but also the spatial and temporal relationships among users. The three types of sensor data together describe the user's interaction states with others. For example, by investigating evaluation results, we observe that the same behavior of shaking hands with others can be detected as distinct results, depending on the relative position of two target users. Specifically, if one user's hand is in successful contact with the hand of the other one, the behavior would be detected as shaking hands and thus non-harassment. However, if one user accidentally makes a move so that one user's hand falls on the other's face, the behavior is then detected as the harassment of slapping. As a result, all three different aspects of users' activity are necessary for harassment detection. As a result, removing any type of data from the input results in a significant decline in accuracy.

6.4 Robustness Evaluation (RQ3)

The effect of window size, in terms of accuracy, efficiency, and computing cost, is discussed in Section 5.2.2. With this motivation in mind, we conduct a set of experiments with varied sliding window size **T** to investigate the effect of window size and perform the parameter sensitivity analysis. Figure 6 demonstrates the performance of HardenVR employing different window sizes with values chosen from the set $[5, 10, 15, 20, 25, 30, 35]$.

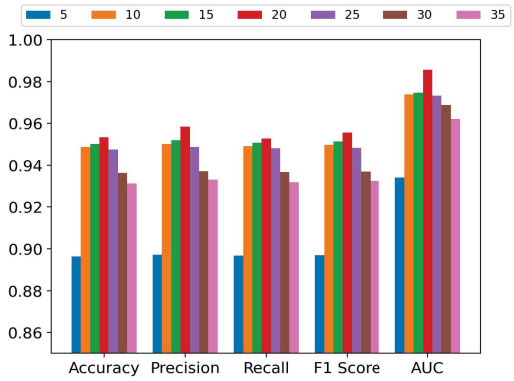


Figure 6: Performance comparison of HardenVR for different window sizes with values in [5, 10, 15, 20, 25, 30, 35]

The figure shows that the best performance is achieved when the window size is set to 20. Once the window size increases to 25 or decreases to 15, the accuracy declines by 0.34% and 0.63%, respectively. Furthermore, a similar trend emerges where the accuracy degradation occurs when the window size continues to grow larger than 25 or shrink smaller than 15. The similar performance declines, along with the change in the window size, appear for the other four metrics as shown in Figure 6.

The degraded performance can be explained by the lack of sufficient information for the model to learn when the window size is too small, or the mixture of behaviors to be hardly recognized when the window size is too big.

Table 7: Performance comparison with the ablation model for both the general and all groups of different sizes *(GroupS-N: the N^{th} group of size S)*

	HardenVR					HardenVR-NM				
	Accuracy	Precision	Recall	F1 Score	AUROC	Accuracy	Precision	Recall	F1 score	AUROC
General	**0.9826**	0.9828	0.9828	0.9828	0.9933	0.9534	0.9584	0.9527	0.9555	0.9856
Group2-1	**0.9810**	0.9812	0.9813	0.9812	0.9912	0.9453	0.9459	0.9432	0.9445	0.9845
Group2-2	**0.9805**	0.9804	0.9807	0.9805	0.9909	0.9690	0.9645	0.9679	0.9661	0.9889
Group2-3	**0.9829**	0.9828	0.9829	0.9828	0.9936	0.9577	0.9575	0.9584	0.9579	0.9862
Group2-4	**0.9827**	0.9828	0.9830	0.9829	0.9935	0.9561	0.9554	0.9560	0.9556	0.9860
Group2-5	**0.9828**	0.9830	0.9829	0.9829	0.9935	0.9530	0.9532	0.9529	0.9530	0.9849
Group2-6	**0.9833**	0.9832	0.9834	0.9833	0.9939	0.9526	0.9528	0.9526	0.9526	0.9843
Group2-7	**0.9831**	0.9834	0.9833	0.9833	0.9937	0.9532	0.9504	0.9538	0.9538	0.9832
Group2-8	**0.9822**	0.9824	0.9826	0.9825	0.9929	0.9588	0.9579	0.9592	0.9585	0.9873
Group2-9	**0.9842**	0.9849	0.9844	0.9846	0.9953	0.9556	0.9568	0.9562	0.9564	0.9848
Group3-1	**0.9838**	0.9839	0.9837	0.9838	0.9946	0.9613	0.9605	0.9615	0.9609	0.9882
Group3-2	**0.9821**	0.9822	0.9824	0.9823	0.9926	0.9489	0.9495	0.9492	0.9493	0.9821
Group3-3	**0.9825**	0.9827	0.9824	0.9825	0.9931	0.9568	0.9562	0.9558	0.9559	0.9855
Group4-1	**0.9823**	0.9824	0.9823	0.9823	0.9930	0.9572	0.9564	0.9569	0.9566	0.9875
Group4-2	**0.9835**	0.9839	0.9837	0.9838	0.9942	0.9520	0.9523	0.9518	0.9520	0.9251

6.5 Ablation Study (RQ4)

To evaluate the effectiveness of the HardenVR model design, we eliminate the MHA (Multi-Head Attention) block outside the encoder to observe the impact. To ensure fairness, except for the specified component, all other experiment settings are kept the same.

- **HardenVR-NM:** The MHA block outside the encoder is removed and the other components remain the same as shown in Figure 4.
- **HardenVR:** All components remain the same as shown in Figure 4.

The ablation test results are presented in Table 7. For both general and group models, the component of MHA plays an indispensable role in harassment detection. The performance is severely impacted by the ablation of components: for the general model, accuracy drops to 95.34% from 98.26%; for the groups of sizes 2, 3, and 4, the accuracy decreases by 2.73%, 2.76%, and 2.83% on average. The performance degradation also happens in the other four listed metrics for both general and group models.

The performance decrease can be explained by the function of MHA, the main essence of the transformer model. MHA, or self-attention, is designed to focus on and analyze how relevant one user's activity is with respect to others in the experiment, in the representation of the attention vector. The interaction relationship among users is exactly the key to harassment detection. Thus, the removal of the component leads to the performance decline in varying degrees on the listed performance metrics.

7 Discussion

In this section, we discuss the potential improvements of our work. Our study represents the first effective attempt at reducing harassment in social VR. However, it is subject to several limitations.

Broadening of social VR operations. In this work, we focus on online harassment in the form of behaviors in social VR. The virtual behaviors, supported by the controller devices, are chosen because they simply represent the unique high immersion and physicalized nature of social VR. Besides, the behaviors under examination are widespread in daily life. On the other hand, social VR devices, for example, Oculus Quest2, actually provide support for hands-free operations with more flexibility and complexity, so as to imitate a wider range of human gestures and movements for various applications and scenarios. Due to the limitations of the Mozilla Hubs platform, they are not covered in this paper. As part of future work, we plan to design the user study to collect the sensor data or record users' activities related to hands-free operations and further address the corresponding challenges of more comprehensive harassment detection. Besides, in order to achieve better alignment with real-world harassment in social VR, we intend to employ a combination of simulation and real-world event observation methodologies for data collection and analysis in future work [53].

Generality beyond behavioral harassment. Harassment mitigation in social VR is of greater importance, particularly as more and more emerging VR applications are being deployed. Depending on the high diversity of users' experiences, contexts, and devices, the identification results and corresponding impacts could be uncertain and highly subjective. In this paper, we carefully navigate through different scenarios and focus on less controversial harassment behaviors. As a starting point, our framework accurately identifies harassment behaviors in social VR. In the future, we hope to expand to more complicated scenarios and set up a comprehensive framework to understand in-depth and handle harassment in various contexts such as text, image, audio, and video.

Additionally, to handle large-scale issues of social VR environments in the real world, we conceive, as part of our future work, a hierarchical design where each room will be individually monitored by an (AI) agent while the extra communication and coordination mechanism be designed among agents to protect all users from harassment. Meanwhile, we envision the development and incorporation of preventive measures within the interface design, with the aim of providing enhanced proactive protection for social VR users.

8 Conclusion

Our study provides a preview of what can be achieved in terms of using captured motion data to mitigate harassment within thriving social VR environments. We collected the SAHARA dataset of users' social interaction behaviors in social VR including hand actions and body movements, and proposed the innovative context-aware framework HardenVR to discern instances of harassment. In our best configuration, we have achieved a classification accuracy of 98.26%. These results suggest that harassment identification based on captured motion data is a promising approach for creating a better environment within social VR.

9 Acknowledgements

We thank the anonymous reviewers for their insightful comments and the participants of our user study for their valuable contributions. This work was supported in part by the U.S. NSF under Grants CNS-2007153 and CNS-2235049.

REFERENCES

[1] S. C. Adam Paszke, Sam Gross and G. Chanan. Pytorch. https://pytorch.org.

[2] R. Alazrai, A. Awad, A. Baha'A, M. Hababeh, and M. I. Daoud. A dataset for wi-fi-based human-to-human interaction recognition. *Data in brief*, 31:105668, 2020.

[3] Amazon. Amazon comprehend. https://aws.amazon.com/comprehend/.

[4] D. Anguita, A. Ghio, L. Oneto, X. Parra, J. L. Reyes-Ortiz, et al. A public domain dataset for human activity recognition using smartphones. In *Esann*, vol. 3, p. 3, 2013.

[5] BBC. Female avatar sexually assaulted in meta vr platform, campaigners say. https://www.bbc.com/news/technology-61573661.

[6] J. Belamire. My first virtual reality groping. https://medium.com/athena-talks/my-first-virtual-reality-sexual-assault-2330410b62ee.

[7] U. Bhattacharya, N. Rewkowski, A. Banerjee, P. Guhan, A. Bera, and D. Manocha. Text2gestures: A transformer-based network for generating emotive body gestures for virtual agents. In *2021 IEEE virtual reality and 3D user interfaces (VR)*, pp. 1–10. IEEE, 2021.

[8] Bigscreen. Bigscreen. https://www.bigscreenvr.com.

[9] L. Blackwell, N. Ellison, N. Elliott-Deflo, and R. Schwartz. Harassment in social virtual reality: Challenges for platform governance. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–25, 2019.

[10] L. Blackwell, N. Ellison, N. Elliott-Deflo, and R. Schwartz. Harassment in social vr: Implications for design. In *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pp. 854–855. IEEE, 2019.

[11] M. Bock, A. Hölzemann, M. Moeller, and K. Van Laerhoven. Improving deep learning for har with shallow lstms. In *2021 International Symposium on Wearable Computers*, pp. 7–12, 2021.

[12] A. Bönsch, S. Radke, H. Overath, L. M. Asché, J. Wendt, T. Vierjahn, U. Habel, and T. W. Kuhlen. Social vr: How personal space is affected by virtual agents' emotions. In *2018 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pp. 199–206. IEEE, 2018.

[13] M. Campbell, B. Spears, P. Slee, D. Butler, and S. Kift. Victims' perceptions of traditional and cyberbullying, and the psychosocial correlates of their victimisation. *Emotional and Behavioural Difficulties*, 17(3-4):389–401, 2012.

[14] K. Chen, D. Zhang, L. Yao, B. Guo, Z. Yu, and Y. Liu. Deep learning for sensor-based human activity recognition: Overview, challenges, and opportunities. *ACM Computing Surveys (CSUR)*, 54(4):1–40, 2021.

[15] R. Cheng, N. Wu, M. Varvello, S. Chen, and B. Han. Are we ready for metaverse? a measurement study of social virtual reality platforms. In *Proceedings of the 22nd ACM Internet Measurement Conference*, pp. 504–518, 2022.

[16] CNN. "ok" hand sign added to list of hate symbols for white supremacy. https://www.cbsnews.com/news/ok-symbol-hand-gesture-anti-defamation-league-bowl-cut-racism-hatred/, 2019.

[17] A. Dey, M. Jenamani, and J. J. Thakkar. Senti-n-gram: An n-gram lexicon for sentiment analysis. *Expert Systems with Applications*, 103:92–105, 2018.

[18] K. Dinakar, R. Reichart, and H. Lieberman. Modeling the detection of textual cyberbullying. In *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 5, pp. 11–17, 2011.

[19] M. Duggan. Online harassment 2017. 2017.

[20] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. KDD'96. AAAI Press, 1996.

[21] G. Freeman, S. Zamanifard, D. Maloney, and D. Acena. Disturbing the peace: Experiencing and mitigating emerging harassment in social virtual reality. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW1):1–30, 2022.

[22] B. Ghojogh and M. Crowley. The theory behind overfitting, cross validation, regularization, bagging, and boosting: tutorial. *arXiv preprint arXiv:1905.12787*, 2019.

[23] GlobalData. Virtual reality market size, share and trends analysis report by end-user type, product type and region, 2021-2030. https://www.globaldata.com/store/report/vr-market-analysis/.

[24] Google. Perspective api. https://www.perspectiveapi.com.

[25] Y. Guan and T. Plötz. Ensembles of deep lstm learners for activity recognition using wearables. *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies*, 1(2):1–28, 2017.

[26] C. N. Hase, S. B. Goldberg, D. Smith, A. Stuck, and J. Campain. Impacts of traditional bullying and cyberbullying on the mental health of middle school and high school students. *Psychology in the Schools*, 52(6):607–617, 2015.

[27] HeartMob. Heartmob. https://stories.righttobe.org.

[28] A. Jain and V. Kanhangad. Human activity classification in smartphones using accelerometer and gyroscope sensors. *IEEE Sensors Journal*, 18(3):1169–1177, 2017.

[29] N. Jalloul, F. Porée, G. Viardot, P. L'Hostis, and G. Carrault. Activity recognition using complex network analysis. *IEEE journal of biomedical and health informatics*, 22(4):989–1000, 2017.

[30] K. Kowsari, K. Jafari Meimandi, M. Heidarysafa, S. Mendu, L. Barnes, and D. Brown. Text classification algorithms: A survey. *Information*, 10(4):150, 2019.

[31] K. Kumari and J. P. Singh. Identification of cyberbullying on multimodal social media posts using genetic algorithm. *Transactions on Emerging Telecommunications Technologies*, 32(2):e3907, 2021.

[32] A. Kupin, B. Moeller, Y. Jiang, N. K. Banerjee, and S. Banerjee. Task-driven biometric authentication of users in virtual reality (vr) environments. In *MultiMedia Modeling: 25th International Conference, MMM 2019, Thessaloniki, Greece, January 8–11, 2019, Proceedings, Part I 25*, pp. 55–67. Springer, 2019.

[33] Z. Li, K. Kamnitsas, and B. Glocker. Analyzing overfitting under class imbalance in neural networks for image segmentation. *IEEE transactions on medical imaging*, 40(3):1065–1077, 2020.

[34] J. Liebers, M. Abdelaziz, L. Mecke, A. Saad, J. Auda, U. Gruenefeld, F. Alt, and S. Schneegass. Understanding user identification in virtual reality through behavioral biometrics and the effect of body normalization. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pp. 1–11, 2021.

[35] C.-Y. Lin, Y.-T. Liu, C.-Y. Lin, and T. K. Shih. Tcn aa: A wi fi based temporal convolution network for human to human interaction recognition with augmentation and attention. *arXiv preprint arXiv:2305.18211*, 2023.

[36] Meta. Horizon worlds. https://www.oculus.com/horizon-worlds/.

[37] Meta. Quest controller. https://learn.microsoft.com/en-us/windows/mixed-reality/altspace-vr/getting-started/oculus-controls.

[38] M. R. Miller, E. Han, C. DeVeaux, E. Jones, R. Chen, and J. N. Bailenson. A large-scale study of personal identifiability of virtual reality motion over time. *arXiv preprint arXiv:2303.01430*, 2023.

[39] M. R. Miller, F. Herrera, H. Jun, J. A. Landay, and J. N. Bailenson. Personal identifiability of user tracking data during observation of 360-degree vr video. *Scientific Reports*, 10(1):17404, 2020.

[40] A. G. Moore, R. P. McMahan, H. Dong, and N. Ruozzi. Personal identifiability and obfuscation of user tracking data from vr training sessions. In *2021 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 221–228. IEEE, 2021.

[41] M. Mozafari, R. Farahbakhsh, and N. Crespi. A bert-based transfer learning approach for hate speech detection in misc social media. In *International Conference on Complex Networks and Their Applications*, pp. 928–940. Springer, 2019.

[42] Mozilla. Hubs cloud. https://hubs.mozilla.com/docs/hubs-cloud-intro.html.

[43] Mozilla. Mozilla hubs. https://hubs.mozilla.com.

[44] V. S. Murahari and T. Plötz. On attention models for human activity recognition. In *Proceedings of the 2018 ACM international symposium on wearable computers*, pp. 100–103, 2018.

[45] T. Mustafa, R. Matovu, A. Serwadda, and N. Muirhead. Unsure how to authenticate on your vr headset? come on, use your head! In *Proceedings of the Fourth ACM International Workshop on Security and Privacy Analytics*, pp. 23–30, 2018.

[46] R. Mutegeki and D. S. Han. A cnn-lstm approach to human activity recognition. In *2020 international conference on artificial intelligence in information and communication (ICAIIC)*, pp. 362–366. IEEE, 2020.

[47] V. Nair, W. Guo, R. Wang, J. F. O'Brien, L. Rosenberg, and D. Song. Berkeley open extended reality recordings 2023 (boxrr-23): 4.7 million motion capture recordings from 105,852 extended reality device users. *arXiv preprint arXiv:2310.00430*, 2023.

[48] NewYorkGovernment. Sexual harassment prevention training. `https://www.ny.gov/sites/default/files/2023-04/SexualHarassmentPreventionTrainingSlides.pdf`, 2023.

[49] I. Olade, C. Fleming, and H.-N. Liang. Biomove: Biometric user identification from human kinesiological movements for virtual reality systems. *Sensors*, 20(10):2944, 2020.

[50] F. J. Ordóñez and D. Roggen. Deep convolutional and lstm recurrent neural networks for multimodal wearable activity recognition. *Sensors*, 16(1):115, 2016.

[51] K. Pfeuffer, M. J. Geiger, S. Prange, L. Mecke, D. Buschek, and F. Alt. Behavioural biometrics in vr: Identifying people from body motion and relations in virtual reality. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pp. 1–12, 2019.

[52] R. Room. Rec room. `https://recroom.com`.

[53] N. Sabri, B. Chen, A. Teoh, S. P. Dow, K. Vaccaro, and M. Elsherief. Challenges of moderating social virtual reality. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pp. 1–20, 2023.

[54] K. Shriram and R. Schwartz. All are welcome: Using vr ethnography to explore harassment behavior in immersive social virtual reality. In *2017 IEEE Virtual Reality (VR)*, pp. 225–226. IEEE, 2017.

[55] P. K. Smith, J. Mahdavi, M. Carvalho, S. Fisher, S. Russell, and N. Tippett. Cyberbullying: Its nature and impact in secondary school pupils. *Journal of child psychology and psychiatry*, 49(4):376–385, 2008.

[56] T. Stiefmeier, D. Roggen, G. Ogris, P. Lukowicz, and G. Tröster. Wearable activity tracking in car manufacturing. *IEEE Pervasive Computing*, 7(2):42–50, 2008.

[57] T. Sztyler and H. Stuckenschmidt. On-body localization of wearable devices: An investigation of position-aware activity recognition. In *2016 IEEE International Conference on Pervasive Computing and Communications (PerCom)*, pp. 1–9. IEEE, 2016.

[58] P. P. Tricomi, F. Nenna, L. Pajola, M. Conti, and L. Gamberi. You can't hide behind your headset: User profiling in augmented and virtual reality. *IEEE Access*, 11:9859–9875, 2023.

[59] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[60] N. Vishwamitra, H. Hu, F. Luo, and L. Cheng. Towards understanding and detecting cyberbullying in real-world images. In *2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA)*, 2021.

[61] VRChat. Vrchat. `https://hello.vrchat.com`.

[62] Z. Wang, K. Ying, J. Meng, J. Ning, and C. Bai. Human-to-human interaction detection. *arXiv preprint arXiv:2307.00464*, 2023.

[63] J. Williamson, J. Li, V. Vinayagamoorthy, D. A. Shamma, and P. Cesar. Proxemics and social interactions in an instrumented virtual reality workshop. In *Proceedings of the 2021 CHI conference on human factors in computing systems*, pp. 1–13, 2021.

[64] R. Xi, M. Hou, M. Fu, H. Qu, and D. Liu. Deep dilated convolution on multimodality time series for human activity recognition. In *2018 international joint conference on neural networks (IJCNN)*, pp. 1–8. IEEE, 2018.

[65] K. Xia, J. Huang, and H. Wang. Lstm-cnn architecture for human activity recognition. *IEEE Access*, 8:56855–56866, 2020.

[66] N. Yuvaraj, K. Srihari, G. Dhiman, K. Somasundaram, A. Sharma, S. Rajeskannan, M. Soni, G. S. Gaba, M. A. AlZain, and M. Masud. Nature-inspired-based approach for automated cyberbullying classification on multimedia social networking. *Mathematical Problems in Engineering*, 2021, 2021.

[67] Y. Zhao, R. Yang, G. Chevalier, X. Xu, and Z. Zhang. Deep residual bidir-lstm for human activity recognition using wearable sensors. *Mathematical Problems in Engineering*, 2018:1–13, 2018.