Going through the motions: AR/VR keylogging from user head motions

Carter Slocum, Yicheng Zhang, Nael Abu-Ghazaleh, Jiasi Chen

UC RIVERSIDE

UNIVERSITY OF MICHIGAN

What are virtual reality head-mounted displays?

A display worn on the head!

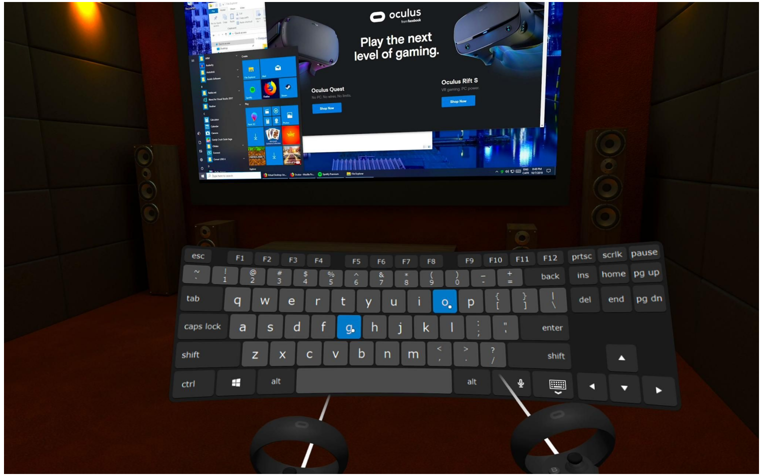
- Tracks the “pose” of the head.
- Displays 3D computer graphics so they appear to the user’s eyes as if they were actually there
- Untethered mobile device

Apple Vision Pro
(announced June 2023)

Applications:

- Games, public safety, workspaces, medicine, etc.
- Telepresence, socialization

Key Ideas


oculus
Play the next level of gaming.
Oculus Quest
Oculus Rift S
esc F1 F2 F3 F4 F5 F6 F7 F8 F9 F10 F11 F12 prtsc scrlk pause
back ins home pg up
tab q w e r t y u i o p del end pg dn
caps lock a s d f g h j k l enter
shift z x c v b n m shift
ctrl alt alt

Key Challenge:

What new ***privacy issues*** are raised by a user wearing a VR headset and ***entering sensitive information?***

Key Contribution:

We demonstrate an ***attack*** using the freely available ***head movement tracking data*** to ***infer what words*** a user is typing, by training a ML model on real user data traces.

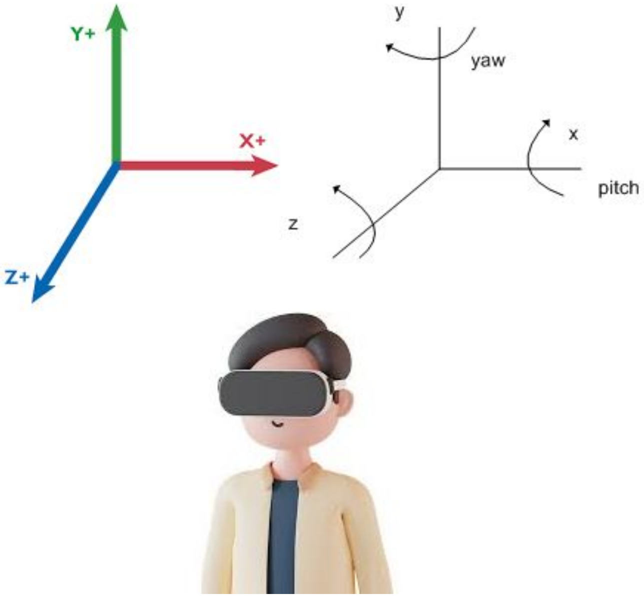
Example video of user typing in VR

- Observations
  - Users move their heads to see what they are typing in VR
  - Even if using hands for typing, the head is still involved

- Is it possible to accurately guess typed words from head movements?

https://drive.google.com/file/d/1hOGYe1dlrBI-8fcU1jPyT4YyZx48zp9B/view?usp=sharing

Background on VR head pose tracking



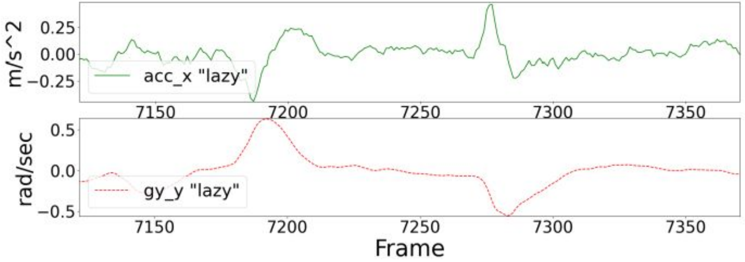
Six degrees of freedom tracking:

- Position (x,y,z) (m/sec^2)
- Rotation (yaw, pitch, roll) (rad/sec)

Sensed by headset’s gyroscope, accelerometer, and cameras.

Examples of user head pose traces

Is it possible to accurately guess typed words from head pose?



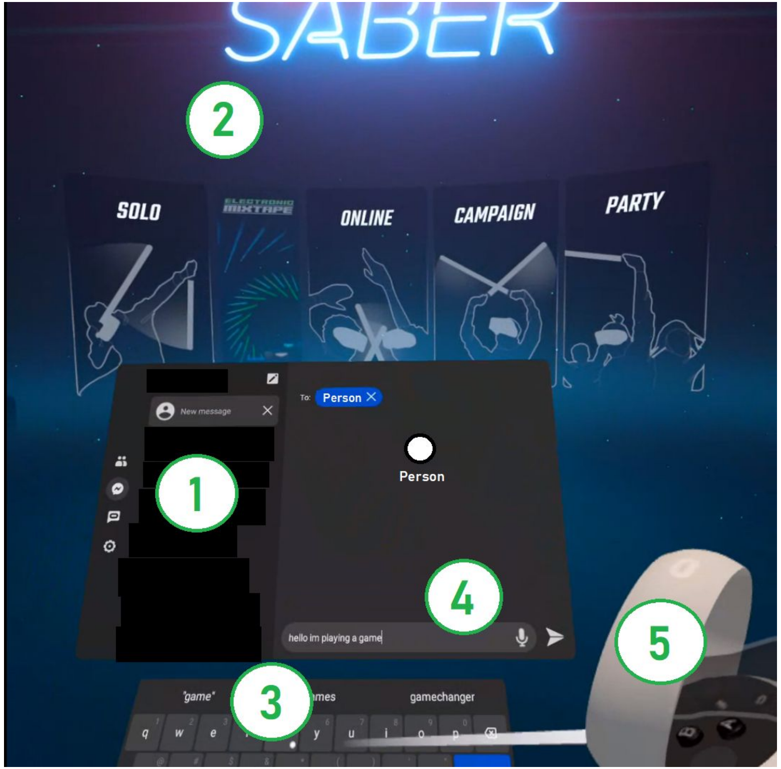
(a) Victim types the word “lazy”

Threat Model

Popular VR programming environments (e.g., Unity) provide all apps with real-time head pose

1) Malicious background app records head pose
2) User types sensitive data
3) Background app infers sensitive typed data

No special permissions needed by attacker



[redacted]

hello im playing a game

1) Foreground App
2) Background App
3) Virtual keyboard
4) Text Entry Field
5) Handheld Controller

Cross-modality attack: differences from related work

Head-Gaze Commit [1]		Hand Tracking [2]	
Input Method	Tracked Data	Input Method	Tracked Data
Head pose + button press	Head pose + button press	Hand pose	Hand pose

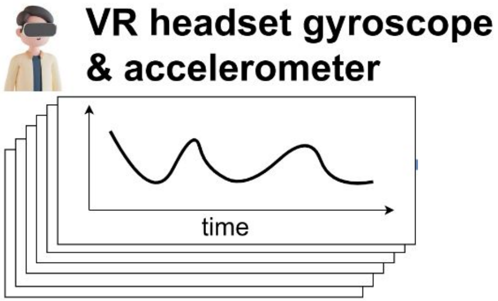
Going Through the Motions (Ours)

Input Method	Tracked Data
Hand pose	Head pose

Cross-Modal!
!=

[1] Shiqing Luo et al: Keystroke inference on mixed reality head mounted displays. *IEEE VR*, 2022.
[2] Ülkü Meteriz-Yıldıran et al: A keylogging inference attack on air-tapping keyboards in virtual environments. *IEEE VR*, 2022

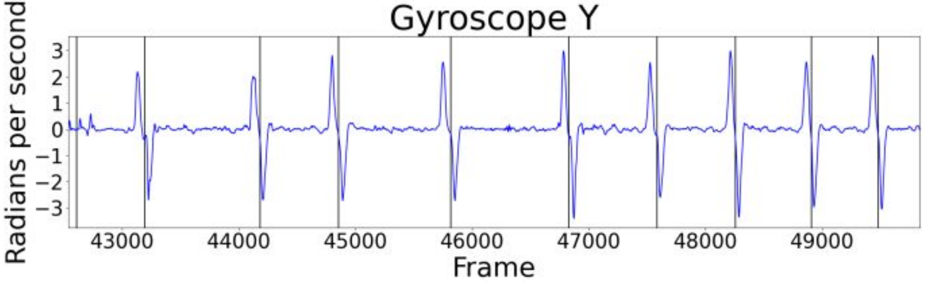
Overview of framework



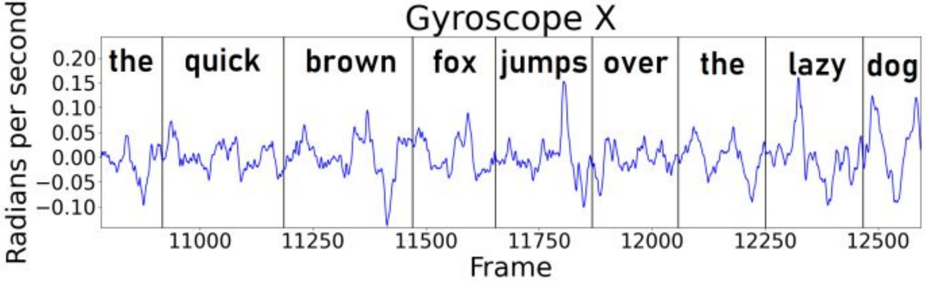
VR headset gyroscope & accelerometer

Challenge: Determining word boundaries

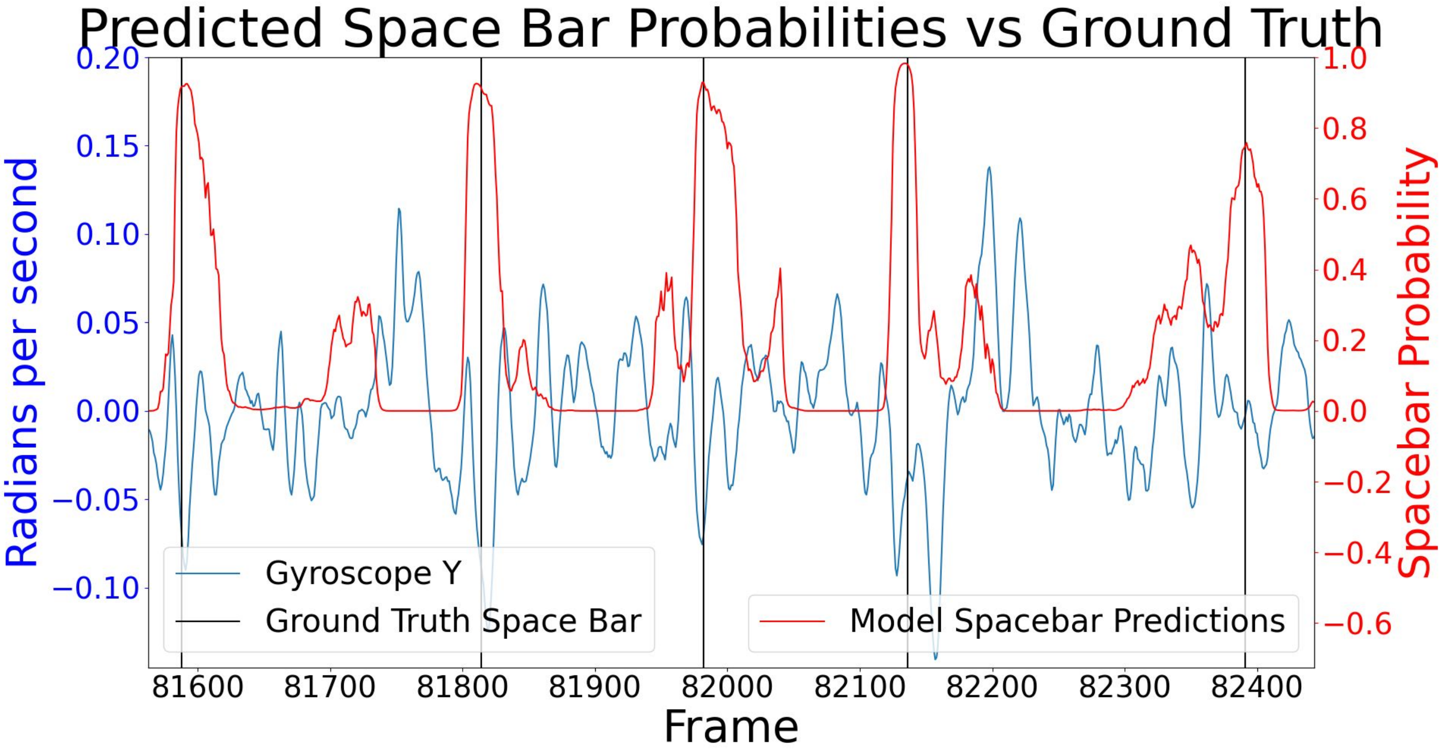
Easy case: words separated by “submit” button



Hard case: words separated by space bar



Idea: Can machine learning help determine word boundaries?



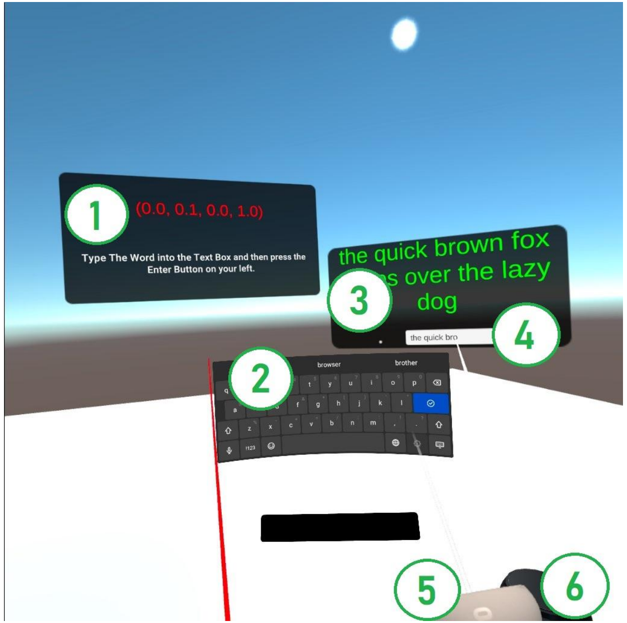
Segmenter design

Similar design also used for word classifier

Evaluation setup

- Data collection app
  (1) The background app records the headset sensor readings even it is not in focus
  (2) Foreground keyboard
  (3) Sentence prompt
  (4) Text input
  (5) System controller (untracked)
  (6) Background app controller (frozen)

- User study
  - Collected traces from ~20 participants
  - ~400 minutes of recorded head pose frames
  - 60+ unique words (common 2 and 6-letter words)
  - Over 600 total words typed



Type The Word into the Text Box and then press the Enter Button on your left.

the quick brown fox ...s over the lazy dog

the quick bro

[redacted]

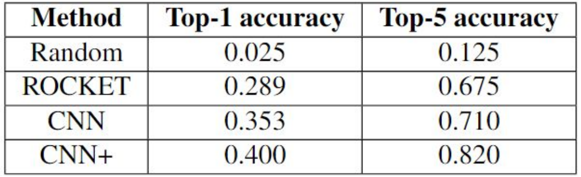
Evaluation setup: Machine learning models

Tried many machine learning methods

- Random: Random guess from the sets of words or character pairs
- KNN with time warping
- ROCKET: Popular time series library
- CNN, CNN+ (winner!)
  - Treat 6DoF pose over time as an image
    - 6xTime pixels per word sample
- Others
  - Various feature engineering tricks (1st order statistics_
  - Key timing correlation

Results: Word and character pair classifier

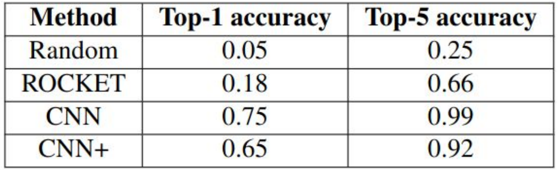
Word classification (all users)



Method	Top-1 accuracy	Top-5 accuracy
Random	0.025	0.125
ROCKET	0.289	0.675
CNN	0.353	0.710
CNN+	0.400	0.820

→ Word classifier significantly outperforms random guesses

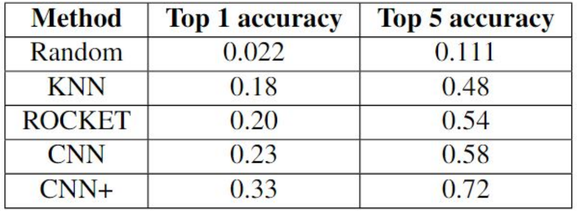
Word classification (personalized attack on single user)



Method	Top-1 accuracy	Top-5 accuracy
Random	0.05	0.25
ROCKET	0.18	0.66
CNN	0.75	0.99
CNN+	0.65	0.92

→ Individuals have highly repeated patterns of typing unique to themselves

Character pair classification (all users)

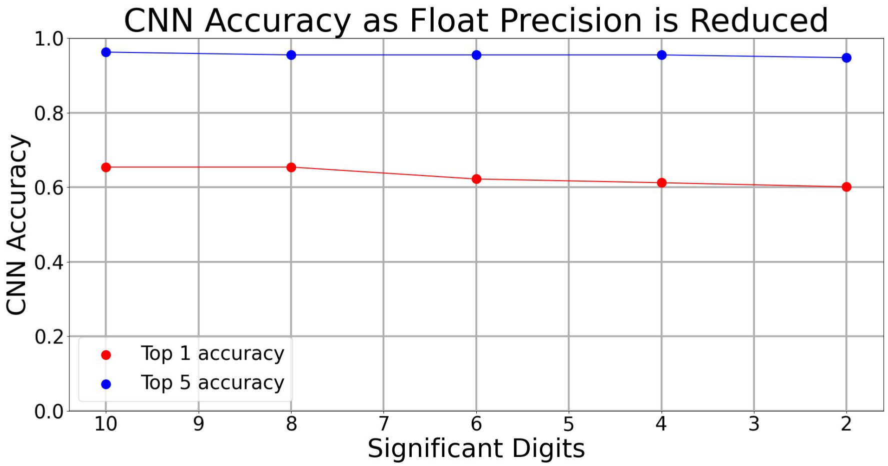


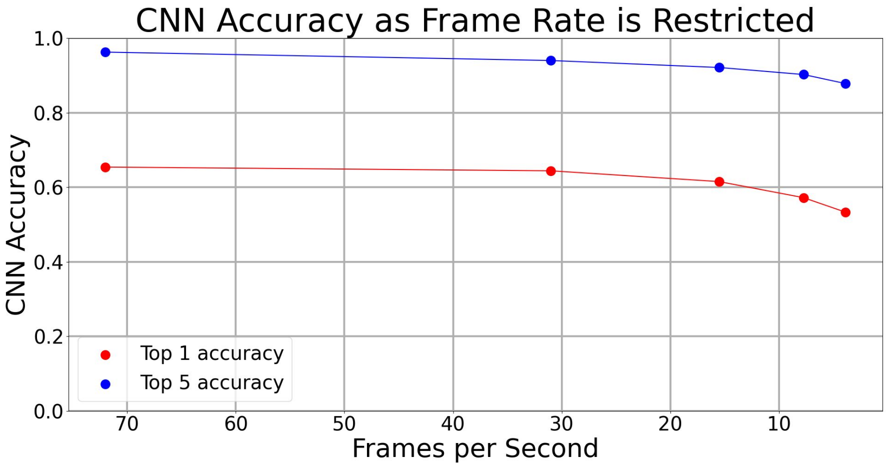
Method	Top 1 accuracy	Top 5 accuracy
Random	0.022	0.111
KNN	0.18	0.48
ROCKET	0.20	0.54
CNN	0.23	0.58
CNN+	0.33	0.72

→ Character pair classifier accuracy is lower, can still help reduce the search space for password cracker

See paper for more results.

Do simple mitigations work?





Not promising!
Better isolation between applications is needed

- Example: System renders a placeholder background while access to head tracking data is cut off.

Limitations

- Sensitive to cascading errors in the end-to-end attack
  - → Use other methods for word boundary estimation
- Need to update the word corpus in training dataset with new words
- Users' familiarity with VR affects their typing behavior
  - Some users naturally move their heads less

Thank you! Questions?

More XR Security from us:

Thursday, Aug 10th @ 1:30pm

“It’s All Fun and Games Until…”, Track 6