Demo abstract: Towards Real-time Rich Scene Analysis Using Vision-guided Wireless Vibrometry

Ziqi Wang, Ankur Sarker, Jason Wu, Derek Hua, Gaofeng Dong, Akash Deep Singh, Mani Srivastava
Electrical and Computer Engineering Department, University of California, Los Angeles
Los Angeles, California 90095, United States of America
wangzq312@g.ucla.edu,as4mz@virginia.edu,{jaysunwu,derekhua,gfdong,akashdeepsingh,mbs}@ucla.edu

ABSTRACT

Intelligent systems commonly employ vision sensors like cameras to analyze a scene. Recent work has proposed a wireless sensing technique, wireless vibrometry, to enrich the scene analysis generated by vision sensors. Wireless vibrometry employs wireless signals to sense subtle vibrations from the objects and infer their internal states. However, it is difficult for pure Radio-Frequency (RF) sensing systems to obtain objects' visual appearances (e.g., object types and locations), especially when an object is inactive. Thus, most existing wireless vibrometry systems assume that the number and the types of objects in the scene are known. The key to getting rid of these presumptions is to build a connection between wireless sensor time series and vision sensor images. We present Capricorn, a vision-guided wireless vibrometry system. In Capricorn, the object type information from vision sensors guides the wireless vibrometry system to select the most appropriate signal processing pipeline. The object tracking capability in computer vision also helps wireless systems efficiently detect and separate vibrations from multiple objects in real time.

CCS CONCEPTS

• **Human-centered computing** → *Ubiquitous and mobile computing systems and tools*; • **Computer systems organization** → *Real-time system architecture.*

ACM Reference Format:
Ziqi Wang, Ankur Sarker, Jason Wu, Derek Hua, Gaofeng Dong, Akash Deep Singh, Mani Srivastava. 2022. Demo abstract: Towards Real-time Rich Scene Analysis Using Vision-guided Wireless Vibrometry. In *The 20th ACM Conference on Embedded Networked Sensor Systems (SenSys '22), November 6–9, 2022, Boston, MA, USA.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3560905.3568060

1 INTRODUCTION

Real-world scenes are often complicated and rapid-changing, and intelligent systems must continuously analyze the scene surrounding them using sensory data to build situational awareness. For objects in a scene, some extrinsic properties like shape, location, and color can be easily acquired from vision sensors. In this domain, video-based scene analysis has achieved great success [1]. However, objects still possess some intrinsic properties invisible to vision sensors. Some of the intrinsic properties can be manifested as tiny movements, i.e., vibrations. For example, vibrations can tell the operating states of a machine for industry monitoring or the usage of appliances in a smart home. Previous works in the wireless sensing community have proposed sensing vibrations using mmWave [2], WiFi [6], or ultra-wideband (UWB) radar [5] to infer the intrinsic states of objects. However, these RF sensing techniques have difficulties building connections between their inference results and real-world objects because they have no information about these objects' extrinsic properties. Their acquired RF data are time series that are difficult to decipher without knowing what the object is. Also, these systems either target a single object or require a blind search to identify the vibrations, which is error-prone and can not work for inactive objects. In this demo abstract, we propose a multimodal sensor fusion system that combines LiDAR, camera, and UWB radar for a rich semantic labeling of a complex scene. With the fusion of the three modalities, we can infer each object's extrinsic properties, as well as their intrinsic states based on their vibrations.

2 SYSTEM DESIGN

Capricorn's architecture is presented in Figure 1, where blue blocks indicate system components, red stands for algorithms, and yellow means exchanged information. Capricorn takes a late fusion approach, where sensors can make individual inferences while sharing and exchanging their inference results via a shared in-memory database via queries (marked as orange arrows).

First ❶, the camera processes its images using state-of-the-art object detection and recognition algorithms such as YOLOv5 [3]. A tracker based on Kalman filtering and the Hungarian algorithm is then applied to obtain object ID, object bounding boxes, and object types. Second ❷, the LiDAR aligns its frame with the camera. Then Capricorn queries the objects' bounding boxes and combines the LiDAR's depth map with these bounding boxes to generate a depth histogram for each object, with which the distances from the sensor to each object can be estimated. Third ❸, the UWB radar sends pulses continuously to probe the scene to generate a

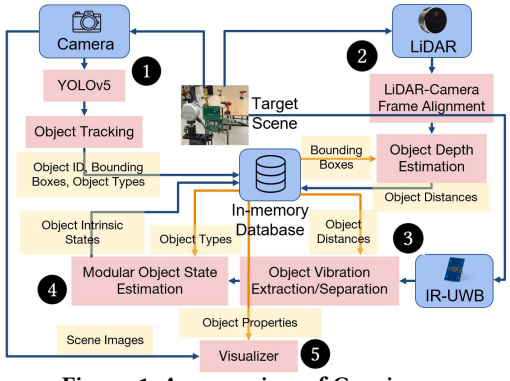

Figure 1: An overview of Capricorn.

SenSys '22, November 6–9, 2022, Boston, MA, USA

ACM ISBN 978-1-4503-9886-2/22/11.
https://doi.org/10.1145/3560905.3568060

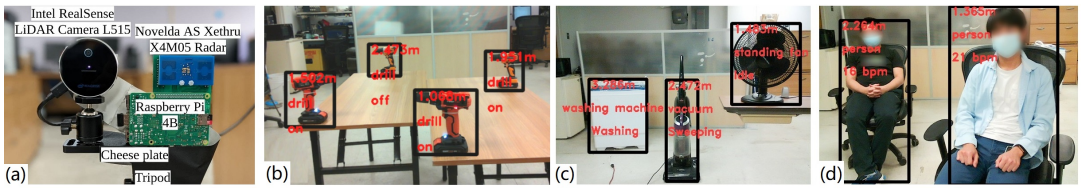

Figure 2: Implementations and results: (a) System hardware (b) Machine states estimation in a workshop (c) Appliance states estimation in a smart home (d) Multi-person respiration rate estimation

two-dimension range profile, where the y-axis is the time, and the x-axis is the distance calculated using the signal-time-of-flight (ToF). When Capricorn queries the objects' distance, Capricorn uses this information to take slices in the x-axis from the 2D range profile (select a signal distance bin for each object). This step assures that we do not have to blindly search for active objects over all the distances and that there is a one-on-one mapping between objects' images and their time series. Fourth ❹, we perform post-processing on each object's RF time series data and generate estimations about their states. This step is designed modularly: the optimal algorithms and models for processing the objects' data are different for different types of objects. Thus, we query each object's type from the database, create a thread for each object, and apply the most appropriate signal processing pipeline. For example, simple SVM models are used to classify the operating states of home appliances, and Variation Mode Decomposition-based respiration rate estimation [7] is performed for human subjects. Finally, we visualize the scene by overlaying all the inference results over the camera's images (❺).

3 IMPLEMENTATION AND RESULTS

Capricorn's multimodal sensor module is implemented with an Intel RealSense LiDAR Camera L515 and a Novelda AS Xethru X4M05 Radar sensor. We implemented the entire software architecture in C++ using ROS [4], where we utilized the Pub-Sub mechanism in ROS for real-time sensor data collection and processing. All the computations occur on an Intel NUC mini PC except that the UWB Radar driver runs on a Raspberry Pi under the same local area network.

Our results show that Capricorn can simultaneously monitor multiple home appliances' operating status and recover vital signals like respirations from multiple people in real-time (see Fig 2(b-d)). Fig 2(b) shows a setting simulating industrial assembly lines or workshops, where it is crucial to monitor the operating states of machines to promote safety and productivity. We placed four drills that are turned on and off in front of Capricorn, exhausting all possible combinations. On average, Capricorn achieves an accuracy of 99.47% classifying the on/off states of the machines. Fig 2(c) shows a smart home setting where we can use Capricorn to understand the operating states of home appliances, e.g., the rotating speed level of a table fan and whether a washing machine is performing washing, is performing drying, or is idle. In Fig 2(d), we show Capricorn can simultaneously monitor multiple subjects' respiration rates at a median error of 0.9603 bpm and associate the breath rate with the image of that person. This capability can be helpful in medical triage situations where we must simultaneously assess multiple wounded's conditions. Capricorn can process the LiDAR-camera information at a latency of 42.81 ± 6.30 ms without any hardware accelerator and process radar information at less than 200 ms for non-human subjects (on a 1.024s buffer) and 2 seconds for human subjects (on a 30-second buffer).

4 DISCUSSIONS AND CONCLUSIONS

In this abstract, we presented a vision-guided wireless vibrometry system for real-time rich scene analysis. Using a more computationally intensive sensing modality (vision) to guide a less intensive modality (RF) seems counter-intuitive at first glance. However, as a unique sensing modality that perceives the world similarly to a human, the camera is pervasive in intelligent systems for scene analysis and is unlikely to be entirely replaced. On that basis, this work demonstrates that the information acquired from computer vision can help us make more sense of the RF data and improve the versatility of RF systems in dynamic environments where the number and the type of the sensing target objects are all uncertain. Correspondingly, the RF sensors can compensate the vision sensors by making inferences about the objects' intrinsic states that are invisible to cameras, resulting in a "richer" scene analysis. In the design of Capricorn architecture, the key concept is that the inferences from one sensor can serve as the prior information for the processing of another sensor. This information sharing and exchange improve the efficiency of the multimodal sensing system.

ACKNOWLEDGMENTS

The authors would like to thank Miss Haorui Sun at UCLA for helping with the experiments. The research reported in this paper was sponsored in part by: the CONIX Research Center, one of six centers in JUMP, a Semiconductor Research Corporation (SRC) program sponsored by DARPA; by the IoBT REIGN Collaborative Research Alliance funded by the Army Research Laboratory (ARL) under Cooperative Agreement W911NF-17-2-0196; by the NIH mHealth Center for Discovery, Optimization and Translation of Temporally-Precise Interventions (mDOT) under award 1P41EB028242; and, by the National Science Foundation (NSF) under award #CNS-1705135. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the ARL, DARPA, NIH, NSF, SRC, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.

REFERENCES

[1] Qaisar Abbas et al. 2018. Video scene analysis: an overview and challenges on deep learning algorithms. *Multimedia Tools and Applications* 77, 16 (2018), 20415–20453.
[2] Chengkun Jiang et al. 2020. mmVib: micrometer-level vibration measurement with mmwave radar. In *Proceedings of MobiCom '20*. 1–13.
[3] Glenn Jocher et.al. 2022. *YOLOv5*. https://doi.org/10.5281/zenodo.7002879
[4] Stanford Artificial Intelligence Laboratory et al. 2018. *Robotic Operating System*. https://www.ros.org
[5] Ziqi Wang et al. 2020. UWHear: through-wall extraction and separation of audio vibrations using wireless signals. In *Proc. of SeySys '20*. 1–14.
[6] Teng Wei et al. 2015. Acoustic eavesdropping through wireless vibrometry. In *Proceedings of MobiCom '15*. 130–141.
[7] Tianyue Zheng et.al. 2020. V2iFi: In-vehicle vital sign monitoring via compact RF sensing. *Proc. of IMWUT* 4, 2 (2020), 1–27.