Real-Time Multimodal Tactile Sensor with Visual and Auditory Feedback

Hyosung Kim
Graduate School of Data Science
Kyungpook National University
Daegu, Republic of Korea
hyosungk98@knu.ac.kr

Junhui Lee
Graduate School of Data Science
Kyungpook National University
Daegu, Republic of Korea
jk06033@knu.ac.kr

Saekwang Nam*
Graduate School of Data Science
Kyungpook National University
Daegu, Republic of Korea
s.nam@knu.ac.kr
ORCID : 0000-0002-7713-8505

***Abstract*—Vision-Based Tactile Sensors (VBTS) offer high spatial resolution but exhibit limitations in capturing high-frequency dynamic events, such as slip and texture, which consequently constrains robotic dexterity. To overcome this limitation, we propose a multimodal tactile sensor system that augments a standard VBTS with a microphone for Fast-Adapting (FA) feedback. The entire system is self-contained on a single NVIDIA Jetson Orin NX board and is operated by a unified ROS 2 node. We introduce a real-time, camera-triggered synchronization method that fuses high-resolution visual data with corresponding high-frequency auditory data, publishing the synchronized information as a unified tactile message. Experimental validation demonstrates two primary advantages. First, the system can detect not only static contact events through its internal camera but also transient dynamic events via the integrated microphone. Second, it successfully discriminates between different surface textures by identifying their unique vibrational signatures, a capability that is also crucial for detecting slip. This work presents an accessible and efficient framework for acquiring rich, dynamic tactile information, thereby advancing the development of more robust and dexterous robotic manipulation.**

***Index Terms*—Tactile sensor, multimodal sensing system, real-time data processing**

I. Introduction

Tactile sensing is fundamental to how humans perceive and interact with their environment, and equipping robots with analogous capabilities is critical for achieving dexterous manipulation [1]. While early tactile sensors, such as capacitive or resistive types, could detect basic contact, they were hampered by manufacturing complexities and low spatial resolution, limiting the richness of the contact information they could provide [2].

Vision-Based Tactile Sensors (VBTS) have emerged to address these shortcomings by utilizing internal cameras to generate high-resolution tactile data from physical deformations. A prominent example is the TacTip [3], which features a soft, deformable skin with an internal array of pins. As the sensor makes contact with an object, an integrated camera captures the displacement of these pins to produce high-resolution tactile feedback. The open-source nature of platforms like TacTip has fostered widespread adoption, providing researchers with an accessible, low-cost tool for modification and experimentation, thereby significantly advancing robotic tactile research [4].

Human skin perceives complex tactile information through two types of receptors: slow-adapting (SA) receptors, which detect sustained pressure and shape, and fast-adapting (FA) receptors, which detect subtle vibrations and texture [5]. Vision-based sensors primarily mimic SA receptors but largely fail to capture FA information. The mean operating cycle of cameras utilized in VBTS is estimated to be between 30 and 60 times per second. While this is sufficient to perform the role of human skin's SA receptors, it is inadequate for the response speed required to perform the role of FA receptors, which must detect rapid changes exceeding hundreds of hertz. Consequently, robots demonstrate a notable vulnerability in dynamic interactions, failing to perceive transient vibrations upon contact with an object's surface or the rough texture of the surface. In this study, we propose the development of a multimodal tactile sensor that integrates heterogeneous sensors to perform the roles of both SA receptors and FA receptors to address these limitations. Specifically, it integrates the high-resolution spatial information provided by existing VBTS with the vibration information detected by a high-sensitivity microphone.

II. Related Work

VBTS research has achieved significant progress in recent years. TacTip offers straightforward interpretation of feature points by monitoring the displacement of its internal pin array, thus establishing itself as a prevalent platform within academic circles due to its extensive compatibility with 3D printing technology [3], [4]. Operating on a different principle, GelSight demonstrates the ability to optically track continuous deformations of a translucent gel surface, thereby reconstructing highly precise 3D surface information, which shows strong performance in analyzing fine textures [6]. In addition to these, various sensor forms have been proposed, such as research increasing sensor form freedom using flexible optical waveguides [7] and OmniTact, which detects contact across

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (RS-2023-00242528, RS-2024-00436182) and by the IITP (Institute for Information & Communications Technology Planning & Evaluation)-ITRC (Information Technology Research Center) grant funded by the Korea government (No. IITP-2025-RS-2024-00437756) and the other IITP grant funded by the Korea government (MSIT) (No. RS-2025-02263277).

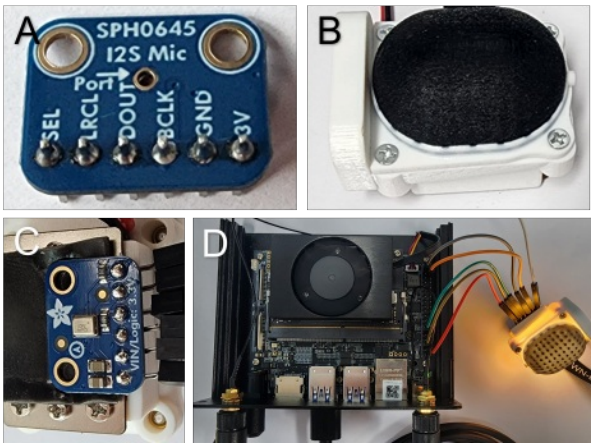

Fig. 1. Hardware components of the multimodal tactile sensor. (A) A microphone serving as the Fast-Adapting (FA) receptor. (B) The TacTip, which functions as the Slow-Adapting (SA) receptor. (C) The microphone is attached to the backside of the TacTip. (D) The NVIDIA Jetson Orin NX board for real-time data integration.

entire curved surfaces [8], thereby expanding the possibilities of vision-based tactile sensing.

Previous attempts have been made to recognize the importance of FA receptors and implement them. Examples include the attachment of microphones to robotic skin for the classification of objects by contact sounds [9] and research utilizing piezoelectric elements to detect impacts and slips at high speeds [10]. While these studies clearly demonstrated the potential of multimodal sensing, they were often systems built for specific purposes, lacking versatility or integration with widely used open-source platforms. Recently, Meta AI developed DIGIT 360, a multimodal VBTS that applies principles similar to GelSight while additionally detecting vibrations or inertia [11]. However, given its status as a commercial product, it is challenging to modify or extend its internal hardware or software for research purposes.

III. Materials and Methods

A. System Overview

Proposed multimodal tactile sensing system, shown in Fig. 1(D), is designed as a standalone system with all functions integrated on a single NVIDIA Jetson Orin NX board computer. The primary objective of this system is to fuse, in real-time, the high-resolution spatial information provided by TacTip (Fig. 1(B)) with the high-frequency vibration information detected by a microphone (Fig. 1(A, C)).

The system's software is a modular framework built upon ROS 2. It is designed to autonomously acquire and precisely time-synchronize data from heterogeneous sensors, which is then disseminated as a unified, readily-utilizable topic. This architecture ensures reproducibility, promotes scalability, and is intended to lower the barrier for adoption by other researchers.

B. Hardware Components

All hardware components are directly interfaced with the NVIDIA Jetson Orin NX board(SBC) (Fig. 1(D)). The detailed specifications for each component are as follows.

a) Host System: The NVIDIA Jetson Orin NX 16GB board was utilized as the central processing unit of the entire system. This board offers the potential to run future on-device machine learning models through its powerful integrated GPU. It provides an optimal environment for directly connecting sensors without a separate controller, featuring both USB and GPIO headers.

b) Vision Sensor: The primary tactile sensing component is a vision-based sensor derived from the open-source TacTip design, which measures contact shape and pressure distribution. Internally, a standard USB webcam captures the sensor's deformation at a 1280×720 resolution and 30 Hz refresh rate, streaming the data to the single-board computer (SBC) via its USB port.

c) Auditory Sensor: An Adafruit SPH0645 I2S MEMS Microphone was utilized to detect subtle textures and slippage vibrations. The high-sensitivity microphone establishes a direct connection with the SBC's 40-pin GPIO header through the utilization of the I2S communication protocol, thereby facilitating the acquisition of audio streams at an elevated sampling rate of 44.1 kHz. The sensor was affixed to the base of the TacTip, meticulously designed to detect vibrations transmitted to it.

C. Software and Data Synchronization

The software framework, as shown in Fig. 2, has been developed using ROS 2 Humble Hawksbill, which operates on the Jetpack 6.2, Ubuntu 22.04 based operating system. All data acquisition and integration processes within the system are efficiently handled within a single ROS 2 node that acts as the Sensor Integration Driver Node. The data processing pipeline of this driver node is designed as follows:

a) Asynchronous Microphone Data Buffering: A dedicated thread within the node is responsible for the collection of microphone data. This particular thread has the capacity to continuously read audio data in real-time at a 44.1 kHz sampling rate from an I2S microphone that is connected via GPIO. The read audio data is sequentially stored in an internal buffer.

b) Camera-Triggered Data Synchronization and Publishing: The node's primary loop functions in accordance with camera frames and employs a trigger-based synchronization method, as outlined subsequently. Initially, the primary loop acquires a new image frame from the USB camera. This frame acquisition action serves as the reference signal for the entire data integration process, thereby functioning as the trigger. Upon successful frame reading, the frame's timestamp is recorded. Subsequently, the system retrieves all audio data accumulated in the microphone data queue up to that point as a single data chunk.

```
#tactile_msgs/msg/MultimodalData.msg
std_msgs/msg/Header header
```

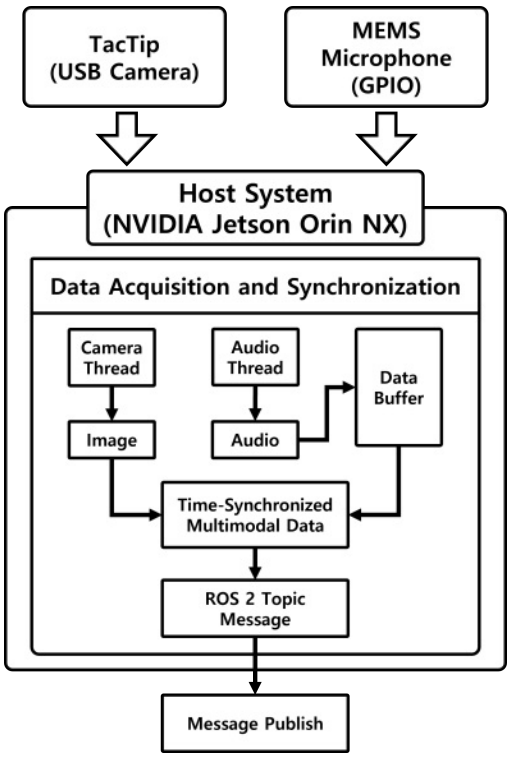

Fig. 2. Data acquisition and synchronization pipeline

```
sensor_msgs/msg/Image image
audio_common_msgs/msg/AudioData audio
```

The completed MultimodalData.msg message is ultimately published to a single topic named /tactile/multimodal, enabling immediate utilization by other applications. This trigger-based synchronization method inherently addresses variations in data processing speeds between the two sensors. This approach guarantees data integrity by aligning data to a definitive reference point: the camera frame.

IV. Experiments and Results

In order to validate the proposed multimodal tactile sensor in this study, two experiments were conducted. The overarching objective of the present series of experiments is to demonstrate that dynamic tactile events, which are difficult or impossible to detect using visual information alone (i.e., the SA receptor), can be successfully captured through acoustic/vibration information (i.e., the FA receptor). The initial experiment centered on the capacity to discern instantaneous contact and separation occurrences. The subsequent experiment examined variations in vibration patterns during slip events, contingent on the nature of the underlying surface material. All data were collected in real-time from the /tactile/multimodal topic via the previously described system.

A. Experiment 1: Transient Contact Detection

a) Objective: The objective of this experiment is to compare the temporal precision of the vision-only and multimodal methods in detecting a transient tap event.

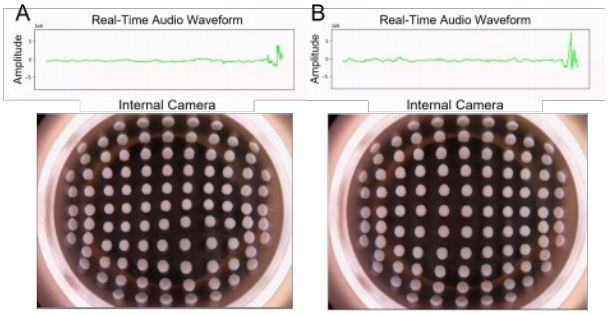

Fig. 3. Results from the transient contact (tap) detection experiment. Subfigures (A) illustrate the sensor's state at the moment of initial contact (pressing), while subfigures (B) illustrate the state at the moment of separation (release).

b) Procedure: With the sensor firmly secured, the experimenter repeatedly tapped its center by rapidly applying and releasing pressure. Throughout this action, image and audio data were recorded synchronously.

c) Results and Discussion: As demonstrated in Fig. 3, the multimodal system exhibits superior temporal resolution in detecting a rapid tap-and-release event. The visual data from the internal camera (bottom row) shows a nearly identical pattern of pin deformation at both the moment of contact (Fig. 3(A)) and release (Fig. 3(B)), making it challenging to distinguish between these two transient events from the images alone. In stark contrast, the auditory data (top row) provides unambiguous temporal cues. A sharp, high-amplitude spike in Fig. 3(A) corresponds to the initial impact, while a second distinct spike in Fig. 3(B) precisely marks the moment of separation. These results confirm that the microphone can capture and differentiate transient events with high fidelity—a capability crucial for robust manipulation that remains a significant challenge for vision-only tactile sensors.

This experiment highlights a key advantage of the proposed system. Whereas conventional vision-based sensors are limited to interpreting dynamic events as a series of state changes between frames, our multimodal sensor can instantaneously detect such events as discrete occurrences within a single data frame using the auditory signal. This capability has the potential to significantly reduce system latency and improve reaction times in real-time robotic control.

B. Experiment 2: Surface Texture Discrimination during Slip

a) Objective: The objective is to verify whether the system's audio data can distinguish between different surface textures during a slip event, particularly when those surfaces are indistinguishable by visual information alone.

b) Procedure: Data was collected from three surfaces prepared with distinct textures: a smooth surface, 80-grit sandpaper, and 2000-grit sandpaper. For each surface, data acquisition was performed by applying pressure with the sensor tip.

c) Results and Discussion: The value of the proposed multimodal approach for texture discrimination is underscored by the experimental results. Visually, the data from the TacTip sensor was nearly indistinguishable across the three surfaces during a slip, as shown in Fig. 4(B). While these images successfully confirmed physical contact, they lacked sufficient information to discern the surface texture.

In stark contrast, the auditory data, shown in Fig. 4(A), captured the distinct physical characteristics of each surface with high fidelity. The audio stream exhibited unique vibrational signatures corresponding to each texture: the smooth surface produced a low-amplitude baseline signal, while the 80-grit (coarse) and 2000-grit (fine) sandpapers both generated high-amplitude, high-frequency vibrations, each possessing a signature pattern reflective of its specific texture.

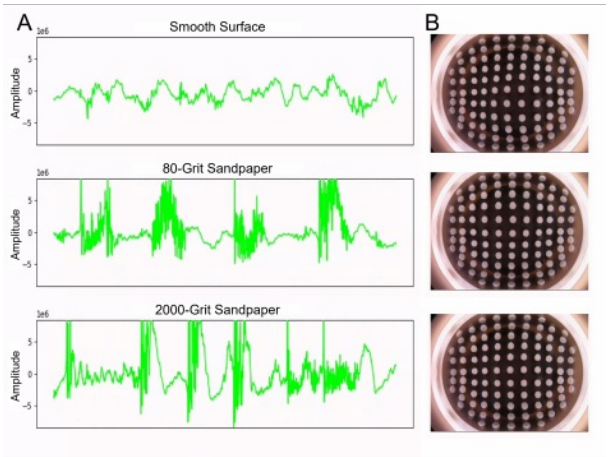

Fig. 4. Comparison of multimodal sensor data when sliding over a smooth surface, 80-grit sandpaper, and 2000-grit sandpaper. (A) The audio waveforms measured for each surface show unique vibrational patterns. (B) The corresponding TacTip images, captured during the slip, show the sensor's deformation.

In conclusion, the auditory data provided unique "vibrational signatures" that successfully distinguished between surface textures where vision-only methods failed. While the sensor is sensitive to ambient noise, the distinct textural signals are robust enough to provide essential information for material identification and the refinement of robotic gripping strategies.

V. Conclusion

In this paper, we presented and validated a multimodal tactile sensor that fuses a vision-based SA channel with a microphone-based FA channel on a single, ROS 2-driven embedded platform. Our experiments confirm that this system effectively captures dynamic tactile events, demonstrating two key advantages over vision-only methods. First, it achieves superior temporal resolution by detecting transient contact events from the auditory signal within a single data frame, overcoming the multi-frame analysis required by vision-alone approaches. Second, it successfully discriminates between visually-indistinguishable surfaces by identifying their unique vibrational signatures in the audio data. These capabilities provide novel opportunities for robots to identify materials and adapt their interaction strategies in real-time.

However, our experimental process also revealed a key limitation: the high sensitivity of the microphone, while beneficial for capturing contact data, also makes it susceptible to ambient noise from sensor movement and the surrounding environment. This noise can potentially corrupt the pure contact signal. To address this, our immediate future work will focus on implementing a differential sensing approach. This involves integrating a secondary, reference microphone to specifically measure and subtract ambient vibrations, thereby isolating the true contact signal. This enhancement is expected to yield more robust multimodal data, contributing to the overarching goal of achieving human-like robotic dexterity.

References

[1] M. Raibert, K. Blankespoor, G. Nelson, and R. Playter, "BigDog, the rough-terrain quadruped robot," in *Proc. 17th IFAC World Congress*, 2008, pp. 10822-10825.

[2] M. S. Arian, C. A. Blaine, G. E. Loeb, and J. A. Fishel, "Using the BioTac as a tumor localization tool," in *2014 IEEE Haptics Symposium (HAPTICS)*, 2014, pp. 443-448.

[3] B. Ward-Cherrier *et al.*, "The TacTip family: Soft optical tactile sensors with 3D-printed biomimetic morphologies," *Soft Robotics*, vol. 5, no. 2, pp. 216-227, 2018.

[4] N. F. Lepora, "Soft biomimetic optical tactile sensing with the TacTip: A review," *IEEE Sensors Journal*, vol. 21, no. 19, pp. 21131-21143, 2021.

[5] R. S. Johansson and J. R. Flanagan, "Coding and use of tactile signals from the fingertips in object manipulation tasks," *Nature Reviews Neuroscience*, vol. 10, no. 5, pp. 345-359, May 2009.

[6] W. Yuan, S. Dong, and E. H. Adelson, "GelSight: High-Resolution Robot Tactile Sensors for Estimating Geometry and Force," *Sensors*, vol. 17, no. 12, p. 2762, 2017.

[7] J. Zhou *et al.*, "Prosthetic finger for fingertip tactile sensing via flexible chromatic optical waveguides," *Materials Horizons*, vol. 10, no. 11, pp. 4940-4951, 2023.

[8] A. Padmanabha, F. Ebert, S. Tian, R. Calandra, C. Finn, and S. Levine, "OmniTact: A multi-directional high-resolution touch sensor," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*, 2020, pp. 618-624.

[9] J. Liu and B. Chen, "Sonicsense: Object perception from in-hand acoustic vibration," *arXiv preprint arXiv:2406.17932*, 2024.

[10] W. Liu, P. Yu, C. Gu, X. Cheng, and X. Fu, "Fingertip piezoelectric tactile sensor array for roughness encoding under varying scanning velocity," *IEEE Sensors Journal*, vol. 17, no. 21, pp. 6867-6879, 2017.

[11] M. Lambeta *et al.*, "Digitizing touch with an artificial multimodal fingertip," *arXiv preprint arXiv:2411.02479*, 2024.