AFRL-AFOSR-JP-TR-2023-0051

Edge Intelligence based Hand Gestures Recognition using Wearable Multimodal

Tran, Thanh Hai
Hanoi University of Science and Technology
No 1 Dai Co Viet, Hai Ba Trung district
Hanoi, , 00000
VN

12/23/2022
Final Technical Report

Air Force Research Laboratory
Air Force Office of Scientific Research
Asian Office of Aerospace Research and Development
Unit 45002, APO AP 96338-5002

REPORT DOCUMENTATION PAGE

PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ORGANIZATION.

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED	
20221223	Final	START DATE 20200930	END DATE 20220929

4. TITLE AND SUBTITLE
Edge Intelligence based Hand Gestures Recognition using Wearable Multimodal

5a. CONTRACT NUMBER	5b. GRANT NUMBER	5c. PROGRAM ELEMENT NUMBER
	FA2386-20-1-4053	
5d. PROJECT NUMBER	**5e. TASK NUMBER**	**5f. WORK UNIT NUMBER**

6. AUTHOR(S)
Thanh Hai Tran

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Hanoi University of Science and Technology
No 1 Dai Co Viet, Hai Ba Trung district
Hanoi 00000
VN

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)	11. SPONSOR/MONITOR'S REPORT NUMBER(S)
AOARD UNIT 45002 APO AP 96338-5002	AFRL/AFOSR IOA	AFRL-AFOSR-JP-TR-2023-0051

12. DISTRIBUTION/AVAILABILITY STATEMENT
A Distribution Unlimited: PB Public Release

13. SUPPLEMENTARY NOTES

14. ABSTRACT
Hand gestures recognition (HGR) has achieved great success and opened a new trend in human-machine interaction in recent years. Deployment of some existing HGR systems in practical applications still meets some challenges such as the limited measurable range of sensors; the lack of important information due to the use of a single modality; the high communication cost, latencies, and privacy burdens due to the training of complex deep models. This project aims to overcome these main issues by developing edge intelligence techniques f or hand gesture recognition using wearable multimodal sensors (e.g. accelerometer and camera) with less annotation effort. In this project, we have designed a wearable multimodal prototype that enables the capture of multimodal information such as RGB and motion data. We then designed a set of twelve dynamic hand gestures commonly utilized in the context of human-machine interaction. We collected datasets of such gestures using the designed prototype with fifty subjects in various environmental conditions. To the best of our knowledge, this dataset can be considered the first benchmark dataset for the research community of gesture recognition from wrist-worn multimodal sensors. We deployed various state-of-the-art CNN models for the comparative study of hand gesture recognition using RGB and motion data. The experimental results showed the challenges of the benchmark as well as the best performance of existing models and the room for future improvement. Besides, in the framework of the project, we improved algorithms for hand pose estimation with temporal information and continuous hand gesture recognition. We also conducted fundamental research on shape analysis and Bayesian inference in a hybrid CNN-LSTM model for time-series prediction. We introduced a framework for easy study on federated learning. The prototype and research results have been published in 12 international conferences and submitted to one IEEE Sensor journal.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES
a. REPORT U	b. ABSTRACT U	c. THIS PAGE U	SAR	36

19a. NAME OF RESPONSIBLE PERSON	19b. PHONE NUMBER *(Include area code)*
AKIRA NAMATAME	3152277010

FINAL REPORT FOR AOARD GRANT FA2386-20-1-4053

Edge intelligence based hand gesture recognition using wearable multimodal sensors for human-machine interaction

Principal Investigator: Assoc. Prof. Thanh-Hai Tran
Email: thanh-hai.tran@mica.edu.vn
Hanoi University of Science and Technology, Vietnam
Address: 1 Dai Co Viet Street, Hai Ba Trung, Ha Noi, Vietnam
P: +84 976 560 526
Co-PI: Dr. Trung-Kien Tran
Key members: Thi-Lan Le, Hai Vu, Thanh-Phuong Nguyen, Huu Thanh Nguyen, Cuong Pham
Period of Performance: 09/30/2020-09/29/2022

December 23, 2022

Abstract

Hand gestures recognition (HGR) has achieved great success and opened a new trend in human-machine interaction in recent years. However, deployment of some existing HGR systems in practical applications still meets some challenges such as the limited measurable range of sensors; the lack of important information due to the use of a single modality; the high communication cost, latencies, and privacy burdens due to the training of complex deep models. This project aims to overcome these main issues by developing edge intelligence techniques for hand gesture recognition using wearable multimodal sensors (e.g. accelerometer and camera) with less annotation effort. In this project, we have designed a wearable multimodal prototype that enables the capture of multimodal information such as RGB and motion data. We then designed a set of twelve dynamic hand gestures commonly utilized in the context of human-machine interaction. We collected datasets of such gestures using the designed prototype with fifty subjects in various environmental conditions. To the best of our knowledge, this dataset can be considered the first benchmark dataset for the research community of gesture recognition from wrist-worn multimodal sensors. We deployed

various state-of-the-art CNN models for the comparative study of hand gesture recognition using RGB and motion data. The experimental results showed the challenges of the benchmark as well as the best performance of existing models and the room for future improvement. Besides, in the framework of the project, we improved algorithms for hand pose estimation with temporal information and continuous hand gesture recognition. We also conducted fundamental research on shape analysis and Bayesian inference in a hybrid CNN-LSTM model for time-series prediction. We introduced a framework for easy study on federated learning. The prototype and research results have been published in 12 international conferences and submitted to one IEEE Sensor journal.

Key words: Hand Gesture Recognition, Wearable Sensor Design, Deep Learning, Edge Intelligence

CONTENTS

LIST OF FIGURES

LIST OF TABLES

1 ACCOMPLISSMENT

1.1 GOAL AND TECHNOLOGICAL OBJECTIVES

Hand gestures have become efficient means of communication between humans and machines. To this end, hand gestures must be acquired by sensors and automatically recognized and matched to commands to control systems. To collect data, RGB-(D) cameras and accelerometers are the most commonly used as ambient and wearable sensors, respectively. Wearable sensors-based hand gesture recognition systems usually use sole modality (e.g., accelerometer), which causes a lack of important information for recognition. As a result, it is preferable to use multimodal sensors.

Regarding hand gesture representation, recently, deep learning has shown its outperformance in many vision-based tasks, including hand gesture recognition. However, accurate models for hand gesture recognition (i.e., deep learning-based methods) are highly complex and usually need to be trained on powerful servers (e.g. in the cloud) that cause communication costs, latencies, and privacy burdens. In addition, a big annotated dataset for training the CNN model is not always available.

The goal of this project is to overcome the above issues by enhancing the flexibility, privacy, and robustness of dynamic hand gesture recognition from multimodal wearable sensors by using edge intelligence techniques. **Technological objectives** include:

- Design a wristband prototype that integrates heterogeneous sensors (camera, accelerometer, and gyroscope), being able to capture multimodal data for robust hand gesture recognition in a scalable environment.
- Design and implement a multi-stream deep neural network that takes multimodal data as the input and explores complementary of each modality for better gesture recognition accuracy.
- Study edge intelligence techniques to deal with privacy concerns and low-resource edge devices.
- Deployment of meta-learning technique to solve with lack of annotated data.

1.2 APPROACHES

1.2.1 LITTERATURE SURVEY

To achieve the objectives, we first surveyed existing works on human-machine interaction using hand gestures from wearable sensors, the datasets, the prototype, and the methodologies. We then analyze and raise the main issues of those works and propose solutions to overcome them.

- **Survey on existing wrist-worn devices and datasets**: The works on wrist-worn devices for action recognition are quite limited.

  In [7], Maekawa et al. designed a wrist-worn device integrating a camera, a micro-

phone, an accelerometer, and a compass. The device was built to collect data for object-based action recognition. The dataset contains 15 daily life activities (ADL), but it was not available for evaluation.

Ohnishi et al. proposed a network to capture and recognize activities of daily living using a wrist-mounted, and a head-mounted camera [8]. The wrist-mounted camera looks at the hand hollow to facilitate the observation of the hand when interacting with objects. Twenty-three ADL classes have been collected and annotated using both cameras.

Yeo et al. [14] introduced a prototype of a wearable camera with a view of the opisthenar (back of the hand) area. The device used a single infrared camera (Leap Motion device) with an active infrared light source to easily remove the background. They collected ten static hand postures and five individual finger-tapping actions for dynamic gestures.

In [4], [2], [3], the authors have deployed an RGB camera worn on the backside of a user's right to capture hand gestures for human-machine interaction application. In [2], only static hand postures have been collected. Chen et al. introduced another work [3], which collected dynamic hand gestures for human-robot interaction applications.

In [13], Wu et al. designed a wrist-worn camera to capture the fingers of the hand for a hand pose estimation task. Ten static gestures of ASL digits and six dynamic gestures of finger tapping have been collected with this device.

In summary, the number of wrist-worn prototypes and datasets of hand gestures is still very limited compared to the ones captured by ambient cameras. The type, the characteristics, and the mounted position of each integrated sensor vary from prototype to prototype. The activities are dependent on applications (HCI or ADL).

- **Survey on human action recognition**: We only summarized methods that relate closely to our work, where hand gestures are captured by wearable sensors.

  In [8], the authors extracted features from each video frame using a deep model VGG-16. Then a weighted Vector of Locally Aggregated Descriptors (VLAD) was applied for Video Pooling on Convolutional Neural Network (CNN) descriptors. The authors also combined deep features with hand-crafted features (i.e., improve Dense Trajectory - iDT) to improve the classification with a Support Vector Machine (SVM).

  Wu et al. proposed a CNN model, namely DorsalNet, that takes RGB and motion history as inputs to generate 3D hand pose [13]. DorsalNet consists of 3 parts: the preprocessing stage with the encoder-decoder network for hand masking and the motion image computation; the two-stream Long-Short Term Memory (LSTM) CNN with the Kalman filter as a feature extractor; and the hand simulator, which reconstructs the finger angles. They adjust an MLP (Multilayer Perceptron) for gesture recognition just after the DorsalNet.

  Le et al. in [5] proposed a method that estimates the hand pose first, then recognizes hand action from estimated hand joints. The hand pose is represented as a graph that can be inputted into a graphical convolutional network (GCN).

Tran et al. [9] deployed two convolutional neural networks in a unified framework for both gesture detection and segmentation from the RGB video stream.

1.2.2 PROTOTYPE AND DATASET

The wrist-worn device is uncommercialized and needs to be produced to adapt to each application. We analyzed technical requirements (e.g., camera framerate, resolution, and angle view) and developed a novel prototype that consists of three main components: 1) the first component composed of a wide-angle camera, an accelerometer, and a gyroscope; 2) the second component is an embedded device (i.e., jetson nano) that is in charge of data acquisition, processing, and storage; 3) the final component is a power supply.

Then we analyzed the naturalness, memorability, and distinctness of human gestures in interaction with machines to design and collect a hand gesture set usable in the context of human-machine interaction. The dataset was collected in various environments with a large enough number of subjects so that it can be a benchmark for evaluating hand gesture recognition methods from wrist-worn sensors. The collected data are synchronized and annotated spatially and temporally to enable multimodal recognition approaches.

1.2.3 RECOGNITION METHODS

We analyzed some main criteria for an application of human-machine interaction using hand gestures from wearable sensors. This aims at choosing machine learning models satisfying the best trade-off between computational time and accuracy. We investigated some state-of-the-art deep learning models such as MobileNet, R3D, R(2+1)D, MoviNet, EfficientNet, and C3D to evaluate both memory requirements and the actual accuracy of those models on the collected dataset.

We analyzed the main issue of those models facing the particular characteristics of the dataset and suggested improvements:

- Consider an additional stream of optical flow calculated from RGB to capture better the movement of gestures.
- Comparative study on the combination of both RGB and Optical Flow streams
- Evaluate the recent model, such as the Transformer, to boost the performance of motion sensor-based gesture recognition.
- Transfer learning and Few-shot recognition are also studied to deal with the lack of annotated data.

The above techniques are studied and experimented in a centralized framework where all data and training processes are conducted on a server. To deal with the problem of user privacy, latency in data transmission, and communication cost, as well as to leverage the computing resource at edge devices, some edge intelligence techniques are studied and experimented with:

- Model compression: when the model is trained and deployed on edge devices with

low computation and memory resources, the model must be as lightweight as possible. Different strategies for model compression are studied, such as pruning, model quantization (aware training or post-training), drop-out, gradient cutting-off, and layer/block reduction).

- Federated learning: to overcome the main issues encountered by centralized learning, federated learning allows to train the models at local edge devices and sends only the weight to the server to build the global model.

To evaluate the proposed methods, an application of human-machine interaction is implemented in real environments with subjects who perform gestures in a continuous manner. As a result, techniques to deal with continuous temporal data and temporal segmentation are studied. There are two main approaches. The first is to apply window sliding, apply a gesture classifier then vote for each window. This approach is quite a computational cost because the classifier is activated all the time. The second approach is to build a small classifier just to detect the presence of a meaningful action before applying the classifier.

1.3 Details of accomplishments

1.3.1 Major activities

To obtain the specific objectives, we have realized the following activities:

- Survey and design wearable multimodal data of dynamic hand gestures.
- Study and propose deep learning models for hand gesture recognition using an accelerometer and a camera.
- Study and develop Edge intelligence techniques for HAR
- Design edge and Cloud platform for HAR from multiple streams
- Application of controlling home appliance system using dynamic hand gestures in smart home

To conduct the above activities, we have organized

- Kick-off meeting: To present the project objectives, budget allocation, main tasks, and schedule to all project members.
- Weekly meeting: Project members present their research. Then members round-table discuss the progress of each task, the difficulties, and the solution.
- Task sharing: The main tasks of the project are divided into sub-tasks which are conducted by one or several members. Data, code, and reports are shared on some platforms such as onedrive, google drive, and our own drive for collaborators.
- Yearly evaluation: The PI and co-PI projects reported the main achievement of the project at the current reporting time and

1.3.2 SIGNIFICANT RESULTS

In this project we have achieved the following key outcomes:

- A new low-cost wearable prototype for capturing multimodal data.
- A new dataset of dynamic hand gestures for human-machine interaction, that could be shared for research purposes.
- Comparative results to evaluate the performance of state-of-the-art models on our dataset. We find that MoviNet is the best model that makes a good trade-off between accuracy and computational time.
- A two-stream MoviNet that improves the recognition accuracy from RGB and Optical Flow.
- A new framework for initiating the study on federated learning

As a result, we have:

- Published 12 international conference papers.
- Submitted 01 journal paper (IEEE Sensor)
- Participated to train 2 Ph.D. students, 02 Master's students, and some undergraduate students.
- Organized 1 workshop on "Multimodal Analysis using advanced deep learning"
- Participated in different events organized by HUST (BKFamily Day, HUST Admissions Counseling.)

1.4 DISSIMINATION OF RESULTS

To introduce the project and its results to a wide audience, we conducted different activities during the project development.

- Firstly, we organized weekly meetings and internal seminars so that all members of the project, as well as the students who joined the project, would be informed of all research activities of the projects and propose different solutions for encountered problems.
- Secondly, to illustrate the output of the project, we have developed a simple Rock Scissors Paper game based on the action recognition method and organized a demonstration in BKFamily Day - an annual event that HUST organizes for the young generation. This demo has attracted a lot of young members. In this way, the next generation will be aware of new technology. Figure 1.1 shows a photo taken from this event.
- Thirdly, we have published the results of the projects in different venues such as MAPR 2022, APSIPA 2022, NICS 2022. Figure 1.2 has been captured in MAPR 2022 when our Ph.D. student Hong-Quan Nguyen presented results of action recognition.
- Fourthly, we organize a workshop on "Multimodal analysis using advanced deep learning" with participants from some universities in Hanoi. We invite researchers and mas-

ter students. On the one hand, we introduced and present the outcomes of the projects. On the other hand, we invited researchers to present their ongoing research related to the topic studied in the project to discuss the improvement and create new collaboration.

- Finally, through the teaching activity of the project members, the methods, and results of the projects have been introduced to a wide range of students, from undergraduate to Ph.D. students. Thanks to this activity, a new Ph.D. candidate working on federated learning will start next year.

Figure 1.1: Demo of Rock paper scissors game in BKFamily day.

Figure 1.2: PhD student Hong-Quan Nguyen presents the project research paper at MAPR 2022 conference.

2 IMPACTS

2.1 IMPACT ON DEVELOPMENT OF THE PRINCIPAL DISCIPLINE(S) OF THE PROJECT

The research topics conducted in this project have strongly impacted the development of our research group. The research group comprises members from mainly three universities and one institute in Hanoi, Vietnam. Two Ph.D. students (Trung-Hieu Le and Hong-Quan Nguyen) come from Viet-Hung University and Dai-Nam University. Their Ph.D. thesis relates closely to the research topics of the project. Particularly, Trung-Hieu Le is supervised by Assoc. Prof. Cuong Pham and Assoc. Prof. Thanh-Hai Tran. Both are key members of the project.

The project members have a common interest in image and video understanding. Especially regarding action recognition from egocentric vision, we have conducted a project funded by Nafosted (National Foundation for Science and Technology Development of Vietnam) that aimed at developing tools to evaluate human gestures for rehabilitation evaluation using a head-mounted camera. This project focuses on human-machine interaction with wrist-worn multimodal sensors. It strengthens research competence in action recognition of our research members with other heterogeneous sensors. Specially,

- **Improvement of hardware design competence**: In this project, we have designed and made a new prototype that consists of one camera and IMU sensors integrated into a unique wrist-worn device. The prototype enables capturing multimodal information about the hand gestures and the surrounding context. These multimodal data streams are then transmitted and stored in an embedded device. The embedded device is in charge of computing and recognizing the gestures for further applications such as health care, human-machine interaction, lifelogging, etc. In the literature, there exist some similar prototypes, but the characterization of the camera, IMU sensors, and mounting positions are different, leading to different observations of the hand gestures and revealing new challenges to the research community. To the best of our knowledge,

we are the first team in Vietnam to create a new prototype and work on an egocentric vision. This research gives us competence in innovative device design suitable to specific applications or responses to some requirements.

- **Dissemination of a new dataset to the research community**: Thanks to the designed prototype, we have collected and made a new dataset available to the research community. This dataset is original and challenging regarding the number of subjects and various environments. It is annotated and evaluated using many states of the art machine learning models on both RGB streams and Optical Flow streams, and motion sensor streams in a synchronous manner. The dataset has been uploaded to the website and is available upon request. The prototype can be utilized to build other datasets for different goals, such as rehabilitation evaluation; supporting Alzheimer's to improve their memory. We also provided the first evaluation results of gesture recognition which can be the baselines for improving further recognition algorithms.
- **Strengthen the research on machine learning / deep learning**: After investigating a range of baseline methods for action recognition, we have discovered the MoviNet model that can make the best trade-off between the accuracy and the memory requirement. We also tried to find the optimal configuration of the models for the research topic. Besides, we compared MoviNet with other state-of-the-art models and recommended model selection for application deployment. Thanks to the project, we had the opportunity to work on deep models for human action recognition, mostly the ones captured by a wrist-worn sensor. We also look for combining multi-streams to increase algorithm performance.
- **Initiation of new research topics**: Besides doing research on CNN models for action recognition that takes temporal information into account, we initiated a new research topic on the Bayesian network for time series prediction. Bayesian models are able to face better the overfitting on small data and can measure uncertainty, which has a negative effect on their generalization abilities. Our group also researched techniques such as network quantization and gradient cutting off to reduce the size of deep models so that they can be deployed on low-resource devices. Especially we have started new research on federated learning. Federated learning (FL) - a collaborative learning technique that leverages the computational power of a large number of edge devices while preserving data privacy. This property is especially crucial for computer vision tasks as they usually require a huge amount of computation and protect the end users' data privacy. We will choose it as one of our research focus in the near future.

2.2 Impact on human resources

The project has an impact on human resources. Specifically, it helps to

- **Innovate knowledge in teaching Computer Vision, Image Processing and Artificial Intelligence**: Thanks to their participation in the main activities of the project, the members of the project have improved their competence and knowledge in sensor design, machine learning, deep learning, and edge intelligence. This competence has

been transferred to students in the course "AI and applications" given by Assoc. Prof. Thanh-Hai Tran and Assoc. Prof. Hai Vu to students of the fourth year of SEEE/HUST; the course in "Image Processing" and "Computer Vision" given by Assoc. Prof. Thi-Lan Le to undergraduate and master students of SEEE/HUST.

- **Motivate young students and researchers in their future career**: Through conducting the activities of the project, project members, mostly young researchers such as Ph.D. students (Hong Quan Nguyen, Trung Hieu Le), master students (Quynh Khanh Dinh Thi, Thi Oanh Ha), graduate (Hoang Nhat Tran) and undergraduate students (Viet-Duc Le, Danh Huy Vu, Van-Thang Tran), have improved a lot the competence, skills, and attitudes to start their future professional carriers. Actually, Hong Quan Nguyen and Trung Hieu Le are lecturers at Viet-Hung University and Dai Nam University. At the time of writing this report, Hong-Quan Nguyen is preparing his Dissertation for defense; Trung Hieu Le goes in the second year of his Ph.D. courses. Two master's students (Quynh Khanh Dinh Thi and Thi Oanh Ha) are also preparing their master's thesis and planning to defend in March 2023. Hoang-Nhat Tran, after a period of working on the project, has gained knowledge in machine learning and computer vision and has been recruited as a research assistant at VinUni in Vietnam. All of the students have gained skills and experiences and are highly appreciated by recruiters in research and development institutions/ organizations in AI/ML. Actually, the project attracts eight new undergraduate students from SEEE and School of Informatics and Applied Maths of HUST to participate in relevant research topics. The group has weekly meetings. We also have one Ph.D. Student candidate (Kieu Tuan Dung, a lecturer at Thuyloi Univesity in Vietnam) working on federated learning, a relevant and future research topic of the project. His thesis is planned to be supervised by Assoc. Prof. Thi-Lan Le and Assoc. Prof. Thanh-Hai Tran.
- **Give students financial support as contracts of the project:** In the project, financial support has been given to some Ph.D., master's, undergraduate, and graduate students to promote them in their research according to the contracts of the project. The students participated in the realization of tasks defined by the PI, co-PI, and key members. This financial support ensures an economic condition for Ph.D. students to concentrate on their research. Regarding graduate and undergraduate students, it helps them to reduce the economic burden of life and learning.
- **Bring opportunities to researchers and students to expose their achievement at science and technology events:** Thanks to the financial support of the project, some students and young researchers gained opportunities to participate in international virtual or on-site conferences. Particularly, the project supported Ph.D. Students (Hong-Quan Nguyen, Thi Lich Nghiem, and Trung Hieu Le) to go to the MAPR'2022 conference organized in Phu Quoc Island, Vietnam; Ph.D. student Trung Hieu Le to the International Conference on Intelligence of Things, Hanoi Vietnam; master's student Quynh Khanh Dinh Thi to the International Conference on Information and Communication Technology Convergence (ICTC)'2021 Korea, undergraduate Viet-Thanh Le to 13th International Conference on Knowledge and Systems Engineering (KSE'2021), Thailand. The projects supported Assoc. Prof. Thi-Lan Le for her conference trip in Thailand,

and Dr. Thanh-Phuong Nguyen for the registration fee at the ICPR'2022 conference in Canada. Without the financial support from the project, researchers and students may miss opportunities to participate in conferences, disseminate the research results as well as open new collaborations with other partners in the relevant research fields.

2.3 Impact on physical, institutional, and information resources

This project had the budget to buy equipment to build 4 prototypes, including the wrist-worn device and the embedded jetson for capturing and storing the data from sensors. The equipment has been accessed for use by students in the projects. The students of the School of Electrical and Electronic Engineering have the opportunity to deploy learning models on real-edge devices and improve their skills working with embedded devices.

3 CHANGES

3.1 PROBLEMS, DELAYS

- **Delay in prototype production and data collection due to Covid-19**: In the first year of the project, the Covid-19 pandemic caused long subsequences such as the closing of the lab, the late equipment purchase, then the delay in prototype production. As a consequence, we were quite late in prototyping and data collection. In addition, once we built up the prototype, the data collection process was slow because we could not recruit volunteers at that period. However, we speeded up the process in the second year of the project and completed the data collection for further steps.
- **Long review process of the submitted journal paper**: we submitted a paper to the IEEE Sensors Journal. We submitted the manuscript on 8 March 2022 with the reference number: Sensors-47833-2022 and received the first decision on 27 April 2022. Then we resubmitted the revised manuscript on 26 June 2022 with the new reference number: Sensors-50700-2022. However, the status has been showing “under review” since that date, taking nearly six months. We have contacted the Editor in Chief of the journal to know more about the paper’s status, and he answered to push up the review process. At the time of writing this report, the status of the paper does not change.

3.2 EXPENDITURE IMPACTS

- **Reallocation of publication fee budget to organize a workshop**: As aforementioned, until now, we have not yet received the notification of the submitted journal paper to the IEEE Sensor journal. As a result, we can not use the budget for publication fees because the financial year ends soon. We allocated one part of this budget to organize a workshop on "Multimodal analysis using advanced deep learning". The aim of this workshop is to disseminate the main achievement of the project to the public (students, researchers, lecturers of SEEE, MITI, and neighboring universities) to exchange knowledge, competence, and skills in research and development in computer vision, egocentric vision, and deep learning. It is an occasion to attract researchers and students to join our research labs and collaborate with us on common research interests.
- **Reallocation of remaining equipment budget and conference trip budget to conduct more experiments**:
    1. There is some equipment that is not available on the market, therefore we must produce by ourselves. We only bought the main component of the prototype and made use of existing cables and accessories to build up the final prototype. Consequently, we did not expense all the planned budget for equipment purchases.
    2. Due to the covid-19 pandemic, some conferences were virtually organized. As a result, the budget for conference trips was reduced.
    3. The trip from France to Canada for the ICPR conference and the one from France to Vietnam for the MAPR conference were partially supported by the partner lab-

oratory of Toulon University. As a result, we economized the trip fee.

We utilized the remaining budget from publication fees, conference trips, and equipment purchases to:

1. Collect more data with more subjects to benchmark the dataset. Therefore, it needs more human resources to develop the module for data pre-processing and data annotation. We have allocated the budget from the conference trip and publication fees from the first year to pay for this activity.
2. Validate the final application for controlling light using hand gestures from wrist-worn sensors in different environments and lighting conditions. We deploy the system in three additional places (with lighting, background, and subjects' age variation) to evaluate the robustness of our system.

3.3 CHANGES TO THE PRIMARY PLACE OF PERFORMANCE FROM THAT ORIGINALLY PROPOSED

All of the main activities were conducted at Hanoi University of Science and Technology. However, to vary the subjects participating in data collection and experiment with the final system of the project for human-machine interaction, we involve subjects from three working places of our project members: Viet-Hung Univesity (Ph.D Hong Quan Nguyen), Dai Nam University (Trung Hieu Le), MITI (Trung Kien Tran).

4 TECHNICAL UPDATES

4.1 PROTOTYPE

In this project, we aim to control devices by pointing at them and giving corresponding control actions. To do that, we build a novel prototype that integrates multiple sensors into a wrist-worn device. Specifically, this wearable device consists of three main components: 1) a wide-angle camera, an accelerometer, and a gyroscope; 2) an embedded computer (i.e. jetson nano) and 3) a power supply. The device is a single watch-like hand band so that users can feel comfortable when wearing it at his wrist. As a prototype, we use a low-cost wide-angle RGB camera for ordinary usage. The camera model is IMX219-160 which gives the highest resolution of 3280x2464 at 15 fps, the camera side field of view of 160^o. This camera is located on the backside of the user's wrist, consequently, it could capture images of the whole hand back as well as the the object to be controlled. The MPU6050 module is responsible for accelerometer and gyroscope signals. It is located under the camera and connected to an embedded computer (i.e., Jetson Nano) to transfer data through a CSI port. Due to the size, the embedded computer and the battery are attached to the operator's waist. Fig. 4.1 illustrates the final design of our prototype. Compared to the existing prototypes, this device gives a possibility to exploit other multimodal sensors such as accelerometer and gyroscope.

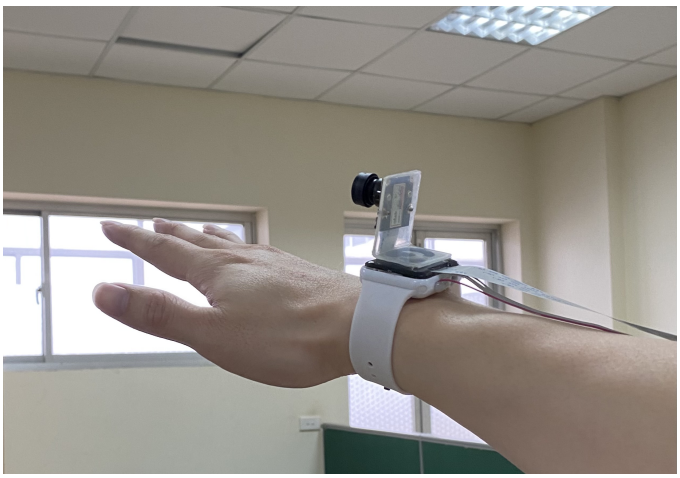

Figure 4.1: Illustration of the prototpe designed in our work

4.2 DATASET

4.2.1 DESIGN OF HAND GESTURE SET

As aforementioned, there are some existing works on human hand gesture recognition using wearable sensors. However, each work has designed and built a proper dataset for a specific

purpose which would be impossible to generalize or re-use in different applications. Moreover, most of them are not publicly available. Our work aims to develop a system for controlling home appliances (e.g., fans, air conditioners, television) through hand gestures captured by wrist-worn devices. As a consequence, a new set of hand gestures is designed. The main required characteristic for any new gesture set in human-machine interaction is that they have to be intuitive, distinguishable, easy to be memorized, and performed by native users.

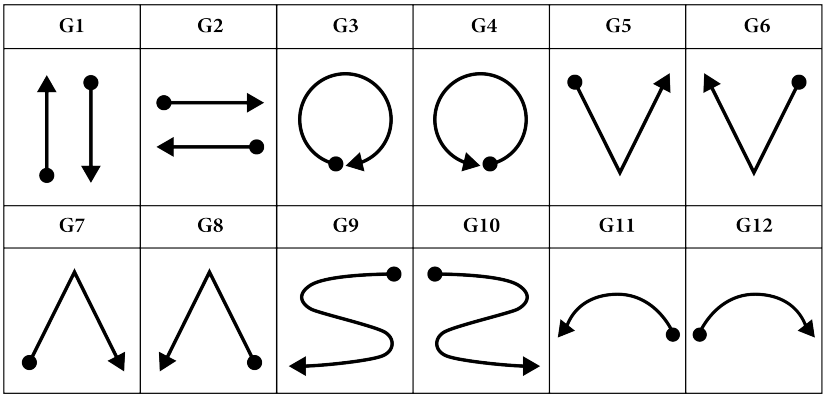

Figure 4.2: Illustration of 12 hand gestures designed in our work

After a careful design process, a set of twelve dynamic hand gestures named from G_1 to G_{12} has been designed. The trajectory of each hand gesture is shown in Fig. 4.2. Each hand gesture can be mapped to one command to control in-home appliances, such as turning on/off the light switch and increasing/decreasing the temperature of the air conditioner. Intuitively, according to the subjects participating in data collection, these gestures are easy to implement and memorize. The shapes of gesture trajectories are distinctive by appearance.

4.2.2 Data collection and annotation

To ease data collection by a large number of subjects in different places, we have produced four similar kits, each of which includes wrist-worn sensors, Jetson nano, and a power supply. In all data acquisition sessions, image data was captured at 30fps with a resolution of 1280x720 pixels, and accelerometer and gyroscope signals at approximately 110hz with the angular value domain set between -2000 and +2000 degrees/s and the acceleration value domain set between -4g and 4g.

We invite 50 volunteers (33 men and 17 women, aged from 10 to 65 years old) to implement twelve designed gestures while standing or sitting in different environments, such as offices, lab rooms, and homes. In each collection session, volunteers are explained how to wear the device and implement the gestures correctly. Each subject performs in his natural manner 12 gestures; each gesture is repeated from 2 to 12 times. All visual frames from the camera, accelerometer data (a_x, a_y, a_z) and gyroscope data (g_x, g_y, g_z) are synchronized and stored in the memory of the embedded device Jetson nano. Furthermore, each gesture's starting and ending times are marked during data acquisition via a keypad or a remote control de-

vice to facilitate the labeling process. Therefore, all gesture instances can be automatically segmented. Totally we conducted 50 sessions of collection for 50 subjects to obtain a multimodal dataset of 5408 gesture samples.

Figure 4.3 shows an example of gesture G_3. The top part of the figure shows some frames extracted from a sequence captured by a third-person view camera that observes the subject in action. The middle part of the figure presents three channels (a_x, a_y, a_z) of accelerometer data (dashed lines) and (g_x, g_y, g_z) of gyroscope data (continuous line). We also scaled accelerometer data by a factor of 50 for better visualization. The bottom part of the figure shows the corresponding frames extracted from the wrist-worn camera. We denote extracting time above each frame.

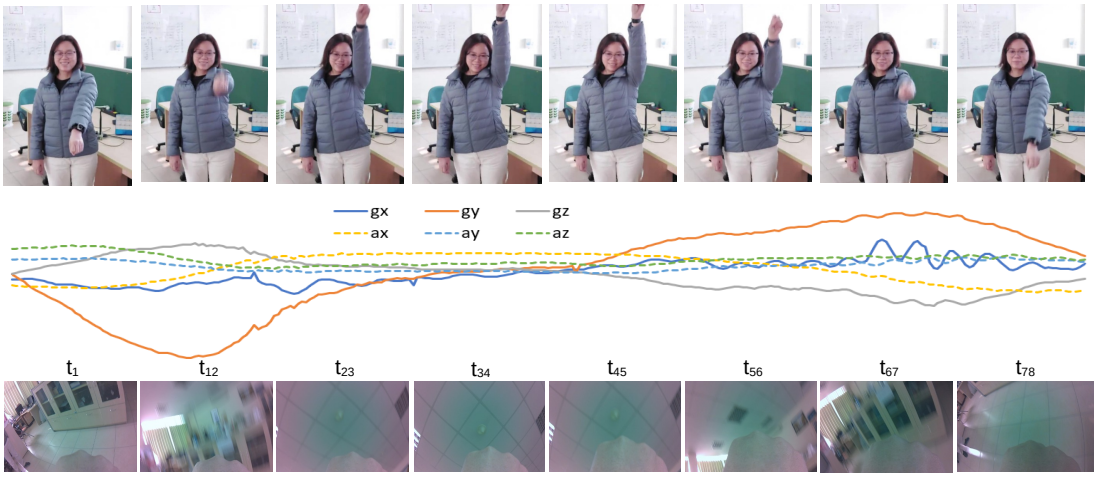

Figure 4.3: An example of gesture G_3: 8 frames uniformly extracted from the original sequence from the third person view are shown in the top row while the accelerometer and gyroscope signals are in the middle row and the eight corresponding frames captured by the prototype are in the bottom row.

We compare our dataset with some relevant existing ones in Tab. 4.1. **Our dataset is significant in terms of the number of subjects, number of gesture instances, number of modalities, and availability**. This dataset and pre-trained models used in the evaluation are available at `https://www.mica.edu.vn/perso/Tran-Thi-Thanh-Hai/MuWiGes.html`.

4.3 Methodologies

4.3.1 CNNs for RGB/OF

In addition to RGB video collected from our prototype, we also consider converting to Optical Flow stream. Optical flow is the pattern of apparent motion of image objects between two consecutive frames caused by the movement of objects or camera. It is a 2D vector field where each vector is a displacement vector showing the movement of points from the first frame to

Table 4.1: Comparison with existing dynamic gesture datasets captured from wrist camera (na. stands for not available).

Dataset	Instances	Activities	Classes	Subjects	Scene	Modality	Availability
Maekawa et al. 2010 [7]	na.	ADL	15	10	Home-like, Lab.	RGB, Acc., Sound	No
Ohnishi et al. 2016 [8]	628	ADL	23	20	na.	RGB	Yes[1]
Jiang et al. 2017 [4]	na.	Air gesture	3	10	na.	sEMG & IMU	No
Chen et al. 2018 [3]	1350	HCI	10	15	Room, Outdoor	RGB	No
Yeo et al. 2019 [14]	na.	Finger Tapping	5	10	na.	IR	Yes[2]
Wu et al. 2020 [13]	na.	Finger Tapping	5	6	Indoor, Outdoor	RGB	No[3]
Our dataset 2021	**5408**	HCI	12	**50**	Home/Office	RGB, Acc., Gyro.	**Yes**

the second frame. For action recognition, optical flow has been used as an efficient feature for action representation. It sometimes produces better performance than RGB stream because it can characterize better movements of objects in the scene. In this project, we compute the dense optical flow using Gunnar Farneback algorithm. Fig. 4.4 illustrates two consecutive RGB images and the corresponding dense optical flow.

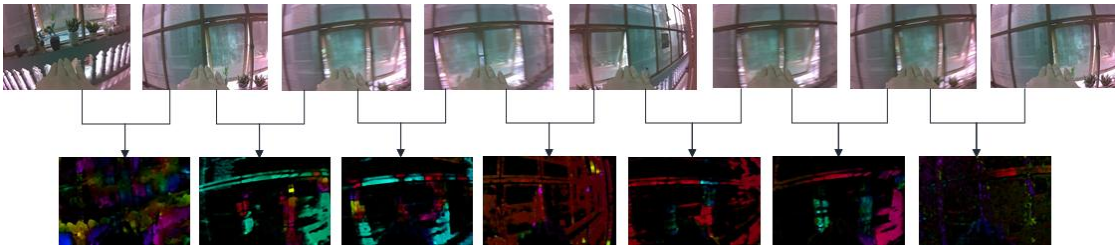

Figure 4.4: Illustration of two consecutive frames and the optical flow..

To evaluate the quality of dataset, different 3D models will be deployed (i.e., C3D, R(2+1)D, R3D, MobileNet, MoviNet, and EfficientNet). These models have proven effective for RGB and OF streams, especially with MoviNet family. We then take advantage of it and proposed to combine the RGB stream with the optical flow stream to boost the recognition rate using MoviNets. The combination of RGB and optical flow has been shown to be more efficient than single-stream because optical flow captures better the motion information while RGB extracts the appearance features of the scene. Our proposed method is a late fusion of two outputs from two CNN streams as follows. The framework is depicted in Fig. 4.5.

The inferences of CNN models with RGB and OF channels on our WiMuGes dataset are listed on Table. 4.2 and Table. 4.3. These reported results indicated that our proposed combination of RGB and OF using moviNet could help impove the robustess of the models

4.3.2 Deep learning for IMU sensors

- **Transformer-based human action recognition from IMU sensors**: In this project, we make use of transformer models to investigate their performance on action recognition using inertial sensors. The transformer model relied upon the attention mechanism to find time series correlations between features and allows for large parallelism of time

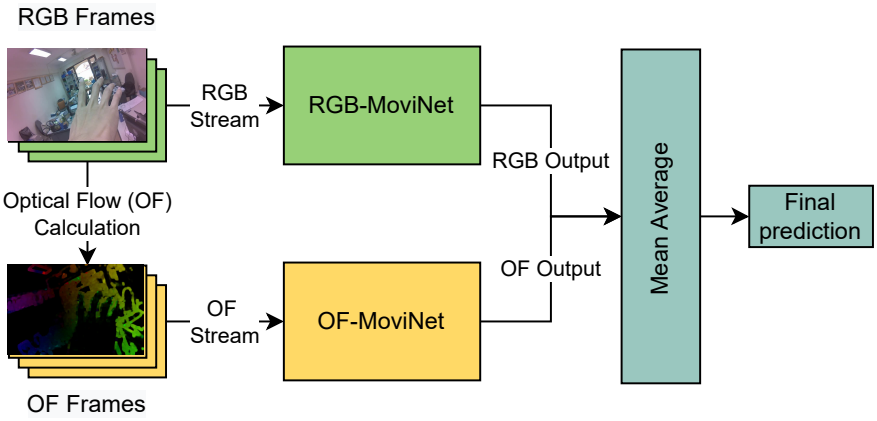

Figure 4.5: The proposed vision-based framework for hand gesture recognition.

Table 4.2: **Cross-subject evaluation**: Comparison of experimental results of experimented CNN models for hand gesture recognition. Acc. and Gyr. stand for Acceleration and Gyroscope.

Model	Modality	Pre-trained	Top-1 Acc (%)	Top-5 Acc (%)	Frame Resolution	Params	GFLOPs
C3D	RGB	Kinetics 700	70.88	96.33	112x112	63.37M	77.3
R3D-50	RGB	Kinetics 700	88.54	99.21	224x224	46.2M	80.1
EfficientNet3D-b0	RGB	Kinetics 600	52.94	89.79	224x224	4.72M	0.06
MobileNet3D_v2_1.0x	RGB	Kinetics 600	67.42	96.26	112x112	2.4M	1.1
R3D-18	RGB	Kinetics 700	76.85	95.35	224x224	33.2M	65.9
R(2+1)D-18	RGB	Kinetics 400	90.46	99.32	112x112	31.3M	81.4
MoviNet-a0	RGB	Kinetics 600	92.44	99.38	172x172	1.9M	1.8
MoviNet-a2	RGB	Kinetics 600	93.28	99.66	172x172	4M	4.9
MoviNet-a2*	RGB	Kinetics 600	94.81	99.66	224x224	4M	4.9
MoviNet-a5	RGB	Kinetics 600	92.78	99.55	172x172	17.5M	23.9
MoviNet-a2*	OF	Kinetics 600	95.59	99.43	224x224	4M	4.9
our two-stream MoviNet-a2*	RGB+OF	Kinetics 600	98.48	99.83	224x224	8M	9.8

Table 4.3: **Cross-scene-subject evaluation**: Comparison of experimental results of experimented CNN models for hand gesture recognition. Acc. and Gyr. stand for Acceleration and Gyroscope.x

Model	Modality	Pre-trained	Top-1 Acc (%)	Top-5 Acc (%)	Frame Resolution	Params	GFLOPs
C3D	RGB	Kinetics 700	64.08	94.74	112x112	63.37M	77.3
R3D-50	RGB	Kinetics 700	81.05	97.23	224x224	46.2M	80.1
EfficientNet3D-b0	RGB	Kinetics 600	68.67	95.14	224x224	4.72M	0.06
MobileNet3D_v2_1.0x	RGB	Kinetics 600	59.90	94.51	112x112	2.4M	1.1
R3D-18	RGB	Kinetics 700	73.30	95.31	224x224	33.2M	65.9
R(2+1)D-18	RGB	Kinetics 400	83.04	98.08	112x112	31.3M	81.4
MoviNet-a0	RGB	Kinetics 600	94.87	98.82	172x172	1.9M	1.8
MoviNet-a2	RGB	Kinetics 600	89.41	98.93	172x172	4M	4.9
MoviNet-a2*	RGB	Kinetics 600	92.73	98.65	224x224	4M	4.9
MoviNet-a5	RGB	Kinetics 600	90.06	97.52	172x172	17.5M	23.9
MoviNet-a0	OF	Kinetics 600	91.27	99.04	172x172	1.9M	1.8
our two-stream MoviNet-a0	RGB+OF	Kinectics 600	96.23	99.38	172x172	3.8M	3.6

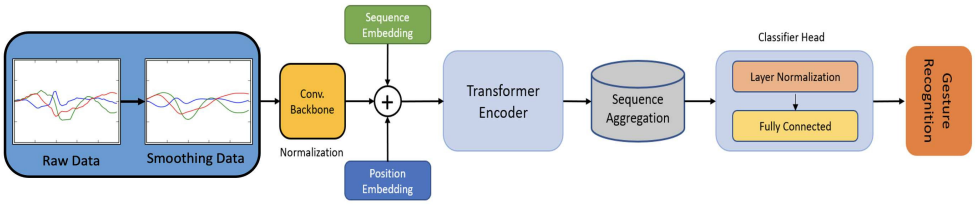

Figure 4.6: The transformer model for action recognition.

series computations, which is different from cyclic lattice neurons that repeat in time series. Another advantage of transformers is longer path lengths between objects in time series, allowing more accurate learning of context in long time series, a claim by Vaswani et al. [11]. Besides using Transformers, we investigate how the size and stride of sliding windows impact the performance of transformer models. We evaluate the studied model on 03 datasets that have been published to the research community. The results showed that the transformer model achieves better results than the method initial on the three evaluated datasets.

The proposed recognition framework is depicted in Fig. 4.6. Given a sample of data $\mathbf{S} \in \mathbb{R}^{k\times 6}$ (3 values of accelerometer and three values of gyroscope), k is the number of data samples (window size). The problem of human action recognition is considered as sequence-to-one where input is a sequence of sensor measurements and output is the label of action. The transformer will be used to extract latent embedding from raw data, and a Multi-layer Perceptron (MLP) can be employed for classification. Fig. 4.6 shows a framework of three steps: pre-processing, transformer encoding, and classification.

We use three available datasets to evaluate the studied method, including CMDFALL [10], C-MHAD [12] and DaLiAc [6]. The data of each gesture/activity used will include six signals: three signals in 3 x, y, z axes of the accelerometer sensor and three signals in x, y, z axes of the gyroscope sensor. Tables 4.5, 4.7, 4.6 show the experimental results obtained by Transformer on three datasets respectively. Details of this work are presented in our paper [5].

- **Time series to image for action recognition from IMU sensors**: Early or conventional methods for human action recognition from inertial sensors usually extract statistical features. However, these features usually require domain knowledge about data. In addition, feature extraction and classification are considered two separate problems. Inertial sensors provide data in the form of multivariate time series. As a result, deep models such as recurrent neural networks RNN, deep belief networks (DBN), stacked autoencoders, and using 1D and 2D CNN can be deployed for action recognition from inertial sensors. Recently, Ahmad et al. proposed different techniques to convert a time series into images such as Gramian Angular Field (GAF) Images, Markov Transition Field (MTF) Images, and Recurrence Plot (RP) Images [1]. They achieved better performance than using raw data.

  In this project, we deploy Gramian Angular Field (GAF) technique to encode inertial

Table 4.4: Information about datasets

CHARACTERISTICS	CMDFALL	CHMAD	DaLiAc
Types of activities	Daily activities	Hand gesture	Daily activities
Number of gestures	S2: 6 S3: 20	5	13
Data types	Acc & Acc	Acc & Gyro	Acc & Gyro
Sampling rate (Hz)	50	50	204,8
Total activities /gestures obtained	2353	1018	266
Average number of samples/activity)	305,7	102,92	17619.64
Split training, test	Even ID train Odd ID test data	K − Fold Cross (K = 10)	Train test split (30% test)

Table 4.5: Comparison of recognition accuracy (%) on the groups S_2 and the group S_3 of CMDFALL dataset

Group	Window size	Accuracy
S2	64	64.33
	96	64.71
	128	64.99
S3	64	58.07
	96	58.17
	128	60.19

Table 4.6: Comparison of recognition accuracy (%) on C-MHAD dataset

Dataset	Window size	Accuracy
C-MHAD	64	88.86
	96	97.93
	128	99.56

Table 4.7: Recognition accuracy (%) on DaLiAc dataset

Dataset	Window size	Accuracy
DaLiAc	64	82.03
	96	85.56
	128	96.86

sensor data into images because it produces the best performance. The angular perspective of the encoded image can be fully utilized by considering the sum between each point to identify the temporal correlation within different time intervals. We used the summation method for Grammian Angular Field [1].

For inertial sensor data, we first convert time series into GAF images and then utilize ResNet - a 2D CNN to do the classification.

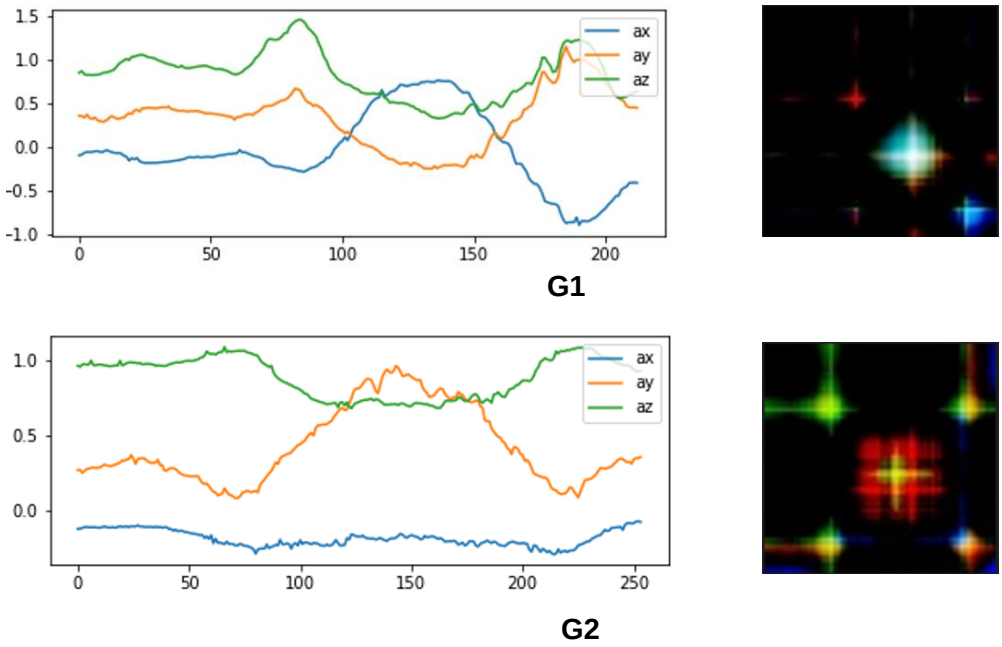

Figure 4.7: Illustration of two GAF images (on the right) generated from the corresponding acceleration data of two different gestures (on the left).

4.3.3 Other researches

In the project, we also conducted research on different topics such as Bayesian network; shape representation; network quantization, network lightweight, and federated learning. At the beginning of our research, we evaluated the different existing datasets. Experiments on

Table 4.8: Comparison of experimental results of experimented CNN models for hand gesture recognition. Acc. and Gyr. stand for Acceleration and Gyroscope.

Model	Modality	Pre-trained	Top-1 Acc (%)	Top-5 Acc (%)	Frame Resolution	Params	GFLOPs
ResNet-18	Acc.	ImageNet	68.9	97.2	224x224	11.18M	3.64
ResNet-50	Acc.	ImageNet	69.0	97.7	224x224	23.53M	8.24
ResNet-18	Gyr.	ImageNet	70.5	98.4	224x224	11.18M	3.64
ResNet-50	Gyr.	ImageNet	70.6	98.5	224x224	23.53M	8.24

our built dataset will be future works.

- **Bayesian network**: Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) provide state-of-the-art performance in various tasks. However, these models are faced with overfitting on small data and cannot measure uncertainty, which has a negative effect on their generalization abilities. In addition, the prediction task can face many challenges because of the complex long-term fluctuations, especially in time-series datasets. Recently, applying Bayesian inference in deep learning to estimate the uncertainty in the model prediction was introduced. This approach can be highly robust to overfitting and allows estimating uncertainty. In this project, we propose a novel approach using Bayesian inference in a hybrid CNN-LSTM model called CNN-Bayes LSTM for time-series prediction. The experiments have been conducted on two real time-series datasets, namely sunspot and weather datasets. The experimental results show that the proposed CNN-Bayes LSTM model is more effective than other forecasting models in terms of Root Mean Square Error (RMSE) and Mean Absolute Error (MAE), as well as for uncertainty quantification. Details of this work are presented in our paper [7].
- **Shape representation**: Analyzing the reflectionally symmetric features inside an image is one of the important processes for recognizing the peculiar appearance of natural and man-made objects, biological patterns, etc. In this work, we point out an efficient detector of reflectionally symmetric shapes by addressing a class of projection-based signatures that are structured by a generalized R-transform model. To this end, we first prove the R-transform in accordance with reflectional symmetry detection. Then different corresponding R-signatures of binary shapes are evaluated in order to determine which corresponding exponentiation of the R-transform is the best for the detection. Experimental results of detecting single/compound contour-based shapes have validated that the exponentiation of 10 is the most discriminatory, with over 2.7% better performance on the multiple-axis shapes in comparison with the conventional one. Additionally, the proposed detector also outperforms most of the other existing methods. This finding should be recommended for applications in practice. Details of this work are presented in our paper [11].

  Besides, a novel method for detecting rotational symmetry is addressed by introducing a new concept of semi-shapes to overcome the main problem of projection-based approaches for studying symmetric rotational properties of an arbitrary shape. It is due to the fact that in the classical approaches, projection cues are periodical with a pe-

riod of preventing exploitation of rotational properties. We then propose the profile of semi-shapes as a signature of the shape together with a simple yet efficient technique to determine the rotation symmetry of an arbitrary shape by considering the correlation of this signature and its circular shift. A new measure is also introduced to determine how good the rotational symmetry would be. Experiments on singlecompound-contour shapes have clearly corroborated the efficacy of our proposal. Details of this work are presented in our paper [8].

- **Federated learning**: Nowadays, Federated Learning (FL), a training paradigm in which data are stored locally and used to train the model on client devices, has emerged thanks to the growing computational power of client devices as well as the concern about transmitting private information. In FL, multiple parties jointly train a model with their local dataset. In this way, the privacy of clients' data is kept. However, FL has to face communication costs during training as the model and model weights have to be sent to the server to update the global model. Several methods have been proposed to address this issue. In this work, we propose model weight compression and encoding during model uploading for Federated Averaging (FedAvg) - a widely used framework in FL. Our weight compression is inspired by the Sparse Ternary Compression algorithm with a modification to be applicable to FedAvg. We also utilize compressed weights' characteristics to encode them; hence the communication cost can be reduced. The experimental results on an image classification task with the MNIST dataset demonstrate that our method is able to reduce communication cost without considerably worsening the model accuracy. Details of this work are presented in our paper [4].
- **Hand pose estimation:** Existing methods for human hand pose estimation usually explore spatial relationships among hand joints in a single image to estimate the 3D hand pose. By doing so, the temporal constraints among hand poses are under-investigated. In this paper, we propose SST-GCN that incorporates both spatial dependencies and temporal consistencies to improve 3D hand pose estimation results. Our method is based on an existing spatial-temporal GCN for 3D pose estimation. In addition, we introduce a new loss function that takes geometric constraints of hand structure into account. Our proposed method takes a 2D hand pose as an input to estimate the 3D hand pose. Finally, we evaluate our method on the First-Person Hand Action Benchmark (FPHAB) dataset. The experimental results show that the proposed method gives promising results in comparison with the original ST-GCN network. Details of this work are presented in our paper [3].
- **Meta-learning** a primary goal of Meta-learning is to understand the interaction between a mechanism of learning and the contexts in which that mechanism is applicable. It assists machine learning systems with the process of model selection by the meta-knowledge acquired from the learning algorithms. In other words, via meta-learning, the networks could "learn how to learn" from prior experience or learned knowledge. This network learns to deal with the tasks via two stages: one acquires meta-knowledge from machine learning systems, and the other adapts that knowledge to the new problems (domain) with the objective of identifying a suitable learn-

ing algorithm or technique for them. Take advantage of meta-learning, many few-shot learning methods have been proposed to solve the problem of data-hungry. The meta-learning framework for few-shot learning follows the key idea of learning to learn. In this project, firstly, we evaluate some SOTA and baseline meta-learning methods for few-shot learning on our proposed dataset (MuWiGes) based on two aspects: (i) Using traditional meta-learning methods for still images converted from video (e.g., MAMN, Reptile, and Meta-Baseline, etc.)(ii) Advanced meta-learning methods for videos based on spatial and temporal alignment (e.g., HyRSM, STRM, ATA, etc.). Secondly, based on the evaluation of these few-shot learning methods according to two criteria: accuracy and model size, we will conduct research towards developing few-shot learning algorithms for wrist-worn edge devices. The results of these studies will be published after MuWiGes dataset is accepted.

4.4 Application

In this project, we develop an application to control home appliances using hand gestures captured by our wrist-worn multimodal device. The following are the main steps to implement the application:

- Step 1: Assignment of hand gestures to commands. For example: turn on, turn off; increase, decrease one level of brightness; increase, decrease two levels of brightness and so on.
- Step 2: Design the diagram of the state transition of the lamps.
- Step 3: Implementation of the controlling system.

In the following, we detail how to implement the controlling system and the evaluation result.

4.4.1 Implementation of the controlling system

The controlling system composes of three components: a wearable device; a transceiver and a lamp. The wearable device is in charge of both data collection; data processing and controlling. The transceiver connects to the device and the lamp through the local area network and zig-bee communication protocol. The transceiver receives gesture commands from the device and transmits the encoded signal to control the lamp.

As the data streams come continuously, we have to develop a continuous gesture recognizer. We adopt the sliding window technique to take a segment of video for recognition. The window size is determined as the mean length of gestures in the training set. Besides, we have to include one non-gesture class to deal with the non-gesture sequences. In total, we have 13 classes to be trained.

4.4.2 Evaluation of the controlling system

We integrated the model R(2+1)D in our controlling system. The testing with other models is ongoing. Fig. 4.8 illustrates a new user using the device with the implementation of a hand gesture. Fig. 4.9 shows the recognition result provided by our system when the user realizes the gestures. On the left is the correct recognition and on the right is the wrong recognition result. We conducted the evaluation on 272 samples implemented by users, including both

Figure 4.8: An new user in a senario of using the device

Figure 4.9: Recognition results by our system

gestures (31 samples) and non-gestures (241). We manually annotated the ground-truth labels and compared them with the predicted labels. The result shows that all 31 gestures are correctly recognized. That means the recall is 100%. However, many non-gestures are also recognized as gestures (165/241 = 68%) false recognition rate. That means the system produces many false alarms. One reason is that non-gestures are too various. Currently, the non-gesture samples for training the model are quite limited. We are conducting experiments with more subjects.

5 Conclusions and Perspectives

5.1 Conclusions

In this report, we have presented our main finding and outcomes:

- Firstly, we have designed a wrist-worn prototype that enables capturing multimodal data (image and motion) of hand gestures and surroundings. The prototype can be utilized not only for human-machine interaction but also for other applications such as healthcare or life logging.
- Secondly, we have designed and collected a dataset of hand gestures for human-machine interaction applications using the designed prototype with a high number of subjects and various environments. This dataset is very special in terms of motion and appearance of the hand because the camera and sensor are mounted at the wrist of the hand, which is the most flexible body part. The multimodal data are synchronized, annotated, and considered the benchmark for evaltheuating hand gesture recognition from wrist-worn sensors.
- We conducted experiments with many base-lines deep models to evaluate and reveal the challenges of the current dataset. The models are studied for both RGB streams and motion streams. We also analyzed the best configuration of CNN models that makes the best trade-off between the accuracy and the memory requirement. Experimental results show promising results but still remain room for improvement.
- We proposed a method that combined two streams (RGB and optical flow) in a unified framework to boost the performance of hand gesture recognition. It helps to increase the accuracy from 94.87% to 96.23%. We also proved that optical flow is very potential when alone. It can provide 91.27%. It shows that appearance remains still an important key for such kind of dataset.
- We conducted other research on Bayesian networks, hand pose estimation, shape characterization, network quantization, meta-learning and initiated the first study in federated learning for action recognition.

Thanks to this grant, a research group that assembles researchers from different faculties and universities in Hanoi has been established to work on a new and interesting research topic of human activity recognition from wearable sensors. We participated to train:

- Two master student in Computer Science (Quynh Khanh Dinh Thi)
- Undergraduate students in Electrical and Electronic Engineering (Thi Thu Hien Le, Viet-Duc Le, Danh Huy Vu, Van Thang Tran)
- Two Ph.D. Students (Trung-Hieu Le, Hong-Quan Nguyen)

All of these people are involved in the project for doing research or/and publishing journal/conference papers.

5.2 PERSPECTIVES

The section Conclusion indicated our initial fundamental research steps toward a final practical application of human-machine interaction using hand gestures from the multimodal wearable sensor. We propose some perspectives from this project, and some of them are being conducted.

- Recently, Deep neural networks (DNNs) have achieved accuracy and impressive performance in various domains. However, DNNs models are faced with overfitting on small data and cannot measure uncertainty problems which has a negative effect on their generalization abilities. For example, DNNs can be unable to take a broad view of new stimuli that are noisy, variable, or ambiguous. Some researchers have shown that DNNs may result in poor recognition when objects are observed in noisy viewing conditions. DNNs models also tend to be too self-confident about their predictions when a confidence interval is provided. A failed prediction can lead to serious outcomes in some domains, such as self-driving cars. Therefore, the Bayesian-based approach that uses probability distribution to estimate the uncertainty in the model prediction can be highly robust to overfitting, giving uncertainty estimation to understand and quantify the uncertainty associated with deep neural network predictions.

  Moreover, Bayesian-based methods do not require a huge amount of training examples. Obtaining sufficient training examples to train deep networks is one of the most challenging issues in action recognition, especially when applying federated learning. In the project, we have incorporated the Bayesian learning scheme for time series prediction. Experiments show the promising results of this approach (this research has been published in MAPR'2022 conference paper). Therefore, in the future, we would like to investigate the use of Bayesian learning for action recognition and FL-based action recognition.

- Meta-learning: We will continue digging into the application of meta-learning-based few-shot learning for hand gesture recognition. This is an entirely new field, and not many researchers invest in in-depth research on it, although it has many applications in daily activities. We will focus more on improving algorithms and meta-learning models for even lower computing power devices like tiny microcontrollers.

- Federated learning: In this project, we initiated a new research topic on federated learning. We have developed a framework allowing the researchers easily evaluate action recognition using federated learning. At the time of writing this report, the framework provides basic functionalities. However, it needs more investment to complete with more federated algorithms in order to deal with noisy clients, a high number of clients, non-idd data distribution, and unbalanced data. Reduction of communication costs will be a new direction in the future.

- Transformer-based model for multimodal recognition: In the project, we have studied and evaluated a range of CNN models for recognition from RGB or motion sensors. Recently, Transformer, a technique inherited from natural language processing, has shown to be very efficient in many tasks. Some works have been applied for video

understanding or accelerometers separately. In the literature, the combination of RGB and accelerometer is still limited. In addition, according to our knowledge, there is not existing work on combining both RGB and motion data using Transformer. Studying the correlation between two modalities must be carefully carried out.

6 PUBLICATION OUTCOMES

The published papers [1,2,3,4,5,6,7,8,9,10,11,12] are attached.

1. Thanh-Hai Tran, Hoang-Nhat Tran, Hong-Quan Nguyen, Trung-Hieu Le, Van-Thang Nguyen, Trung-Kien Tran, Cuong Pham, Thi-Lan Le, Hai Vu, Thanh Phuong Nguyen and Nguyen Huu Thanh, A pilot study on hand posture recognition from wrist-worn camera for human machine interaction, 2021 International Conference On Advanced Technologies For Communications, 14 - 16, October, 2021, Ho Chi Minh City, Vietnam, 10.1109/ATC52653.2021.9598223.
2. Thanh-Hai Tran, Van-Hieu Do, Improving continuous hand gesture detection and recognition from depth using convolutional neural networks, The International Conference on Intelligent Systems Networks. Springer, Singapore, 03/2021, https://doi.org/10.1007/978-981-16-2094-2_10
3. Le, Viet-Thanh, Thanh-Hai Tran, Van-Nam Hoang, Van-Hung Le, Thi-Lan Le, and Hai Vu. SST-GCN: Structure aware Spatial-Temporal GCN for 3D Hand Pose Estimation. In 2021 13th International Conference on Knowledge and Systems Engineering (KSE), pp. 1-6. IEEE, 2021, 10.1109/KSE53942.2021.9648765
4. Dinh, Thi Quynh Khanh, Thanh-Hai Tran, and Thi-Lan Le, Communication cost reduction using sparse ternary compression and encoding for FedAvg. In 2021 International Conference on Information and Communication Technology Convergence (ICTC), pp. 351-356. IEEE, 2021, 10.1109/ICTC52510.2021.9620887
5. Trung-Hieu Le; Thanh-Hai Tran; Cuong Pham, Human action recognition from inertial sensors with Transformer, 2022 International Conference on Multimedia Analysis and Pattern Recognition (MAPR), 10/2022, Phu Quoc, Vietnam, 10.1109/MAPR56351.2022.9924794
6. Hong Quan Nguyen, Thuy Binh Nguyen, Trung Kien Tran, Van Nam Hoang, Thi Lan Le, Thanh Hai Tran, Hai Vu, End-to-end deep learning-based framework for driver action recognition, 2022 International Conference on Multimedia Analysis and Pattern Recognition (MAPR), 10/2022, Phu Quoc, Vietnam, 10.1109/MAPR56351.2022.9924944
7. Thi-Lich Nghiem, Viet-Duc Le, Thi-Lan Le, Pierre Maréchal, Daniel Delahaye, Andrija Vidosavljevic, Applying Bayesian inference in a hybrid CNN-LSTM model for time-series prediction, 2022 International Conference on Multimedia Analysis and Pattern Recognition (MAPR), 10/2022, Phu Quoc, Vietnam, 10.1109/MAPR56351.2022.9924783
8. Thanh Tuan Nguyen, Thanh Phuong Nguyen, Thanh-Hai Tran, Projection of semi-shapes for rotational symmetry detection, The 26th International Conference on Pattern Recognition, August 21-25, 2022, Montréal Québec. 10.1109/ICPR56361.2022.9956227
9. Thanh Nam Nguyen, Thanh-Hai Tran, Hai Vu, An automatic tool for yoga pose grading using skeleton representation, Conference on Information and Computer Science (NICS), October 31-November 01, 2022, Ho Chi Minh City, Vietnam (online unavailable)
10. Thanh Tuan Nguyen, Thanh Phuong Nguyen, Thanh-Hai Tran, Detecting reflectional

symmetry of binary shapes based on generalized R-transform, 2022 International Conference on Multimedia Analysis and Pattern Recognition (MAPR), 10/2022, Phu Quoc, Vietnam, 10.1109/MAPR56351.2022.9924894

11. Trung-Hieu Le, Quoc-Tuan Nguyen, Thanh-Hai Tran, Cuong Pham, Combined local and global features for action recognition from motion sensors, Lecture Notes on Data Engineering and Communications Technologies book series (LNDECT,volume 148), International Conference on Intelligence of Things, 08/2022, Hanoi, Vietnam,
12. Ho, Ha-Dang, Hong-Quan Nguyen, Thuy-Binh Nguyen, Sinh-Thuong Vu, and Thi-Lan Le. "Dynamic Hand Gesture Recognition from Egocentric Videos based on SlowFast Architecture.", Proceedings of 2022 APSIPA Annual Summit and Conference, 7-10 November 2022, Chiang Mai, Thailand
13. Hong Quan Nguyen, Trung-Hieu Le, Trung Kien Tran, Hoang-Nhat Tran, Van Minh Truong, Thanh Hai Tran*, Thi-Lan Le, Hai Vu, Cuong Pham, Thanh Phuong Nguyen, Huu Thanh Nguyen, Dynamic hand gesture recognition from wrist-worn device for human-machine interaction: Prototype, Dataset and Methods, IEEE Sensors (Submitted)

REFERENCES

[1] Z. Ahmad and N. Khan. Inertial sensor data to image encoding for human action recognition. *IEEE Sensors Journal*, 21(9):10978–10988, 2021.

[2] F. Chen, J. Deng, Z. Pang, M. Baghaei Nejad, H. Yang, and G. Yang. Finger angle-based hand gesture recognition for smart infrastructure using wearable wrist-worn camera. *Applied Sciences*, 8(3):369, 2018.

[3] F. Chen, H. Lv, Z. Pang, J. Zhang, Y. Hou, Y. Gu, H. Yang, and G. Yang. Wristcam: A wearable sensor for hand trajectory gesture recognition and intelligent human–robot interaction. *IEEE Sensors Journal*, 19(19):8441–8451, 2018.

[4] S. Jiang, B. Lv, W. Guo, C. Zhang, H. Wang, X. Sheng, and P. B. Shull. Feasibility of wrist-worn, real-time hand, and surface gesture recognition via semg and imu sensing. *IEEE Transactions on Industrial Informatics*, 14(8):3376–3385, 2017.

[5] V.-D. Le, V.-N. Hoang, T.-T. Nguyen, V.-H. Le, T.-H. Tran, H. Vu, and T.-L. Le. A unified deep framework for hand pose estimation and dynamic hand action recognition from first-person rgb videos. In *MAPR*, pages 1–6. IEEE, 2021.

[6] H. Leutheuser, D. Schuldhaus, and B. M. Eskofier. Hierarchical, multi-sensor based classification of daily life activities: comparison with state-of-the-art algorithms using a benchmark dataset. *PloS one*, 8(10):e75196, 2013.

[7] T. Maekawa, Y. Yanagisawa, Y. Kishino, K. Ishiguro, K. Kamei, Y. Sakurai, and T. Okadome. Object-based activity recognition with heterogeneous sensors on wrist. In *International Conference on Pervasive Computing*, pages 246–264. Springer, 2010.

[8] K. Ohnishi, A. Kanehira, A. Kanezaki, and T. Harada. Recognizing activities of daily living with a wrist-mounted camera. In *CVPR*, pages 3103–3111, 2016.

[9] T.-H. Tran and V.-H. Do. Improving continuous hand gesture detection and recognition from depth using convolutional neural networks. In *The International Conference on Intelligent Systems & Networks*, pages 80–86. Springer, 2021.

[10] T.-H. Tran, T.-L. Le, D.-T. Pham, V.-N. Hoang, V.-M. Khong, Q.-T. Tran, T.-S. Nguyen, and C. Pham. A multi-modal multi-view dataset for human fall analysis and preliminary investigation on modality. In *ICPR*, pages 1947–1952, 2018.

[11] 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.

[12] H. Wei, P. Chopada, and N. Kehtarnavaz. C-mhad: Continuous multimodal human action dataset of simultaneous video and inertial sensing. *Sensors*, 20(10):2905, 2020.

[13] E. Wu, Y. Yuan, H.-S. Yeo, A. Quigley, H. Koike, and K. M. Kitani. Back-hand-pose: 3d hand pose estimation for a wrist-worn camera via dorsum deformation network. In *ACM Symposium on User Interface Software and Technology*, pages 1147–1160, 2020.

[14] H.-S. Yeo, E. Wu, J. Lee, A. Quigley, and H. Koike. Opisthenar: Hand poses and finger tapping recognition by observing back of hand using embedded wrist camera. In *ACM*

Symposium on User Interface Software and Technology, pages 963–971, 2019.