Universiteit Leiden
The Netherlands

Bachelor Datascience and Artificial Intelligence

Collecting a State-of-the-Art Event-Based Dataset for In-Vehicle Gesture Recognition

Cindy Wang

Dr. Qinyu Chen & Guorui Lu

BACHELOR THESIS

Leiden Institute of Advanced Computer Science (LIACS)
www.liacs.leidenuniv.nl

15/01/2026

Abstract

Touch interaction with in-vehicle infotainment systems can distract drivers because they often require a visual glance away from the road. Mid-air hand gesture recognition provides a more suitable form of interaction and allows drivers to control system functions without touching the display and with reduced visual demand. However, most existing in-vehicle gesture recognition systems rely on frame-based cameras. These cameras often perform poorly in automotive environments, particularly during fast hand motion and under challenging lighting conditions, where motion blur and sensor noise reduce data reliability. Event-based cameras offer an alternative approach, because their high temporal resolution and wide dynamic range make them more robust to rapid motion and challenging lighting conditions. However, there are currently no available in-vehicle event-based gesture datasets. This thesis addresses this gap with a pilot in-vehicle event-based hand gesture dataset from a DAVIS346 camera, which includes 5 subjects, 16 gesture classes, and five recording conditions: daylight static, daylight driving, nighttime static, sunset static, and nighttime driving.

A convolutional spiking neural network (CSNN) is used to assess the collected dataset. Experiments show reasonable performance when training and testing use the same condition, with 84% accuracy for daylight-only evaluation, 72% for nighttime-only evaluation, 77% for static-only evaluation, and 66% for driving-only evaluation. Generalization across conditions proves difficult; daylight-to-nighttime transfer reaches 13% accuracy, and static-to-driving transfer reaches 12% accuracy. Cross-subject evaluation reports accuracies between 15% and 26%. These results show that performance remains high under matched recording conditions, but drops when conditions or subjects change. This work establishes a baseline for in-vehicle event-based gesture recognition and identifies key challenges for future research.

Acknowledgments

I would like to thank my supervisor, Professor Qinyu Chen, and my co-supervisor, PhD candidate Guorui Lu, for their guidance and support throughout this project. I am also thankful to all the volunteers who participated in the data collection; this dataset would not have been possible without their contributions.

Contents

1 Introduction

1.1 Context and motivation

In-vehicle infotainment (IVI) systems support multiple interaction modalities, such as touchscreen interfaces [FER+14], voice controls [Mou20], and gesture-based interactions [PLGAAC14]. Although these technologies enhance the driving experience, they also raise questions about how much they might distract the driver and the possible impact on road safety [Pic05].
One approach to reduce this type of distraction is to rely on methods that require less visual effort and touch interaction. Gesture-based interaction has been proposed as a potential solution, because gestures can be performed with one hand, involve a small movement range, and require less direct visual attention to the display [GMY+16, TCF+24]. As a result, gesture-based interfaces can reduce visual and manual interference and have the potential to support safer interaction inside the vehicle.
Most existing in-vehicle gesture recognition systems rely on conventional frame-based sensors [KLR+21, OBT14, PLGAAC14, RGF23]. However, under fast hand motion or challenging illumination, these sensors are prone to motion blur, limited dynamic range, and latency–power trade-offs [Del16], which reduce data reliability in real driving scenarios.
Besides these sensor-level limitations, real in-vehicle environments introduce additional sources of visual complexity that are rarely reflected in existing gesture datasets. In practice, gestures are performed in a visually dynamic cabin rather than in front of a static background, and the illumination conditions also vary rapidly. For example, a driver may perform a gesture during transitions from a tunnel to direct sunlight, or at night under the combined influence of dashboard lighting and passing streetlights, which creates sudden brightness changes.
Event-based vision offers an alternative approach to address these challenges, because it operates asynchronously and responds only to pixel-level brightness changes that exceed a given threshold [GDO+22]. As a result, event-based cameras achieve microsecond temporal resolution, a wide dynamic range, and low power consumption, which make them well suited to automotive scenarios with rapid illumination changes and fast hand motion.
However, there is no available dataset that combines in-vehicle hand gesture recordings with event-based cameras. To address this gap, this thesis presents a pilot dataset that was recorded with an event-based DAVIS346 camera in a real in-car setting. The data were collected under different realistic driving and lighting conditions, and this work investigates whether event-based vision can be used effectively for in-vehicle gesture recognition. Moreover, a Convolutional Spiking Neural Network (CSNN) is employed as a baseline to assess the dataset.

1.2 Research question

The absence of an event-based hand gesture dataset inside a real vehicle reveals a clear gap in current research on in-car infotainment interaction. The central research question is as follows: Can event-based vision be used effectively for in-vehicle hand gesture recognition under realistic driving and varied lighting conditions?
The following sub-questions are formulated to investigate the main research question:
RQ1: How does recognition accuracy vary under the same recording conditions for an event-based camera combined with a CSNN in in-vehicle gesture recognition?

RQ2: How much does recognition accuracy vary when the model is trained on daytime data and tested on nighttime data?
RQ3: How much does recognition accuracy vary when the model is trained on static recordings and tested on driving recordings?
RQ4: How much does recognition accuracy vary when the model is evaluated on unseen subjects?

1.3 Contributions

This thesis makes several contributions, it introduces a pilot event-based gesture dataset which is recorded inside a real vehicle in five conditions: daylight static, daylight driving, nighttime static, sunset static, and nighttime driving situations. It develops a baseline CSNN to assess whether event-based data can support in-car gesture classification, and it tests how well this model transfers between different recording conditions. Moreover, it discusses the challenges which were observed during the data collection process and outlines considerations for future dataset design.

1.4 Thesis overview

This work is structured as follows: Section 2 reviews the background of event-based vision, summarizes existing datasets, and gesture-recognition approaches. Section 3 describes the creation of the self-collected in-car gesture dataset, including the recording setup, subjects, conditions, and the final gesture set. Section 4 outlines the CSNN baseline model. Section 5 shows the experiments which evaluate the recognition performance under mixed conditions, such as day-to-night, and static-to-driving transitions. Section 6 discusses the main limitations of the dataset and possible improvements for future work. Finally, Section 7 concludes the thesis.

2 Related work

2.1 In-vehicle gesture recognition with frame-based cameras

Most in-vehicle gesture recognition research relies on conventional frame-based cameras. The DriverMHG dataset provides synchronized RGB, infrared, and depth recordings of micro hand gestures that were performed on a steering wheel inside a driving simulator [KLR+21]. This simulator setup allows data collection under multiple lighting configurations, but it does not capture the continuous illumination changes, motion and vibration that occur during real driving.
Reyes et al. conducted experiments in a real vehicle using a Time-of-Flight depth (ToF) camera, and reported recognition accuracy of up to 90% with a Convolutional Long Short-Term Memory Neural Network model [RGF23]. However, the authors report that ToF measurements are sensitive to noise and that partial depth information is often lost during data capture. They also emphasize the need for further evaluation under fully realistic driving conditions, where illumination and motion vary continuously.
Similarly, Ohn-Bar and Trivedi developed a real-time vision-based system, but their data collection took place in a parking lot [OBT14], which means that their data provide a limited representation of natural driving behavior and the complex variability of real-world environments. These limitations motivate the exploration of alternative methods that can operate robustly under challenging automotive conditions.

2.2 Event-based vision and its application in automotive environments

Event-based sensors differ from conventional frame-based sensors; instead of capturing full images at fixed intervals, they respond to brightness changes at individual pixels, generate an event whenever the change exceeds a threshold, and output a tuple of (x,y,t,p), where (x,y) denotes the pixel location, t the timestamp, and p the polarity of the change [GDO+22]. Event-based vision sensors have temporal resolution on the order of microseconds and achieve a dynamic range (> 130 dB) well above that of conventional frame-based cameras (60 dB), which makes them suitable to capture fast motion under challenging lighting conditions, such as during tunnel exits or under strong sunlight [ZTO+18].
Event-based sensors have also been explored in real-world automotive environments. For example, the Multi Vehicle Stereo Event Camera Dataset (MVSEC) [ZTO+18] and the more recent Driving Stereo Event Camera (DSEC) [GAGS21] provide driving datasets from event cameras under challenging illumination conditions, such as direct sunlight, dusk, and night. These datasets focus on outward-facing road scenes rather than in-car interaction, but they show that event cameras can handle realistic driving conditions where frame-based cameras often struggle. The success of such event cameras in handling rapid illumination changes and motion in outward-facing automotive scenarios suggests potential benefits for in-car applications.

2.3 Event-based gesture recognition datasets

Event-based cameras have been applied to gesture recognition in several prior studies. The DAVIS-Gesture128 dataset captures 11 gesture classes from 29 participants using a DAVIS128 camera under three lighting conditions [ATB+17]. Participants performed gestures in front of a static back-

ground. Similarly, the ASL-DAVIS dataset provides a similar experimental setup, with recordings of American Sign Language gestures from five participants in a stable office environment [Pix]. These datasets follow a similar recording setup. Gestures are performed in front of static backgrounds and under controlled lighting conditions, which reduces background activity and limits illumination variation. As a result, the recognition task is simplified and the datasets are suitable for controlled benchmarking. However, this controlled setting does not reflect the visual complexity in real-world automotive scenarios. Table 1 summarizes the key differences between existing gesture datasets and this work.

Dataset	Subjects	Gestures	Environment	Lighting	Vehicle motion
DriverMHG [KLR+21]	25	7	Simulator	5 conditions	Static
Reyes et al. [RGF23]	83	6	Real car	Varied	Static
Ohn-Bar [OBT14]	8	19	Real car	Outdoor	Slow driving
DAVISGesture128 [ATB+17]	29	11	Lab	3 conditions	N/A
ASL-DAVIS [Pix]	5	24	Office	Stable	N/A
This work	5	16	Real vehicle	3 conditions	Static + Driving

Table 1: Comparison of datasets.

Without the in-vehicle dataset which was collected in this work and its evaluation, it is difficult to determine whether event-based vision still has its advantages over frame-based vision in real automotive environments.

3 In-vehicle gesture dataset

3.1 Data collection equipment and recording configuration

Throughout the entire project, all recordings were conducted using the iniVation DAVIS346 event camera [ini19] and the DV software platform [ini25]. The camera has a spatial resolution of 346x260 pixels. For each gesture class, a separate continuous recording of 15 seconds was captured, and participants repeatedly performed the same gesture at a self-selected speed. All recordings were saved in the AEDAT4 format.
Participants were required to perform all gestures with their right hand, because the dataset was collected in the Netherlands, where vehicles are left-hand drive. In this configuration, the right hand offers a larger interaction space near the center display, whereas the left side is mainly constrained by the door and the window.
The data collection consisted of two phases which were conducted on separate days within a one-month period. The first phase was a preliminary indoor session to identify the most comfortable gestures for participants. Based on participant feedback, the second phase involved in-vehicle recordings using the refined gesture set. The camera position was kept fixed within each phase to ensure consistency.

3.2 Dataset preparation

3.2.1 Gesture design

Each subject might perform the same gesture in different ways. To address this variability, a predefined gesture set (Table 4) was used to ensure consistency during data collection. In total 24 gesture variations were designed; some gestures represented the same command, but differed in execution style. For example, a leftward swipe could be performed with one finger or with the entire hand moving left. GIFs of the gestures were provided to the participants to help them perform each gesture in a standardized manner.

3.2.2 Indoor recording and gesture selection

A preliminary experiment was conducted in an indoor setting to determine the final gesture set. Three participants (one female, two males, aged between 20 and 30, all right-handed and holding valid driving licenses) took part under the same lighting conditions.
The camera was mounted on a table approximately 25 cm from the participant, positioned slightly to the right side and angled toward the gesture area. Participants were seated and, before each gesture recording, watched the corresponding GIF to learn the gesture. For the actual recording, they assumed a driving posture with their hands positioned in front of them at approximately steering-wheel height and performed the gesture repeatedly for about 15 seconds using their right hand. To capture natural gesture variability, participants were instructed to perform the gestures at a self-selected comfortable speed, as long as the required gesture style was followed and remained consistent. After completing each gesture, the participants provided feedback on which variants felt more convenient and intuitive. Based on their feedback, the eight least preferred variants were removed. Figure 1 shows the final set of hand gestures for the in-vehicle dataset collection, the density of the event-based representations depends on the speed of the gesture. Faster movements

generate more events and denser patterns, whereas slower movements produce fewer events and sparser patterns. The RGB images are provided as a visual reference to help interpret the gestures and are not used in the experiments or the neural network.

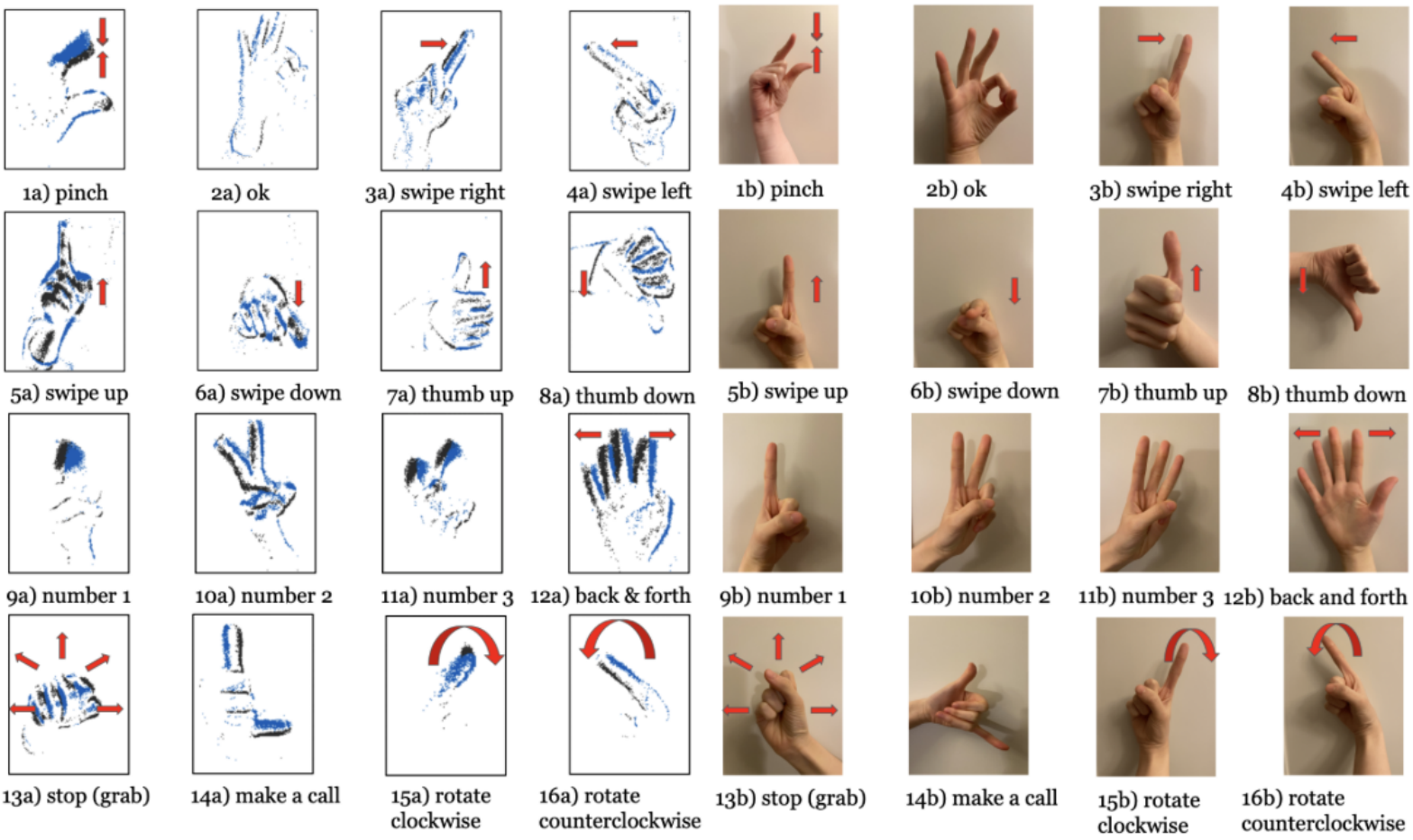


Figure 1: Visualization of the 16 collected hand gesture classes. (1a–16a) Event-based representations. (1b–16b) Corresponding RGB frames for visual reference.

The final dataset and the corresponding motion descriptions are summarized in Table 2. For the pilot dataset, each gesture was assigned a descriptive label representing a typical in-vehicle interaction command (see Table 5).

ID	Gesture	Motion description
0	pinch	Thumb and index finger pinch together
1	OK	Hand enters the camera view first, then forms the "OK: gesture
2	swipe_right	Only the index finger swipes to the right
3	swipe_left	Only the index finger swipes to the left
4	swipe_up	The index finger swipes upward
5	swipe_down	The index finger swipes downward
6	thumb_up	Thumbs-up gesture
7	thumb_down	The thumb points downward while the hand moves slightly
8	number1	The index finger is extended
9	number2	The index and middle fingers are extended
10	number3	Three fingers are extended
11	back_and_forth	Five fingers move left and right repeatedly
12	stop_grab	The hand performs a grabbing motion
13	make_a_call	The thumb and pinky are extended; the hand moves slightly
14	rotate_clockwise	The index finger rotates clockwise
15	rotate_counterclockwise	The index finger rotates counterclockwise

Table 2: List of hand gestures and their corresponding motion descriptions.

3.3 Real in-vehicle data collection

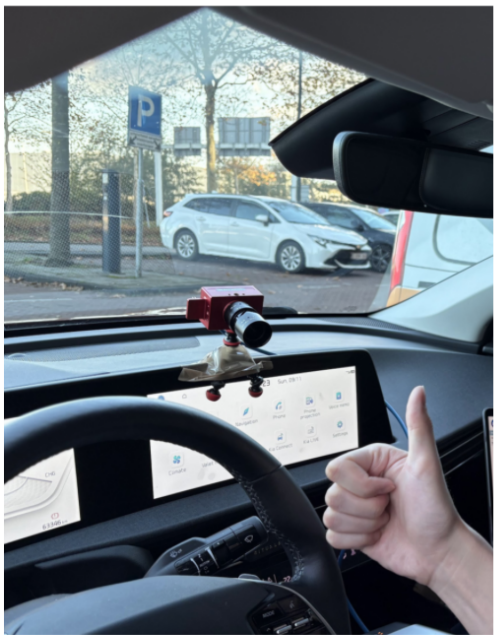
3.3.1 Recording setup and conditions

Five recording conditions were designed to evaluate the collected dataset under varying illumination and motion settings. These conditions combined three illumination scenarios (daylight, nighttime, and sunset) with two motion states (static and driving): daylight static, daylight driving, nighttime static, nighttime driving, and sunset static. In static conditions, the vehicle remained stationary, whereas in driving conditions, recordings were collected when the vehicle was in motion. The sunset static condition was recorded with the vehicle stationary under natural sunset illumination during the early evening.

The dataset was collected in a Kia EV6, which provided a realistic in-car environment for data collection. Inside the vehicle, the camera was mounted on the top edge of the central display using a small stand secured with tape (see Figure 2). The camera was directed toward the area below the display where in-vehicle gestures are performed and angled slightly downward to capture hand movements without the need for the driver to adjust their arm position.

(a) Front view of the camera setup inside the vehicle during daylight.

(b) Front view of the camera setup inside the vehicle during nighttime.



(c) Side view showing the camera position relative to the central display.

Figure 2: Overview of the in-vehicle recording setup from two perspectives, the event camera is mounted above the central display.

3.3.2 Participants

For the in-car recordings, five participants took part in the data collection, with ages ranging from 20 to 30 years. The group included two females and three males. Four participants were right-handed, and one participant was left-handed, but performed all gestures with the right hand as instructed. Four participants held a valid driving license. Each participant contributed data under one or more recording conditions. Table 3 summarizes the recording conditions which were contributed by each participant, where D denotes daytime, N denotes nighttime, and S denotes sunset.

Subject	D Static	D Drive	N Static	N Drive	S Static
Subject 1	X	X	X		
Subject 2			X	X	
Subject 3	X		X		
Subject 4	X		X		
Subject 5					X

Table 3: Recording conditions covered by each subject.

As shown in Table 3, the distribution of recording conditions was uneven due to practical constraints. Although four participants held a valid driving license, only two participated in the driving conditions due to safety and scheduling considerations. As a result, not all participants completed all conditions.

A total of ten recording sessions were conducted across all participants and conditions. Each session contained recordings of 16 distinct gestures. In total, the dataset consists of 160 gesture samples. The preprocessing of the collected data is described in Section 4.

4 Methodology

This section describes the full pipeline, from data collection to preprocessing, model training, and evaluation (see Figure 3). In this pipeline, the collected event-based recordings serve as the starting point for all subsequent processing steps. Raw event streams are first recorded in a real vehicle, segmented into fixed-length temporal clips, and then converted into voxel-grid representations, which are then fed as input to a CSNN. Finally, the trained model is evaluated using multiple experimental settings.

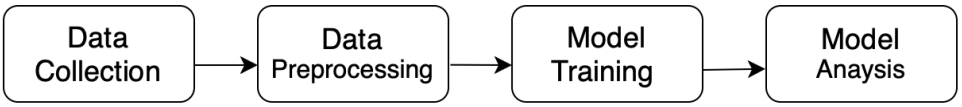


Figure 3: Overview of the methodological pipeline. Event-based gesture recordings are collected in a real vehicle, preprocessed into voxel-grid representations, and used to train a CSNN baseline, which is evaluated under different experimental settings.

4.1 Data collection

During data collection, participants were seated in the driver's seat of the vehicle. Before each recording, participants viewed the GIF demonstration to familiarize themselves with the required gesture. Recording started after the participant confirmed readiness and a standardized verbal start cue was given by the experimenter, and it ended with a verbal stop cue. The event-based camera recorded only during the gesture trial, and each recording was saved individually in AEDAT4 format using the DV software platform. Between each recording, participants were given time to review the next gesture and decide when they were ready to proceed.
For the static conditions, participants remained seated in a stationary vehicle, and each gesture was recorded in a separate 15 seconds trial. For the driving conditions, participants actively drove the vehicle, and gestures were performed during the separate 15 seconds recordings.

4.2 Data preprocessing

The event stream in this work is stored in AEDAT4 format. Each event in the raw data is generated by the camera hardware and contains the pixel location, timestamp, and polarity. To create training samples, each 15 seconds gesture recording is segmented into shorter clips using a sliding-window approach. A window length of 500 milliseconds and a stride of 250 milliseconds are used, such that neighboring clips share 50% of their temporal duration. In the implementation, event timestamps are measured in microseconds, and we convert the window length and stride to 500,000 microseconds and 250,000 microseconds.
This overlap is used because the exact start and end of a gesture within a 15 seconds recording are not clearly defined. If there is no overlap, a gesture can be cut off at the clip boundaries. Let t_{first} and t_{last} denote the timestamps of the first and last event in a recording, and let $T = t_{\text{last}} - t_{\text{first}}$ denote the effective event duration. The number of extracted clips is given by

$$N = \left\lfloor \frac{T - W}{S} \right\rfloor + 1, \tag{1}$$

where W denotes the window length, S the stride, and N the number of clips. For an ideal 15 seconds recording, this setting produces a maximum of 59 clips, but the exact number may differ when fewer events occur.
After temporal segmentation, each 500 millisecond clip is converted into a voxel-grid representation of size (5, 260, 346), where the first dimension corresponds to the number of temporal bins and 260x346 corresponds to the camera resolution. The bins are used to capture the temporal progression of the gesture. For example, gestures such as a left swipe and a right swipe may appear similar if all events within the 500 milliseconds window are aggregated into a single frame. When the window is divided into multiple bins, the model can observe the temporal order of events, and this helps distinguish gestures that have similar spatial patterns, but different motion directions.

4.3 Model architecture

The classification model is a CSNN inspired by prior work on event-based spiking architectures [PTS+24]. The model combines convolutional layers for spatial feature extraction with spiking neurons that learn spatial and temporal relations in event-based inputs.
An overview of the network architecture is shown in Figure 4. The model takes a voxel-grid representation of the event stream, and the network is organized into four convolutional blocks; the convolutional layers extract spatial patterns related to hand shape, and the Leaky Integrate-and-Fire (LIF) neurons handle temporal information. Through the membrane potential, LIF neurons integrate evidence across consecutive time bins and gradually forget older inputs, which allows the network to capture short-term temporal structure. As the network goes deeper, the number of feature channels increases to represent more complex gesture patterns. At the same time, average pooling reduces the spatial resolution. After the final convolutional block, dropout is applied to reduce overfitting. The extracted features are then flattened and passed to a fully connected layer to produce gesture class predictions. All spiking layers use an arctangent surrogate gradient to enable backpropagation through the non-differentiable spike function. The model is trained using the Adam optimizer with an MSE spike-count loss, where higher firing rates are encouraged for the correct class and lower firing rates for incorrect classes.

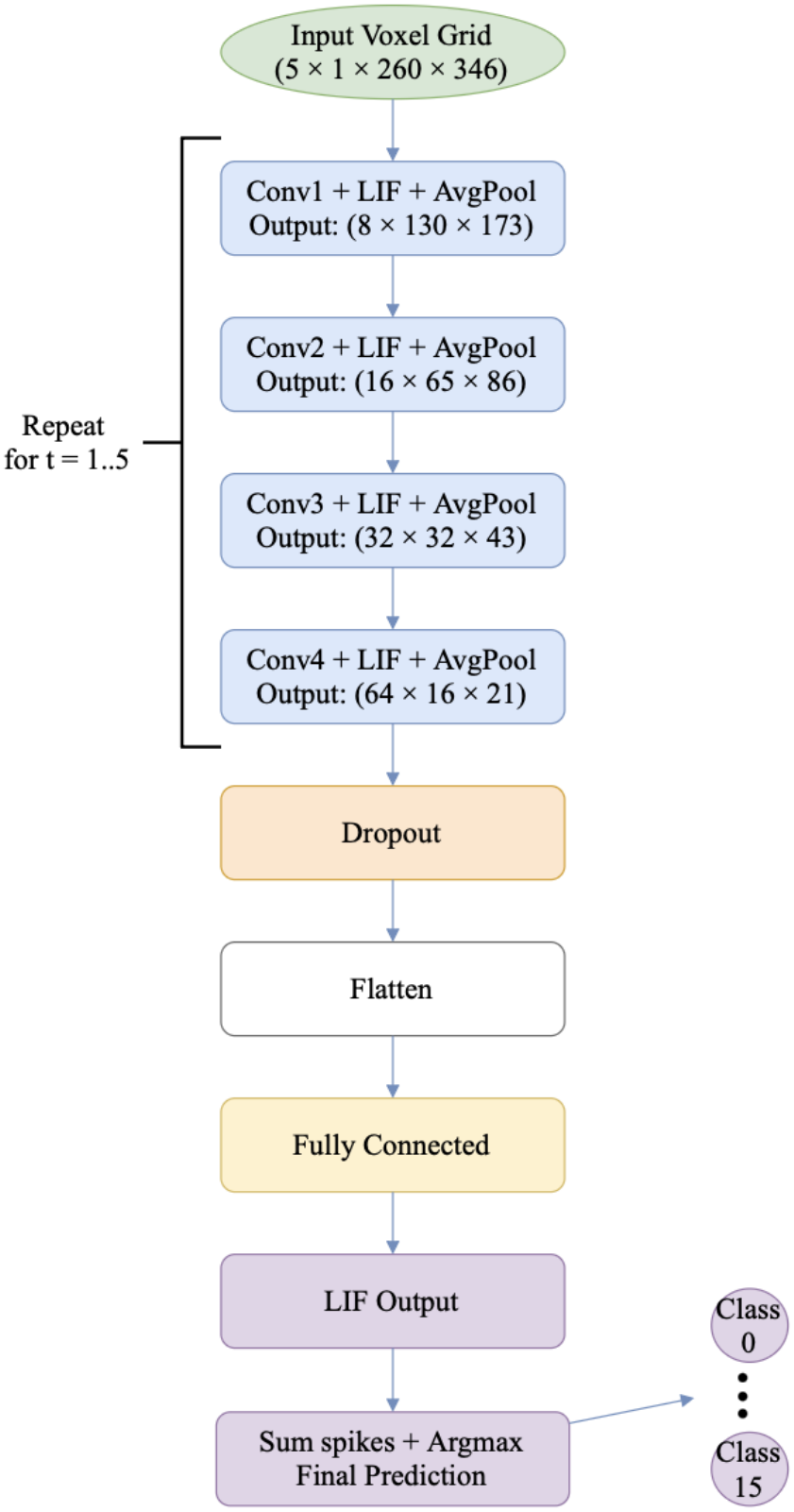


Figure 4: CSNN architecture for 16-class gesture recognition.

4.4 Model analysis

The primary evaluation metric is the classification accuracy, which is defined as the ratio of correctly classified samples to the total number of samples in the test set. For each input sample, the CSNN produces spiking outputs for each gesture class at each discrete time step. In the experiments, the network is evaluated over five time steps, and the output neurons may emit spikes at each step. To obtain a single prediction, the spikes of each class are aggregated across all time steps. The class with the highest total spike count is selected as the predicted gesture label.

5 Experiments & results

For all experiments except Experiment 1.2 (Leave-one-subject-out), the dataset is randomly shuffled and divided into three disjoint subsets. 70% of the samples are used for training, the subsequent 15% are used for validation, and the remaining 15% are reserved for testing. For all experiments, the random shuffling is performed with a fixed random seed (42) to ensure reproducibility.
In Experiment 1.2 (Leave-one-subject-out), the data are partitioned according to subject identity. All samples from one subject are held out exclusively for testing. Samples from the remaining subjects are used for model development. Within this training pool, 80% of the samples are used for training and the remaining 20% are used for validation.
In addition, some experiments involve unequal numbers of samples among different classes or recording conditions because of the limited data collection. Therefore, numerical values in the confusion matrices may differ across experiments.

5.1 Experiment 1.1: Mixed-condition baseline

This experiment evaluates the performance of the CSNN model when data from all recording conditions are combined into a single training and evaluation set. Under this mixed-condition setting, the model achieves an overall accuracy of 69.82%. The confusion matrix in Figure 5 shows a clear diagonal structure for many gesture classes, which indicates that correct predictions dominate for a substantial portion of the dataset. This result shows that the model remains reasonable when recordings from different illumination and motion conditions are combined.

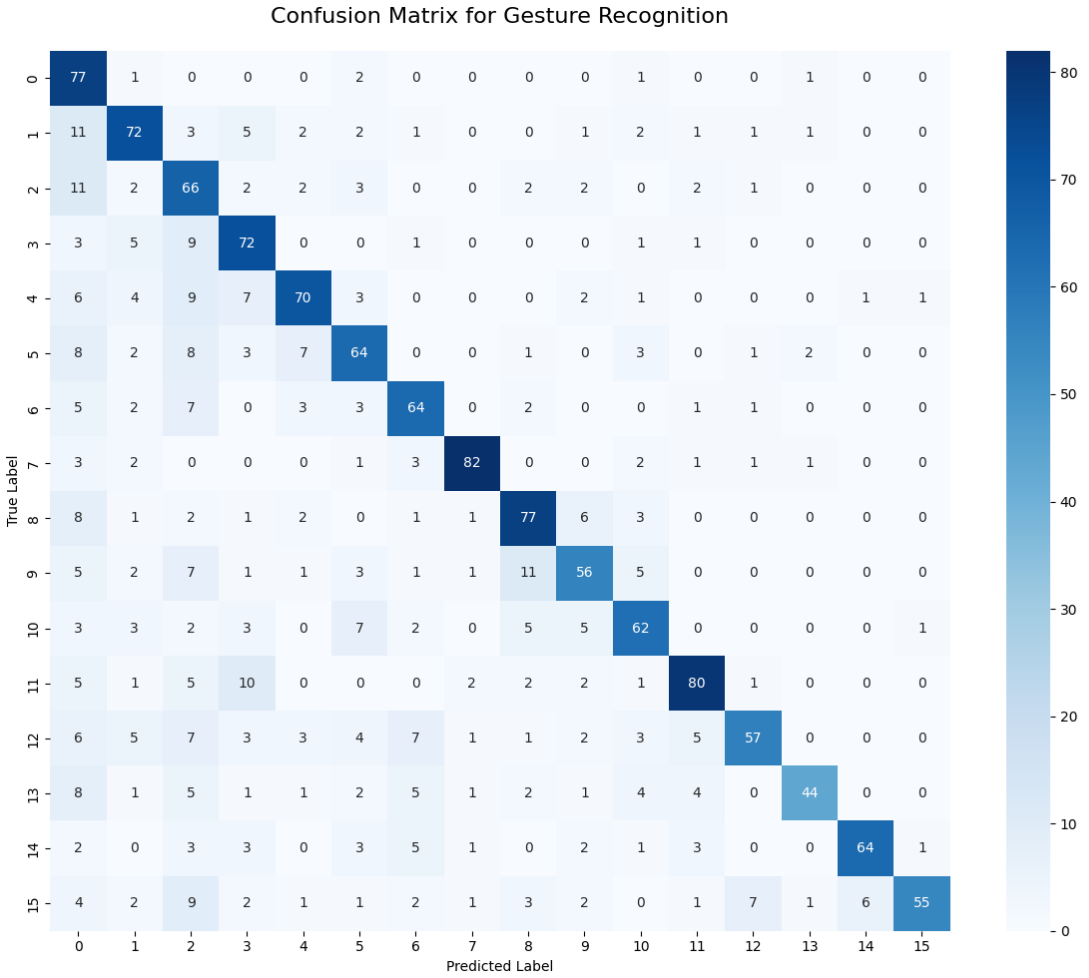


Figure 5: Confusion matrix of the CSNN model's classification results on the mixed dataset after epoch 25 (accuracy:69.82%).

Several gestures are recognized particularly well in this setting. For example, the thumb_down gesture (Class 7) and the back_and_forth gesture (Class 11) achieve 82 and 80 correct predictions, respectively. In contrast, the make_a_call gesture (Class 13) shows the lowest number of correct predictions, with only 44 correctly classified samples. The confusion matrix indicates frequent misclassification of this gesture as pinch (Class 0), swipe_right (Class 2), and thumb_up (Class 6). A plausible explanation is that the make_a_call gesture relies on subtle finger configurations and small-amplitude motions. As a result, this gesture produces less distinctive event patterns that overlap with those of other gestures, which makes classification more difficult under mixed recording conditions.

5.2 Experiment 1.2: Leave-one-subject-out

This experiment evaluates the ability of the CSNN model to generalize to unseen subjects under realistic in-vehicle conditions. Compared to the mixed-condition baseline, the recognition accuracy drops in the leave-one-subject-out setting, with test accuracies ranging from 15.16% to 26.15% across different subjects (Figures 6–10). This result highlights the difficulty of cross-subject generalization

in event-based in-vehicle gesture recognition.
Although all participants performed the same predefined gesture set, clear differences are observed. Differences in execution speed and recording conditions lead to subject-specific event representations. As a result, gesture patterns from a subset of subjects do not reliably transfer to an unseen subject, especially when illumination and vehicle-motion conditions differ.

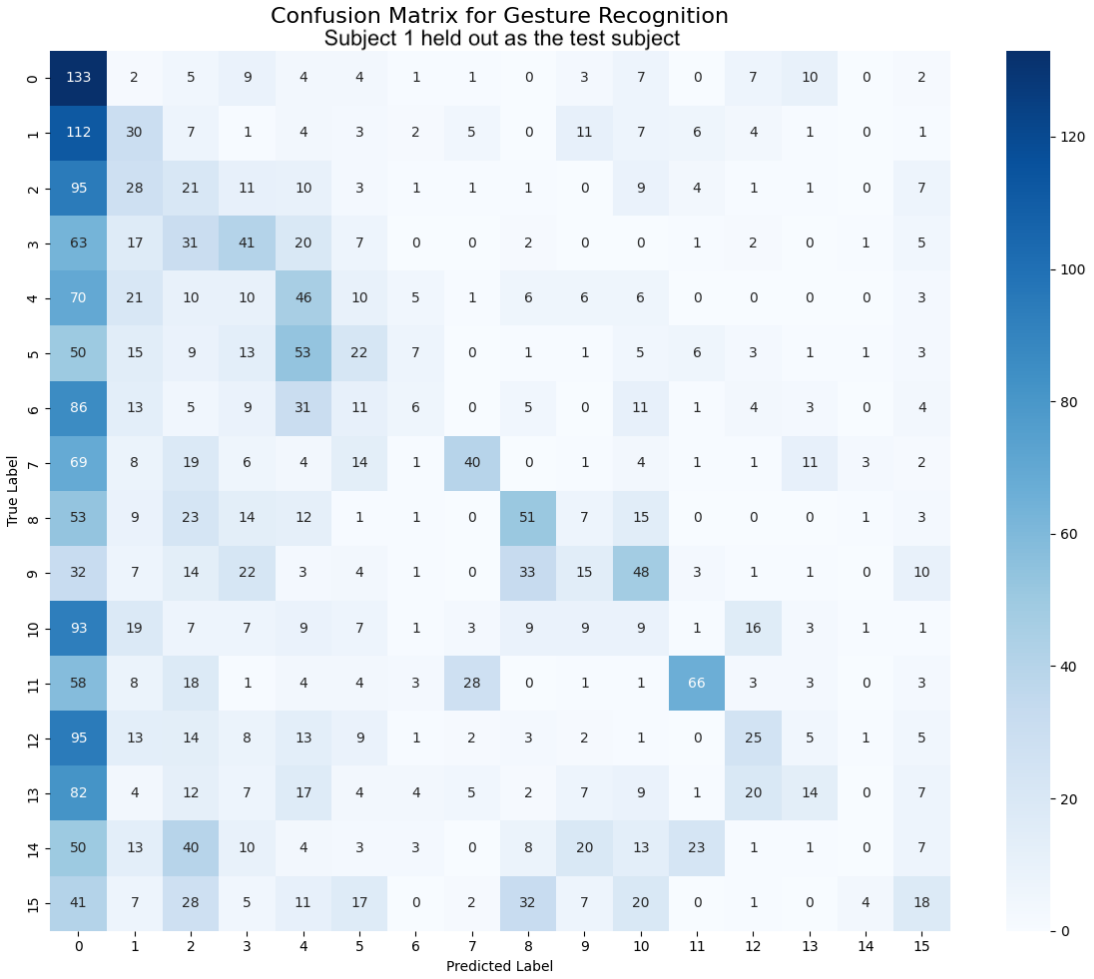


Figure 6: Confusion matrix for the leave-one-subject-out evaluation with Subject 1 held out for testing (accuracy: 17.41%).

Figure 6 shows the leave-one-subject-out results for Subject 1. The pinch gesture (Class 0) is recognized well, with 133 correctly classified samples. At the same time, several other gestures are frequently misclassified as pinch. For example, the OK gesture (Class 1) is predicted as pinch in 112 cases.

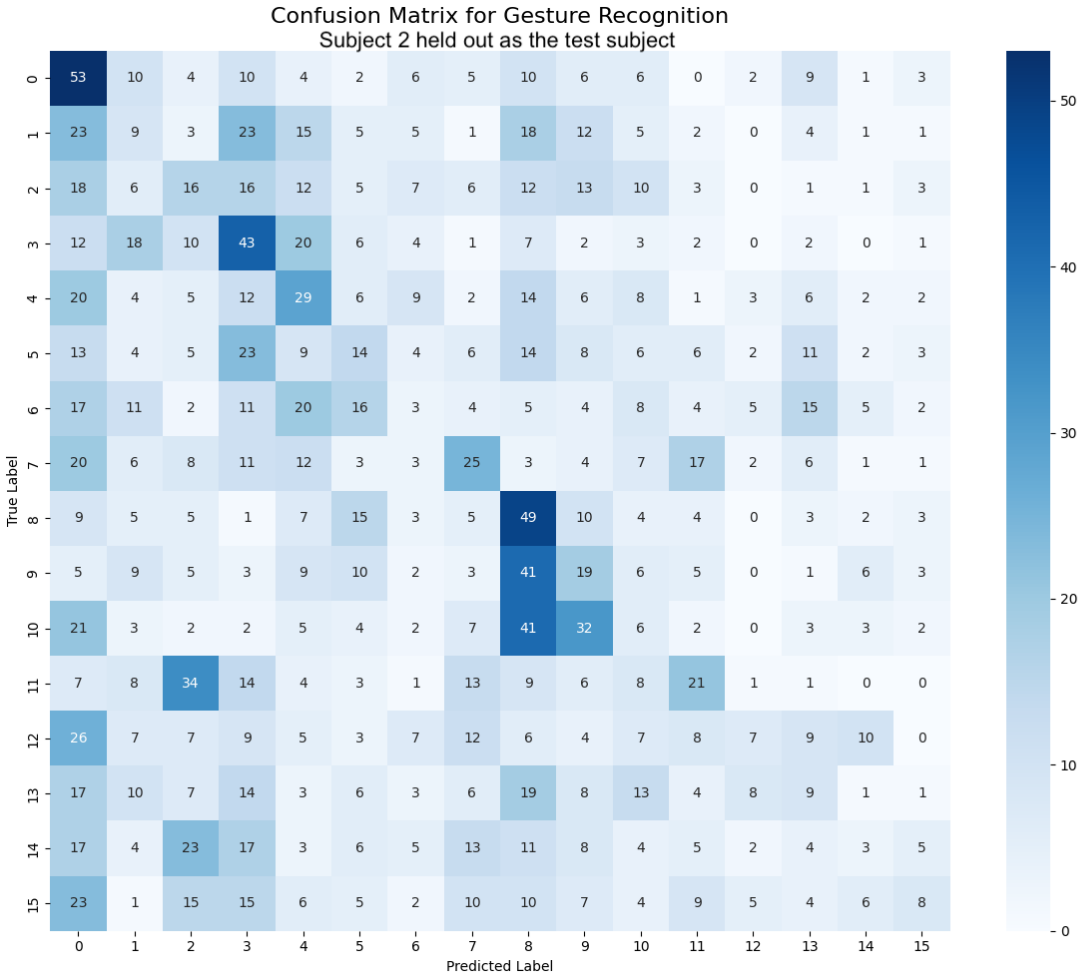


Figure 7: Confusion matrix for the leave-one-subject-out evaluation with Subject 2 held out for testing (accuracy: 15.16%).

Figure 7 shows the lowest accuracy within the five leave-one-subject-out experiments. In this case, the test data mainly come from nighttime recordings, including both static and driving settings. Although the pinch gesture (Class 0) and the number 1 gesture (Class 8) still achieve relatively higher numbers of correct predictions (53 and 49 samples, respectively), many other gestures are poorly recognized. For example, the thumb_up gesture (Class 6) and the rotate_clockwise gesture (Class 14) are each correctly classified only three times.

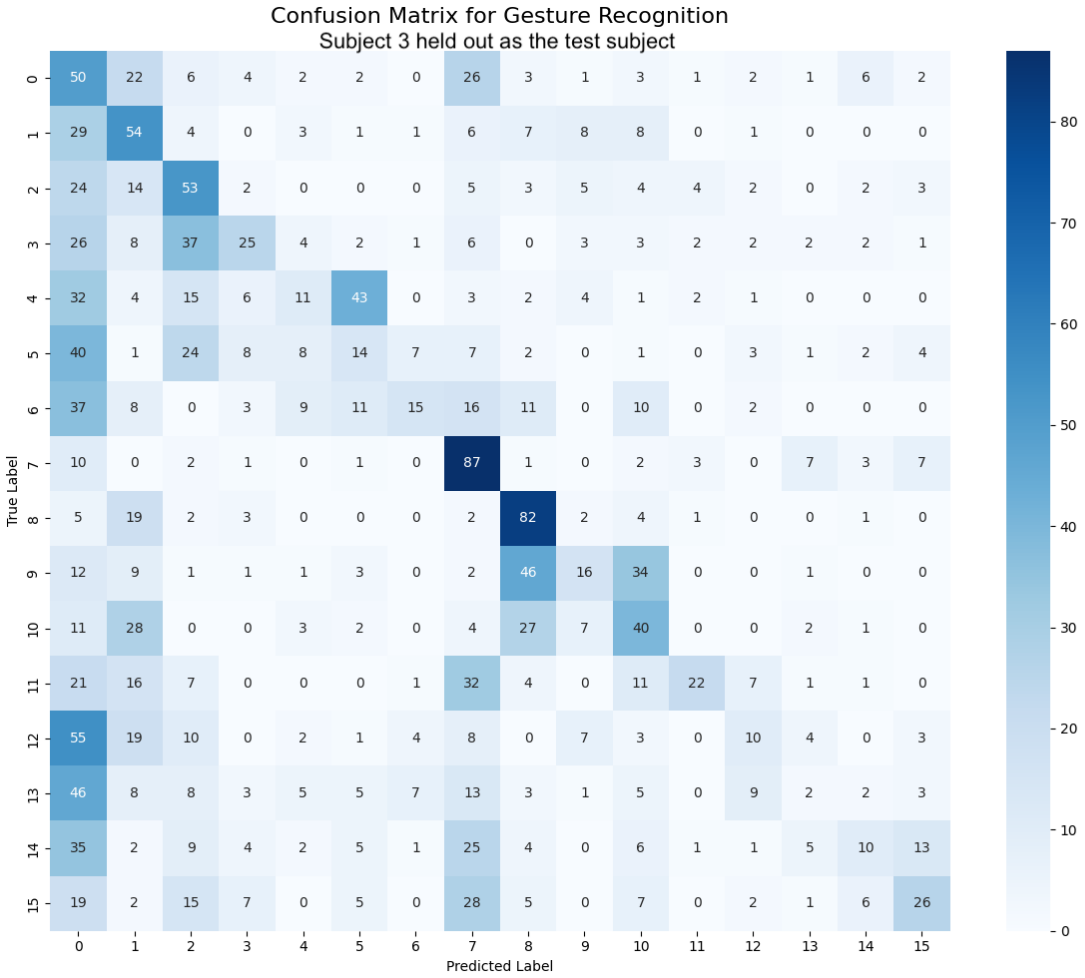


Figure 8: Leave-one-subject-out evaluation for Subject 3 (accuracy: 26.15%).

Both Subject 3 and Subject 4 were recorded under the same daylight static and nighttime static conditions. In the leave-one-subject-out evaluation, Subject 3 achieves the highest accuracy of 26.15%. Subject 4 achieves the third highest accuracy at 23.26%. Although these accuracies are close, the confusion matrices reveal different class-level behaviors.

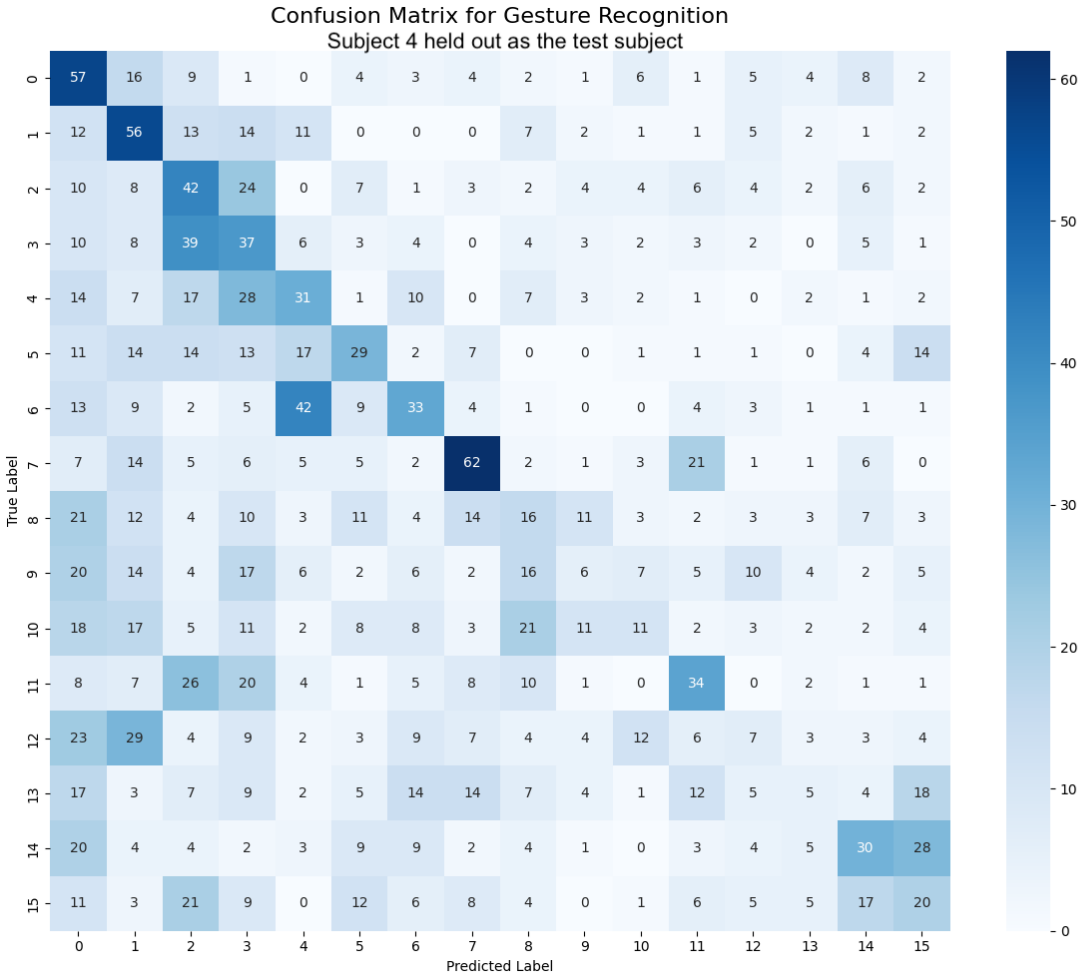


Figure 9: Leave-one-subject-out evaluation for Subject 4 (accuracy: 23.26%).

Figures 8 and 9 show several common patterns in the confusion matrices for Subject 3 and Subject 4. For both subjects, the thumb_down gesture (Class 7) is recognized well, with 87 and 62 correct predictions, respectively. The pinch gesture (Class 0) also shows similar performance, with 50 correct predictions for Subject 3 and 57 for Subject 4. The OK gesture (Class 1) follows a comparable trend, with 54 and 56 correctly classified samples. A notable difference appears for the number 1 gesture (Class 8). This gesture is correctly classified 82 times for Subject 3, but only 16 times for Subject 4. The number 1 gesture depends on a clear extension of the index finger, which varies across subjects and leads to different recognition outcomes.

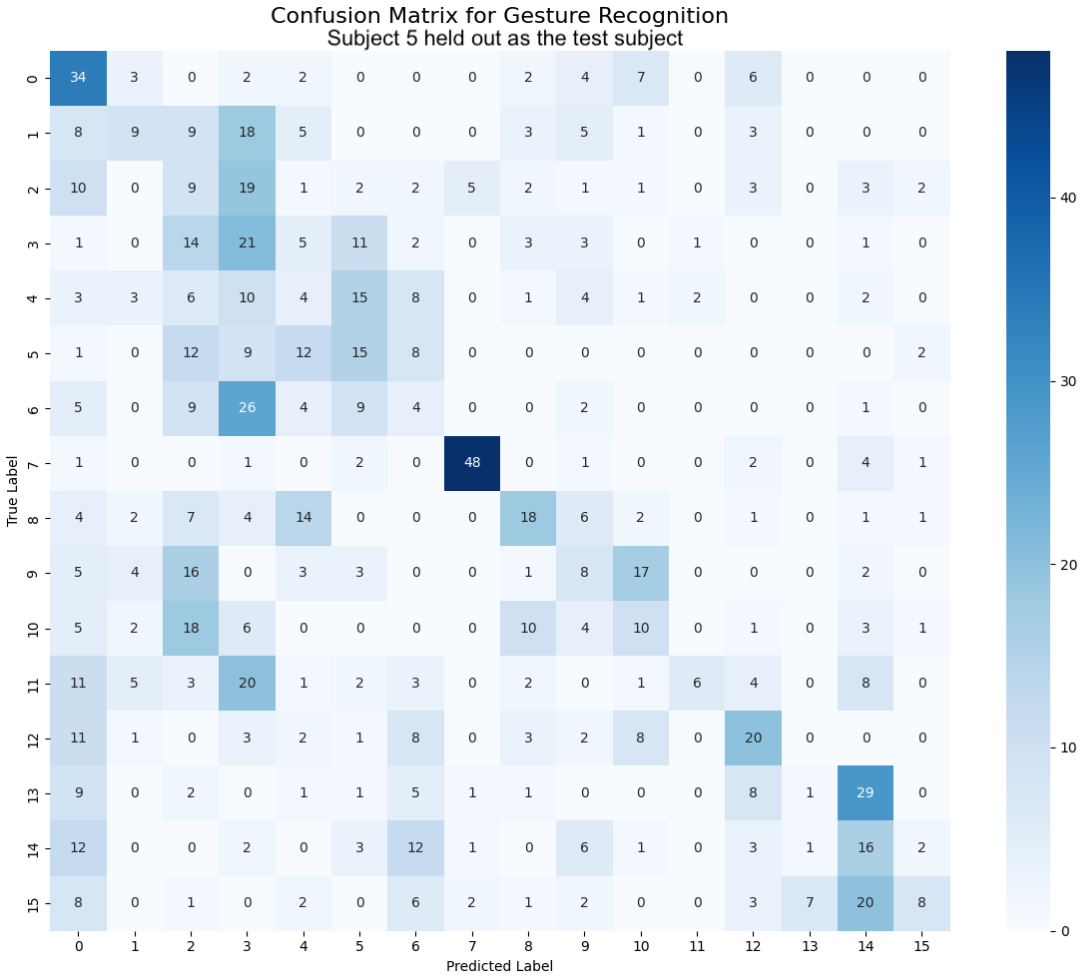


Figure 10: Leave-one-subject-out evaluation for Subject 5 (accuracy: 24.01%).

Figure 10 shows the confusion matrix for Subject 5. The recordings for this subject were collected under daylight-to-nighttime static conditions. The model achieves an accuracy of 24.01%, which is the second highest of the five leave-one-subject-out experiments.

The thumb_down gesture (Class 7) is recognized most reliably, with 48 correctly classified samples. In contrast, the make_a_call gesture (Class 13) shows the weakest performance, with only one correctly classified sample. Most make_a_call samples are misclassified as pinch (Class 0) or stop_grab (Class 12).

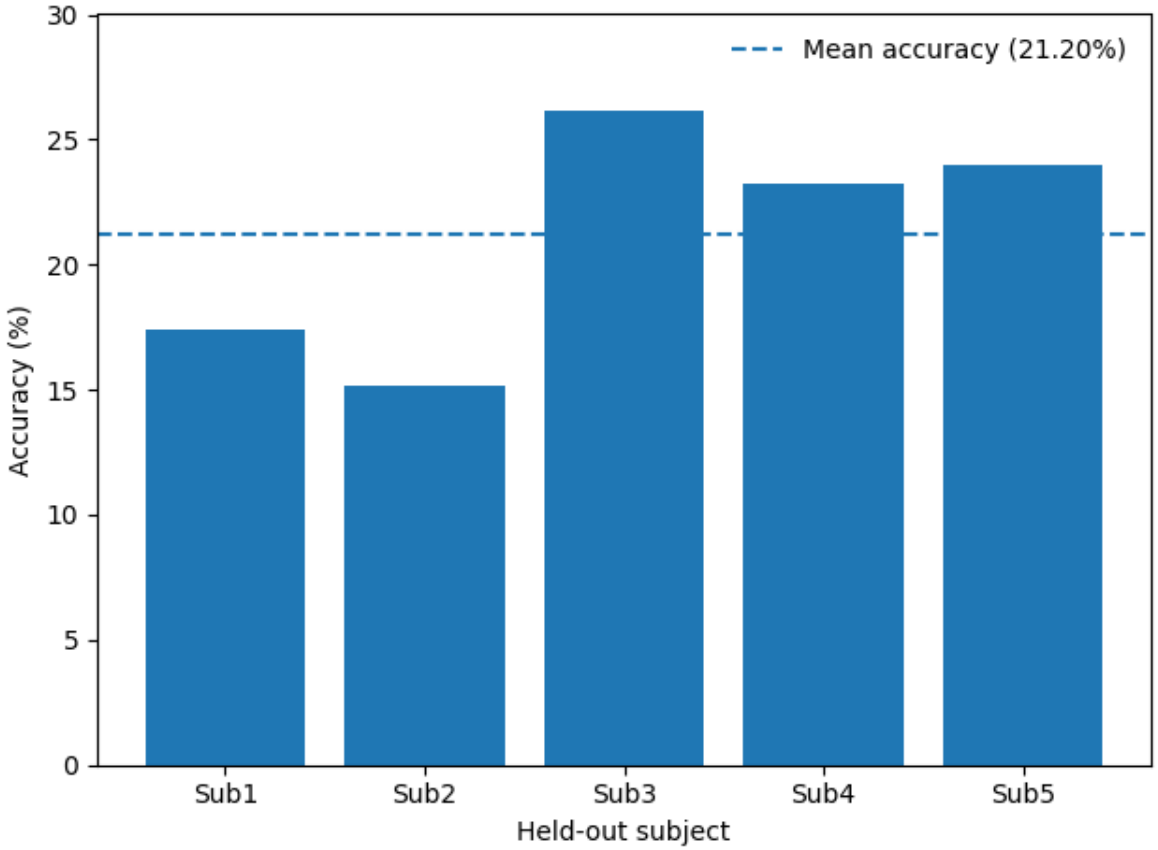


Figure 11: Test accuracy of the CSNN model for each leave-one-subject-out fold. One subject is used for testing in each fold, and the remaining subjects are used for training and validation. The dashed horizontal line indicates the mean accuracy.

Figure 11 summarizes the test accuracy across the five leave-one-subject-out experiments, with a mean accuracy of 21.20%. Subjects 3 (26.15%), 4 (23.26%), and 5 (24.01%) achieve accuracies above the mean. Subjects 1 (17.41%) and 2 (15.16%) perform below the mean.
The observed performance differences relate to the recording conditions of each subject. Subjects 3, 4, and 5 were recorded under daylight static and nighttime static conditions that are well represented in the training data. Subject 2 was the only participant who contributed nighttime driving recordings.

5.3 Experiments 2 and 3: daylight-only and nighttime-only

Experiments 2 and 3 evaluate the performance of the CSNN model when training, validation, and testing are restricted to daylight-only and nighttime-only recordings, respectively (Figures 12 and 13).

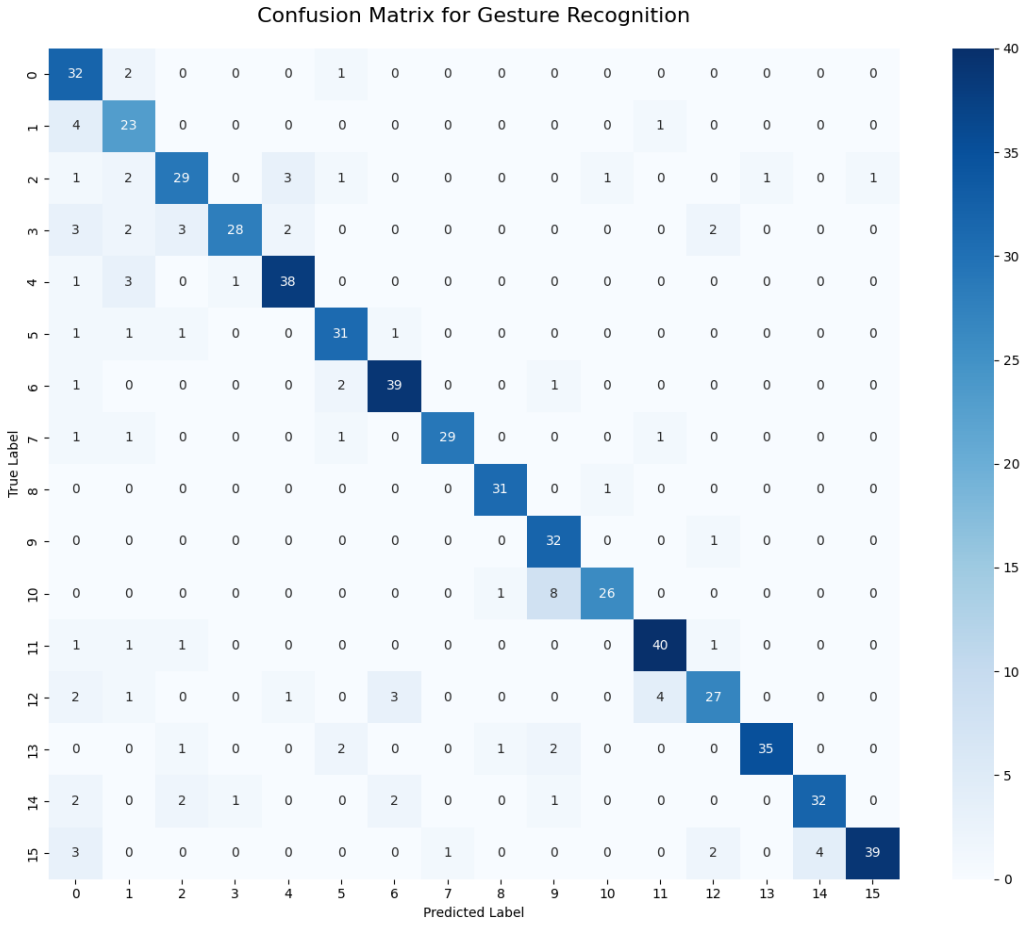


Figure 12: Daylight-only condition after 25 epochs (accuracy: 84.05%).

Under daylight-only conditions, the model achieves an accuracy of 84.05%. As shown in Figure 12, the overall recognition performance is higher than the mixed-condition baseline (69.82%), because it has a stronger diagonal structure in the confusion matrix. This suggests that under consistent illumination conditions, the model can more effectively distinguish between gesture classes. Misclassifications mainly occur between gestures with similar hand configurations or subtle motion differences. For instance, OK (Class 1) and swipe-left (Class 3) are sometimes predicted as pinch (Class 0), because these gestures require the index finger movements that resemble those of the pinch gesture. In addition, number 3 (Class 10) is often confused with number 2 (Class 9), which is likely caused by the similar finger extension patterns in the event data.

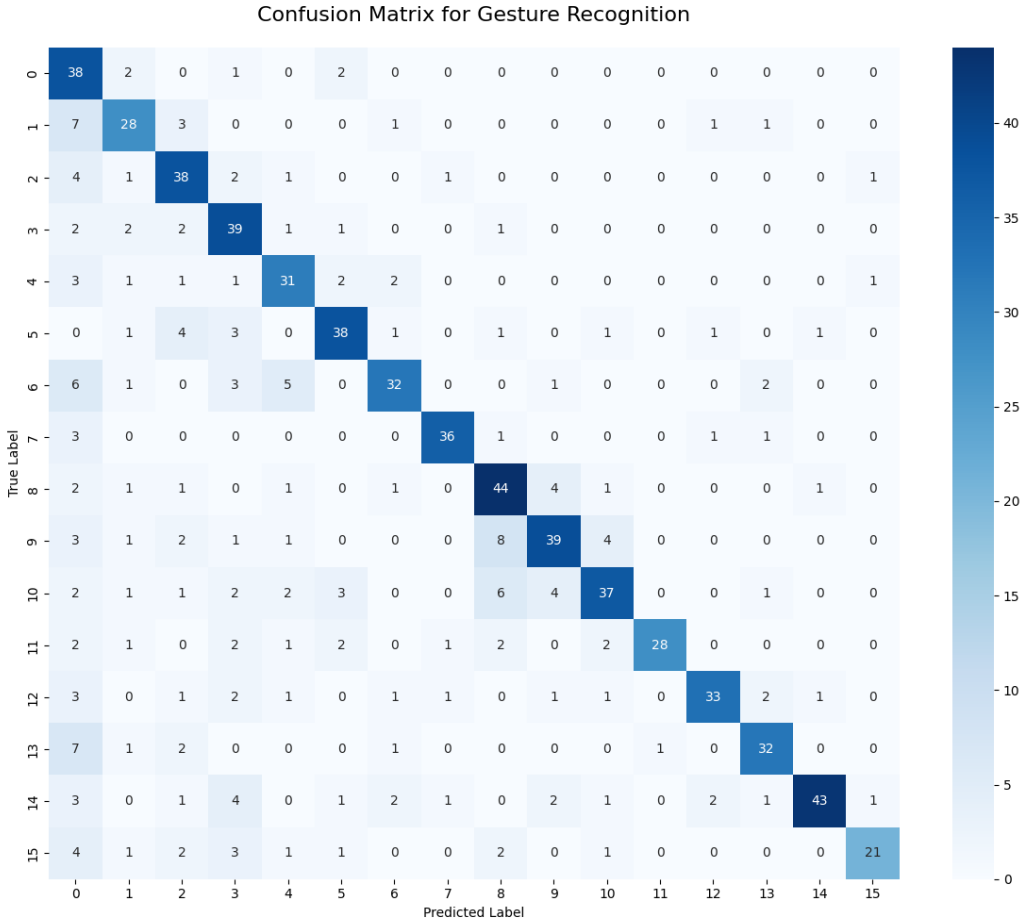


Figure 13: Nighttime-only condition after 25 epochs (accuracy: 72.43%).

Under nighttime-only conditions, the model achieves an accuracy of 72.43%. As shown in Figure 13, OK (Class 1), thumb_up (Class 6), and make_a_call (Class 13) are often misclassified as pinch, which suggests that under reduced illumination these gestures produce event patterns that are less distinctive and closer to some grasp-like motions. Rotate_counterclockwise (Class 15) was well recognized under daylight conditions but not under nighttime-only conditions.

5.4 Experiment 4: Daylight-to-nighttime generalization

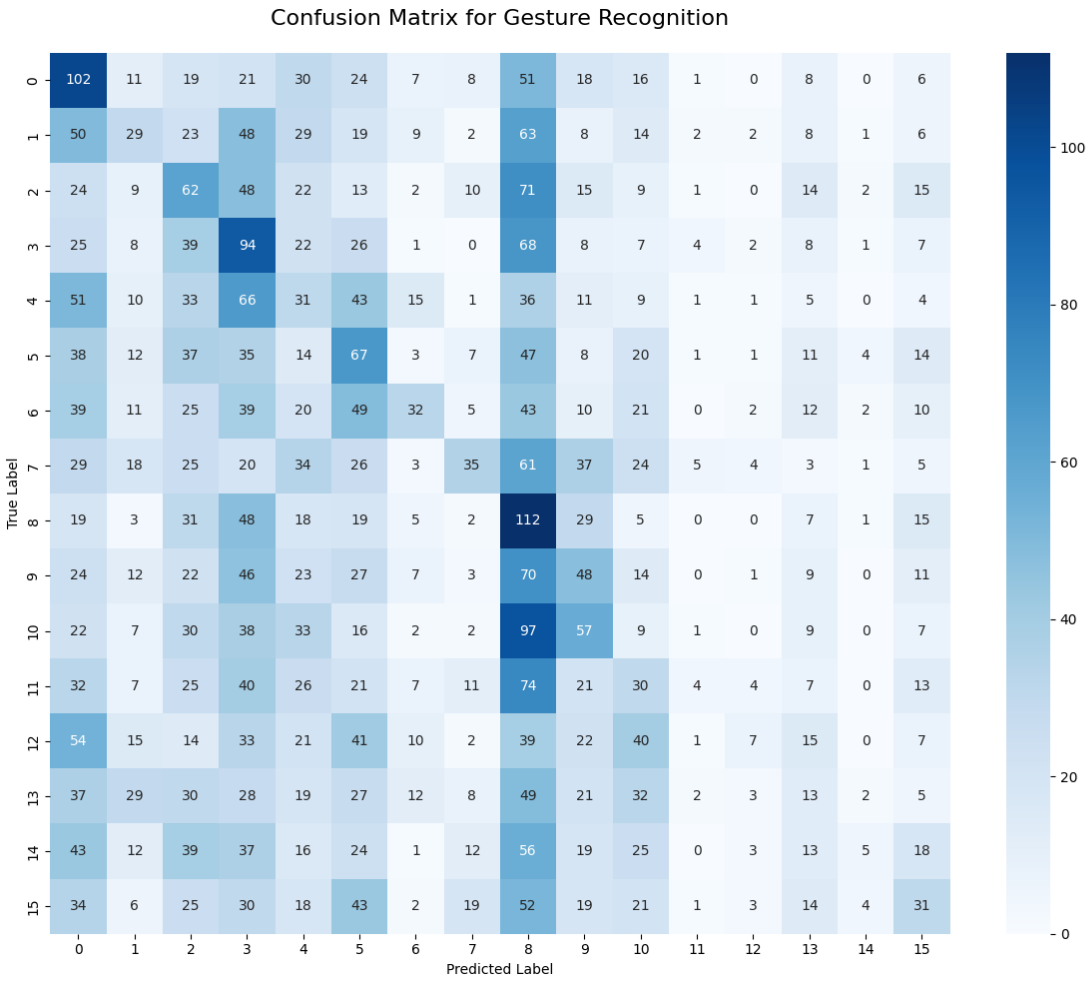


Figure 14: Confusion matrix of the CSNN model under daylight-to-nighttime generalization conditions (accuracy:13.29%).

Under day-to-night generalization conditions, the model achieves an accuracy of 13.29%. As shown in Figure 14, the confusion matrix does not show a clear diagonal structure, which means that the model which is trained on daylight recordings does not generalize well on nighttime data. In addition, predictions are frequently concentrated on the number 1 gesture (Class 8), with samples from multiple other gesture classes misclassified into this category.
Experiments 2 and 3 show stable performance when training and testing are restricted to the same illumination conditions, with higher accuracy under daylight than nighttime recordings.

5.5 Experiment 5 and 6: static-only and driving-only

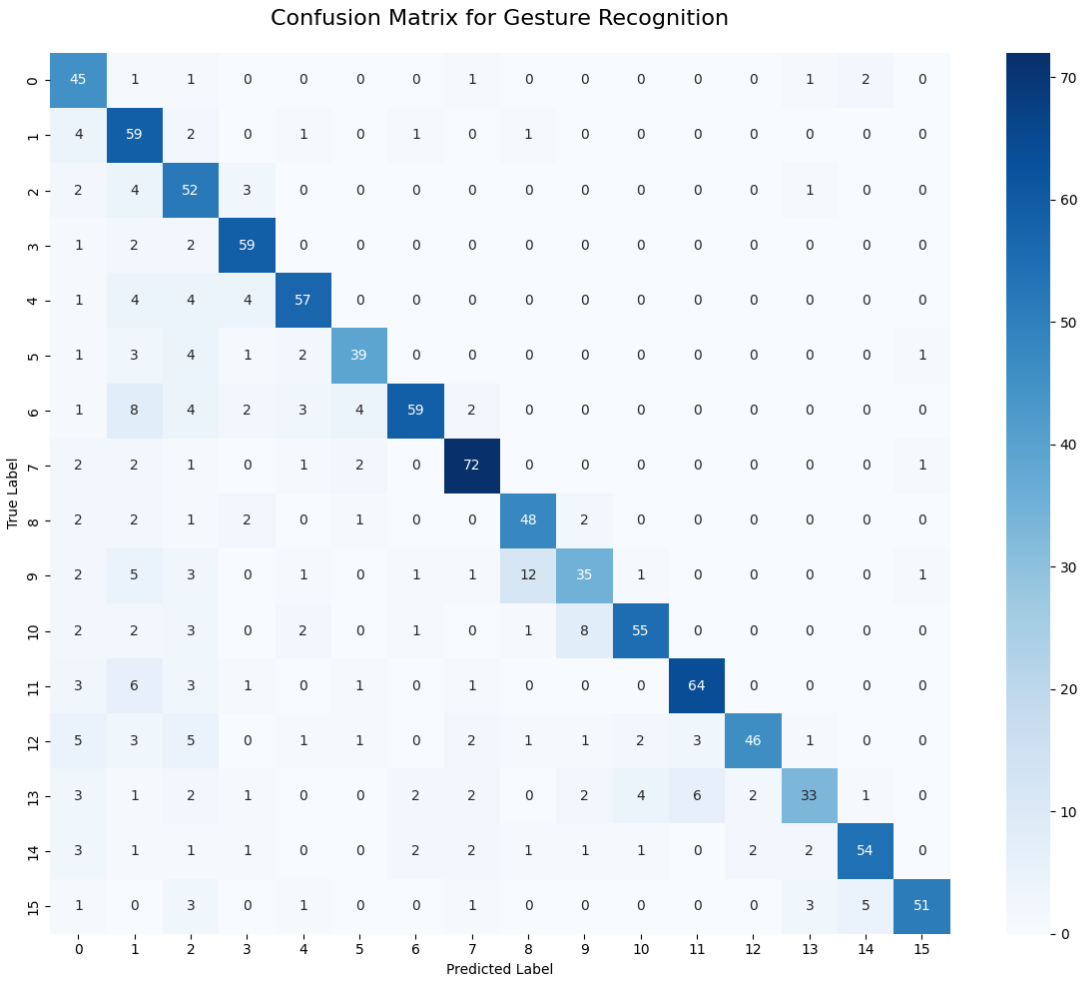


Figure 15: Static-only condition after 25 epochs (accuracy: 77.46%).

Under static-only conditions (Figure 15), the model achieves an accuracy of 77.46%. Overall performance is stable, with a clear diagonal structure in the confusion matrix. Most gestures are recognized reliably, especially thumb_down (Class 7), which reaches 72 correct predictions. However, some gestures are still challenging even in static conditions. In particular, number 2 (Class 9) and make_a_call (Class 13) show lower accuracy. For example, number 2 (Class 9) is frequently confused with OK (Class 1) and number 1 (Class 8).

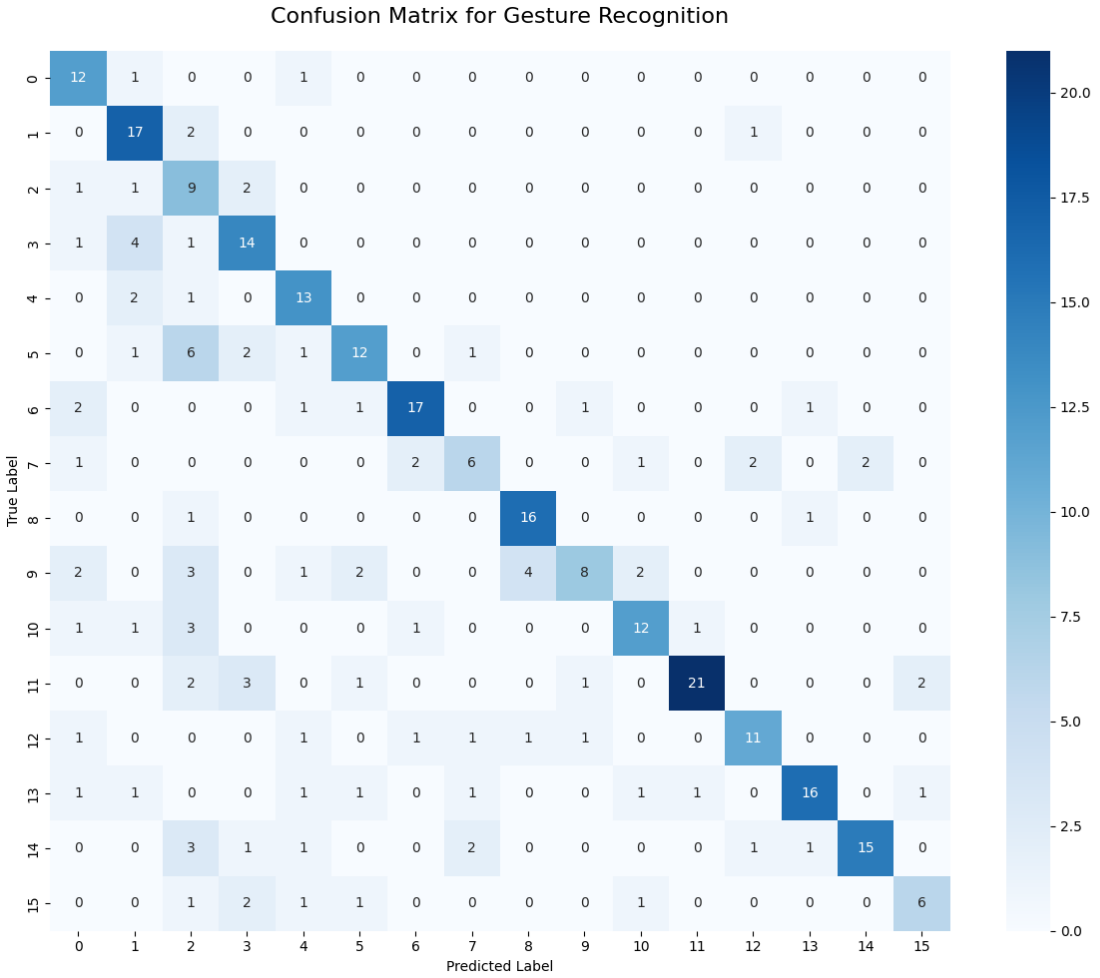


Figure 16: Driving-only condition after 25 epochs (accuracy: 66.34%).

Under driving-only conditions (Figure 16), the accuracy drops to 66.34%, which is 11.12 percentage points lower than in the static setting. It is important to note that the number of driving samples is smaller than the number of static samples, because only two subjects contributed to the driving recordings, whereas four subjects contributed to the static recordings, so the confusion matrix of the driving-only condition should not be directly interpreted as evidence that the driving gestures are more difficult than static gestures.

Despite the lower overall accuracy, some gestures are still recognized more consistently in the driving condition. From the confusion matrix, back_and_forth (Class 11) has the highest number of correct predictions.

5.6 Experiment 7: Static-to-driving generalization

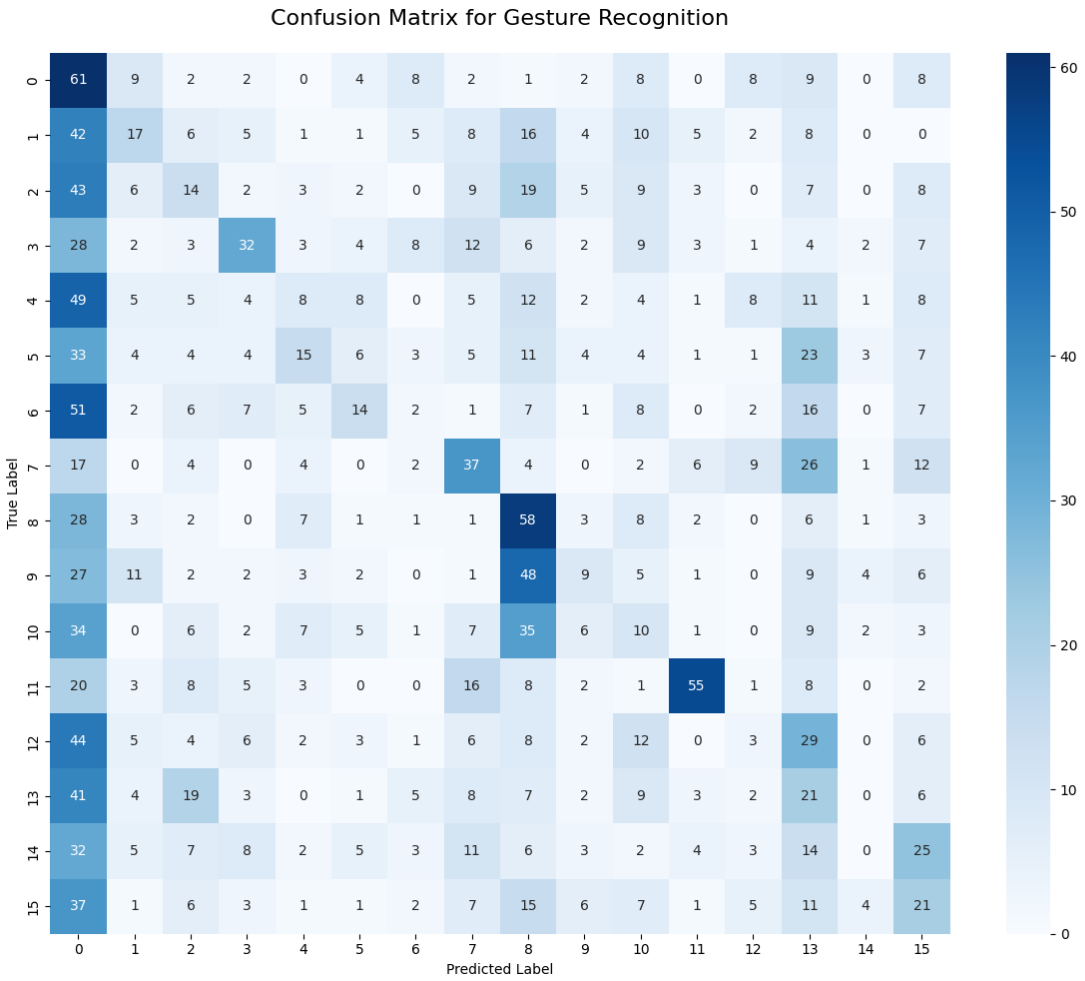


Figure 17: Confusion matrix of the CSNN model under static-to-driving generalization conditions (accuracy: 12.22%).

Under static-to-driving generalization conditions, the model is trained on static recordings and evaluated on driving recordings. The overall accuracy drops sharply compared to the static-only and driving-only experiments.
As shown in Figure 17, many gesture classes are confused with each other, this indicates that the gesture patterns which are learned from the static conditions do not transfer well to driving scenarios.

6 Discussion & Future Work

When training and testing data are from the same recording condition, the CSNN demonstrates relatively stable performance. Accuracies of 84% and 72% are achieved under daylight-only and nighttime-only conditions, respectively. Similarly, the model reaches 77% accuracy for static-only evaluation and 66% for driving-only evaluation. These results indicate that, under consistent illumination and motion settings, the model can learn discriminative event-based representations for in-vehicle hand gestures.

In contrast, recognition accuracy drops when the evaluation involves changes in illumination, vehicle motion, or subject, with daylight-to-nighttime accuracy of 13% and static-to-driving accuracy of 12%. Furthermore, cross-subject evaluation produces accuracies between 15% and 26% across held-out subjects. These results show that the model does not generalize well to different recording conditions and subjects with the current dataset. Therefore, this work could be improved in several aspects to further enhance performance.

6.1 Dataset limitations and collection considerations

The dataset is limited in scale and subject diversity. Only five participants were recorded, which is insufficient to capture the variability in motion amplitude and execution style that can be expected across different drivers. The low leave-one-subject-out performance suggests that the subject-specific patterns play a major role in the learned representations. With such limited subject coverage, it is difficult for the model to learn gesture features that generalize reliably to unseen users.

In addition, the coverage of recording conditions across participants is uneven, as shown in Table 3. Only two participants contributed to the driving recordings, and some conditions are only represented by a single subject. As a result, comparisons between static and driving data, or between daylight and nighttime data, reflect not only differences in recording conditions but also differences in which subjects contributed data.

6.2 Future directions for dataset design

Future data collection should aim for more balanced coverage of subjects and recording conditions. It would be ideal if each participant recorded all defined conditions, or at least a balanced subset. This would allow clearer isolation of illumination effects and driving-related effects without confounding subject identity.

The diversity of driving scenarios can be further expanded. The current dataset does not explicitly vary driving speed, road type, traffic density, or weather conditions. These factors can influence the visual dynamics of the scene and are likely to affect event density and background activity. A wider range of driving scenarios with basic metadata such as approximate speed ranges or road types, would support more detailed analysis in future work.

7 Conclusion

This thesis explored the use of event-based vision for in-vehicle hand gesture recognition. To address the lack of suitable data in this area, a pilot dataset was collected with a DAVIS346 event camera which was mounted inside a real vehicle. The dataset covers multiple realistic conditions, including daylight, nighttime, static, driving, and also transitions between these settings. A CSNN was implemented as a baseline to evaluate if event-based data can support gesture classification in this context.

The results show strong within-condition performance: when training and testing data come from the same recording environment, the model achieves high accuracy and the confusion matrices demonstrate clear class separation. This suggests that event-based vision can capture meaningful patterns for in-car gestures under consistent conditions. In contrast, cross-condition evaluation (day-to-night and static-to-driving) leads to an accuracy drop and a loss of diagonal structure in the confusion matrices. Rather than attributing this outcome solely to the baseline architecture, it is more likely caused by the limitations of the pilot dataset, because the dataset is imbalanced across conditions; nighttime and driving data are underrepresented and contributed by fewer participants, and this makes it difficult for the model to learn stable features.

Overall, this dataset should be considered as a pilot study rather than a complete or large-scale benchmark. The size of the dataset is limited, and only a small number of subjects were recorded under some of the defined conditions, and this means that the variability of gesture execution, driving behavior, and environmental factors are not fully represented in the current data. To solve this problem, a larger dataset would be needed that includes more participants and more recording sessions. In addition, more advanced neural networks can be explored to better understand how well event-based gesture recognition generalizes under different challenging in-vehicle conditions.

References

[ATB+17] Arnon Amir, Brian Taba, David Berg, Timothy Melano, Jeffrey McKinstry, Carmelo Di Nolfo, Tapan Nayak, Alexander Andreopoulos, Guillaume Garreau, Marcela Mendoza, Jeff Kusnitz, Michael Debole, Steve Esser, Tobi Delbruck, Myron Flickner, and Dharmendra Modha. A low power, fully event-based gesture recognition system. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7388–7397, 2017.

[Del16] Tobi Delbruck. Neuromorphic vision sensing and processing. In *2016 46th European Solid-State Device Research Conference (ESSDERC)*, pages 7–14, 2016.

[FER+14] Ahmed Farooq, Grigori Evreinov, Roope Raisamo, Erno Mäkinen, Tomi Nukarinen, and Atif Abdul Majeed. Developing novel multimodal interaction techniques for touchscreen in-vehicle infotainment systems. In *2014 International Conference on Open Source Systems Technologies*, pages 32–42, 2014.

[GAGS21] Mathias Gehrig, Willem Aarents, Daniel Gehrig, and Davide Scaramuzza. Dsec: A stereo event camera dataset for driving scenarios. *IEEE Robotics and Automation Letters*, 6(3):4947–4954, March 2021.

[GDO+22] Guillermo Gallego, Tobi Delbrück, Garrick Orchard, Chiara Bartolozzi, Brian Taba, Andrea Censi, Stefan Leutenegger, Andrew J. Davison, Jörg Conradt, Kostas Daniilidis, and Davide Scaramuzza. Event-based vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(1):154–180, 2022.

[GMY+16] Shalini Gupta, Pavlo Molchanov, Xiaodong Yang, Kihwan Kim, Stephen Tyree, and Jan Kautz. Towards selecting robust hand gestures for automotive interfaces. In *2016 IEEE Intelligent Vehicles Symposium (IV)*, pages 1350–1357, 2016.

[ini19] iniVation. Davis346 camera specifications. https://inivation.com/wp-content/uploads/2019/08/DAVIS346.pdf, August 2019. Accessed: 2025-09-19.

[ini25] iniVation. Aedat file formats — documentation. https://docs.inivation.com/software/software-advanced-usage/file-formats/index.html, 2025. Accessed: 2025-09-19.

[KLR+21] Okan Köpüklü, Thomas Ledwon, Yao Rong, Neslihan Kose, and Gerhard Rigoll. Drivermhg: A multi-modal dataset for dynamic recognition of driver micro hand gestures and a real-time recognition framework, 2021.

[Mou20] Jere Mourujärvi. Voice-controlled in-vehicle infotainment system, Apr 2020.

[OBT14] Eshed Ohn-Bar and Mohan Manubhai Trivedi. Hand gesture recognition in real time for automotive interfaces: A multimodal vision-based approach and evaluations. *IEEE Transactions on Intelligent Transportation Systems*, 15(6):2368–2377, 2014.

[Pic05] C.A. Pickering. The search for a safer driver interface: a review of gesture recognition human machine interface. *Computing Control Engineering Journal*, 16(1):34–40, February 2005.

[Pix] Pix2nvs. Github - pix2nvs/nvs2graph: Repo for "graph-based object classification for neuromorphic vision sensing" iccv2019.

[PLGAAC14] Francisco Parada-Loira, Elisardo González-Agulla, and José L. Alba-Castro. Hand gestures to control infotainment equipment in cars. In *2014 IEEE Intelligent Vehicles Symposium Proceedings*, pages 1–6, 2014.

[PTS+24] Ria Patel, Sujit Tripathy, Zachary Sublett, Seoyoung An, and Riya Patel. Using csnns to perform event-based data processing classification on asl-dvs. *arXiv (Cornell University)*, August 2024.

[RGF23] Guillermo Reyes, Amr Gomaa, and Michael Feld. It's all about you: Personalized in-vehicle gesture recognition with a time-of-flight camera. In *Proceedings of the 15th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, AutomotiveUI '23, page 234–243. ACM, September 2023.

[TCF+24] Moustafa Tabbarah, Yusheng Cao, Ziming Fang, Lingyu Li, and Myounghoon Jeon. Sonically-enhanced in-vehicle air gesture interactions: evaluation of different spearcon compression rates. *Journal on Multimodal User Interfaces*, 18(2–3):159–177, May 2024.

[ZTO+18] Alex Zihao Zhu, Dinesh Thakur, Tolga Ozaslan, Bernd Pfrommer, Vijay Kumar, and Kostas Daniilidis. The multivehicle stereo event camera dataset: an event camera dataset for 3d perception. *IEEE Robotics and Automation Letters*, 3(3):2032–2039, February 2018.

A Complete list of the gesture variations

ID	Gesture	Gesture execution description
0	pinch	Thumb and index finger pinch together
1	OK_1	Hand stays in a stable "OK" shape; thumb and index finger form a circle
2	OK_2	Hand enters the camera view first, then forms the "OK" gesture
3	swipe_right_1	Only the index finger swipes to the right
4	swipe_right_2	Five fingers swipe right; the hand returns as a fist
5	swipe_right_3	The hand starts as a fist, then opens while swiping to the right
6	swipe_left_1	Only the index finger swipes to the left
7	swipe_left_2	Five fingers swipe left; the hand returns as a fist
8	swipe_left_3	The hand starts as a fist, then opens while swiping to the left
9	swipe_up	The index finger swipes upward
10	swipe_down	The index finger swipes downward
11	thumb_up	Thumbs-up gesture
12	thumb_down_1	The thumb points downward while the hand moves slightly
13	thumb_down_2	The hand begins as a fist, then extends the thumb downward
14	number1	The index finger is extended
15	number2	The index and middle fingers are extended
16	number3	Three fingers are extended
17	number4	Four fingers are extended
18	number5	All five fingers are extended
19	back_and_forth	Five fingers move left and right repeatedly
20	stop_grab	The hand performs a grabbing motion
21	make_a_call_1	The thumb and pinky are extended ("call" gesture); the hand moves slightly
22	rotate_clockwise	The index finger rotates clockwise
23	rotate_counterclockwise	The index finger rotates counterclockwise

Table 4: The complete list of the 24 gesture variations and descriptions of how each gesture is performed.

A List of final hand gestures and their functions

ID	Gesture	Function
0	pinch	zoom
1	OK	confirm
2	swipe_right	go to next page
3	swipe_left	go to previous page
4	swipe_up	increase volume
5	swipe_down	decrease volume
6	thumb_up	like
7	thumb_down	dislike
8	number1	numeric gesture
9	number2	numeric gesture
10	number3	numeric gesture
11	back_and_forth	activate voice assistant
12	stop_grab	stop
13	make_a_call	make a call
14	rotate_clockwise	increase temperature
15	rotate_counterclockwise	decrease temperature

Table 5: List of hand gestures and their corresponding functions.