International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 02 Issue: 04 April 2024
Page No: 889 - 894
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2024.0125

Human Activity Recognition Using CNN-LSTM-GRU Model

Garima Pandey[1], Abhishek Kumar Karn[2], Manish Jha[3]
[1]Assistant Professor, Galgotias University, Greater Noida, Uttar Pradesh, India.
[2, 3] UG Student, Galgotias University, Greater Noida, Uttar Pradesh, India., India.
Emails: *Hagarima.pandey@galgotiasuniversity.edu.in[1], kumarabhishekkarn096@gmail.com[2], jham2476@gmail.com[3]*

Abstract
Human Activity Recognition (HAR) is a fundamental task in the field of computer vision and machine learning, with applications spanning from healthcare monitoring to human- computer interaction. This research paper presents a novel approach to HAR utilizing a hybrid model combining Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks, referred to as the VGG-LSTM model. The proposed VGG-LSTM model leverages the power of deep learning to address the challenges associated with HAR, including capturing spatial features and modeling temporal dependencies in complex human activities. In this research, we employ the VGG architecture as the feature extractor to capture discriminative spatial information from input sensor data, such as images or videos. Furthermore, the LSTM layer is integrated to model the temporal dynamics of human activities. This enables the model to effectively recognize and differentiate between a wide range of human activities, such as walking, running, sitting, and more, in real-world scenarios. The research demonstrates the effectiveness of the VGG-LSTM model on benchmark datasets, achieving state-of-the-art performance in human activity recognition tasks. The model's accuracy, robustness, and ability to generalize to diverse scenarios make it a promising solution for applications in healthcare, sports analytics, security, and beyond. The contributions of this paper lie in the development of a powerful hybrid model that combines spatial and temporal information seamlessly, improving the accuracy and applicability of HAR systems. The results underscore the potential of the VGG-LSTM model in advancing human activity recognition technology, with implications for improving the quality of life and safety in various domains.

Keywords: *Human Activity Recognition, Deep Learning, Convolutional Neural Networks, LSTM, VGG, Spatial Features, Temporal Dependencies.*

1. Introduction
Human Activity Recognition (HAR) based on Inertial Measurement Unit (IMU) data has gained significant prominence due to its ability to monitor not only human activities but also the behavior of devices, machine components, and even pets. This approach ensures high levels of privacy and user comfort. While various methods for accurately classifying user activities using IMU sensor data have been proposed, many of them pose significant challenges, demanding substantial resources, domain expertise, and other barriers. The advent of deep learning has revolutionized HAR, rendering the task of activity recognition far more accessible. Deep learning has emerged as a dominant force in machine learning, significantly impacting various fields. However, it has not received as much attention in the context of HAR. In a typical HAR scenario, a user equipped with a device, such as a standalone sensor, smartwatch, or smartphone, containing gyroscopes and accelerometers, continuously transmits sensor data to a monitoring server. Modern smart devices, with their enhanced processing units, larger memories, and superior sensors, can perform activity recognition independently. Deep learning has simplified the training of models to recognize specific activities from raw sensor data swiftly and efficiently. This is in stark contrast to traditional machine learning algorithms like Support Vector Machines (SVM) and feature extraction techniques like the Histogram of Gradients (HOG) used in the

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 02 Issue: 04 April 2024
Page No: 889 - 894
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2024.0125

past, which required extensive data preparation, domain knowledge, and preprocessing. Deep learning models, such as Convolutional Neural Networks (CNNs) for spatial feature extraction, and Long Short- Term Memory (LSTM) networks for temporal modeling, have been proposed individually. Each approach possesses its unique strengths and weaknesses, tailored for specific applications. However, there is a compelling opportunity to enhance the robustness of activity recognition by combining the strengths of both networks. This research paper introduces a novel CNN-LSTM classifier for human activity recognition. While CNNs and LSTMs have been extensively explored in isolation, our study aims to harness the synergies between these two architectures, particularly in the context of human activity recognition. In the following sections, we provide a comprehensive background, review prior research, and detail our methodology and its implementation. We evaluate the performance of the CNN- LSTM model against other models using both the Intelligent Signal Processing Lab (iSPL) dataset and the UCI Human Activity Recognition (HAR) dataset. Finally, we conclude our research in Section V.

2. Background and Related Works

Human Activity Recognition (HAR) is a technology that characterizes human actions, allowing computer systems to anticipate and cater to users' needs. It categorizes various human activities such as walking, standing, working, or lying down as sequences of actions performed by individuals over a certain period.

2.1. Related Works

In the realm of Human Activity Recognition (HAR), re- searchers have delved extensively into various aspects, encompassing sensor technology and algorithmic approaches. Notably, the forefront of HAR research has witnessed a shift towards the adoption of artificial intelligence-driven recognition methods, made possible through the application of machine and deep learning techniques. This transition typically follows the extraction of relevant human motion features from the sensor data. HAR typically relies on data sources such as wearable devices, cameras, and millimeter-wave radar. One notable contribution comes from Ronald Mutegeki et al. [1], who introduced a CNN-LSTM approach achieving impressive accuracy levels, with 99% accuracy on the internal iSPL dataset and 92% accuracy on the publicly available UCI HAR dataset, surpassing other deep neural network architectures and traditional machine learning models that depend on manually crafted features. NA Choudhary [2] presents an efficient CNN-LSTM model for recognizing human activities, particularly in uncontrolled environments. Their approach employs adaptive batch sizes during training and validation and attains remarkable results with minimal data preprocessing and augmentation. The model reaches a peak accuracy of 99.29% with an average loss of $0.08 \pm 0.136\%$, and it validates successfully on two public datasets, health and Motion Sense, with impressive accuracies of 99.5% and 99.8%, respectively, highlighting the model's robustness and high- performance capabilities. Sakorn Mekruksavanich [3] introduces a hybrid model known as a multichannel CNN-LSTM network. This model's performance is assessed using key evaluation metrics such as accuracy, precision, recall, and F1-score on the publicly available DHA dataset, which contains accelerometer data from smartwatches. Impressively, the multichannel CNN-LSTM model outperforms other deep learning approaches, achieving an accuracy of 96.87. Additionally, researchers like NA Choudhary [4] and Shibo Zhang [5] propose an efficient CNN-LSTM model for recognizing daily human activities using smartphone sensor data. They create a contemporary CNN-LSTM model that efficiently handles hierarchical features through time-distributed feature extraction layers and LSTM memorization schemes. MST. Alema Khatun [6] employs a CNN-LSTM approach with different datasets, including MHEALTH and UCI-HAR, demonstrating the model's comparative performance. The pro- posed model achieves high accuracy, reaching 99.93% with H-Activity data, 98.76% with MHEALTH data, and 93.11% with UCI-HAR data, showcasing its effectiveness in human activity recognition. Furthermore, Zhu [7] introduces a deep learning (DL) model that combines 1-D convolutional neural

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 02 Issue: 04 April 2024
Page No: 889 - 894
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2024.0125

networks (1D-CNNs) with long short-term memory (LSTM). Experimental findings indicate that this model effectively captures spatiotemporal patterns in radar data, resulting in superior recognition accuracy while maintaining a relatively lower level of complexity compared to existing 2D-CNN approaches. Sakorna [8], through experimentation with the DHA dataset, finds that their proposed multichannel CNN-LSTM model surpasses other deep learning techniques, achieving a notably high accuracy score of 96.87. Tan [9] introduces the Enhanced Learning Architecture (ELA), which combines a Gated Recurrent Unit (GRU), a Convolutional Neural Network (CNN) stacked on the GRU, and a Deep Neural Network (DNN). The inclusion of an additional feature vector containing 561 time-domain and frequency-domain parameters contributes to the models performance. The DNN component acts as a fully connected layer, integrating the outputs of the three models for activity classification. S. Challa [10] introduces a multi-branch CNN-BiLSTM network for automatic feature extraction from raw sensor data, requiring minimal preprocessing. This model combines CNNs and BiLSTMs to capture both local features and long-term dependencies in sequential data, demonstrating high accuracy on benchmark datasets. Suneth Ranasinghe [11] provides an insightful overview of HAR applications, discussing their strengths and weaknesses. The article also highlights publicly available datasets tailored for assessing these recognition systems. Furthermore, it con- cludes by drawing a comparison between the current methodologies, offering insights into potential research questions for future advancements in real-world scenarios. Lastly, Charikleia Chatzaki [11] focuses on developing an efficient computational and analysis pipeline for precise recognition of Activities of Daily Living (ADLs) and falls. This effort results in two optimized feature sets (OFS1 and OFS2) achieved through iterative testing, removing less effective features, and showcasing promising results for ADL and fall recognition.

3. Method and Experiment Setup

The UCI HAR dataset is a collection of 3D (x, y, z) raw signals obtained from the accelerometer and gyroscope of a smartphone positioned at the waist of 30 subjects aged between 19 and 48 years [6]. The participants engaged in six activities: Walking, Walking Upstairs, Walking Downstairs, Sitting, Standing, and Laying. The dataset consists of 7,352 training samples and 2,947 test samples. Figure 1& 2 shows the Data Split for Train.

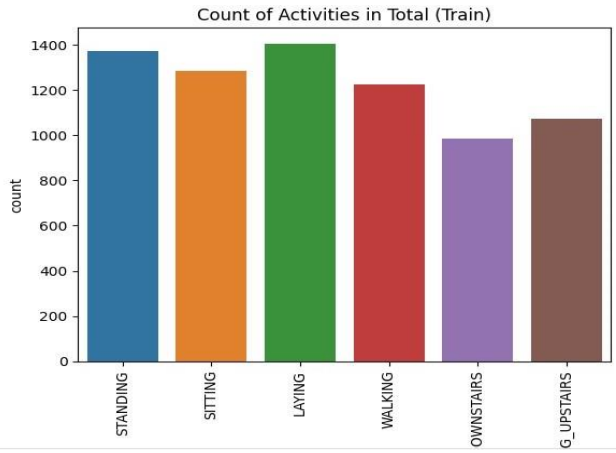

Figure 1 Data Split for Train

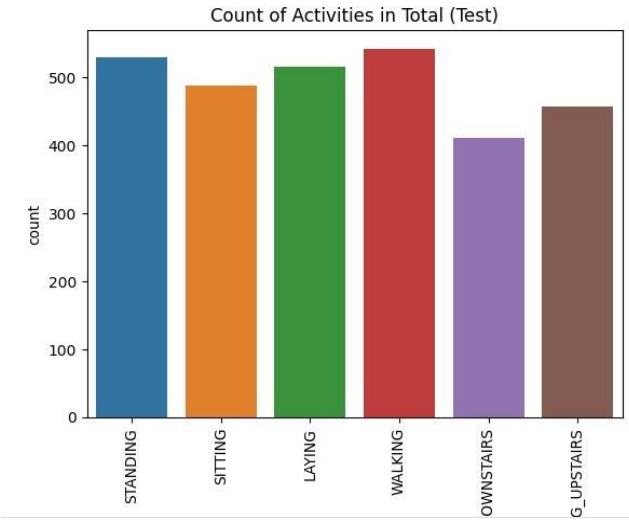

Figure 2 Data Split for Train

In the preprocessing phase, noise filters were applied to the sensor signals. The data was then segmented into fixed-width sliding windows lasting 2.56 seconds with a 50% overlap (128 readings per window). To isolate the gravitational and body motion components within the sensor acceleration signal, a Butterworth low-pass filter was employed. This separation resulted in distinct signals for body acceleration and gravity. In processing nine signals—

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 02 Issue: 04 April 2024
Page No: 889 - 894
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2024.0125

total acceleration (ax, ay, az), angular velocity from the gyroscope (gx, gy, gz), and linear acceleration excluding gravitational effects (lax, lay, laz)—we conducted a controlled experiment. This involved employing a singular 1D convolution layer with 64 filters, a kernel size of 3, and ReLU activation. In pursuit of a controlled and well-defined experimental framework, we initiated our study with a foundational 1D convolution layer employing Tensor Flow's Keras library. This layer featured 64 filters, a kernel size of 3, and a Rectified Linear Unit (ReLU) activation function. The subsequent addition of a 1D max-pooling layer, with a pool size set to 2, and a flatten layer aimed to format the extracted features for seamless integration into the subsequent Long Short-Term Memory (LSTM) layer. Figure 3 shows Model Architecture. Following the LSTM layer, a Gated Recurrent Unit (GRU) layer was seamlessly integrated to further capture and understand intricate patterns within the temporal sequences. The GRU layer, with its gating mechanism, enhanced the model's ability to retain important information over extended sequences. The combined output from the LSTM and GRU layers flowed into a fully connected (FC) output layer. Equipped with a Softmax activation function, this layer facilitated the classification of input data into predefined classes. Our experiments encompassed a 3-class classification for the iSPL dataset and a 6-class classification for the UCI HAR dataset.

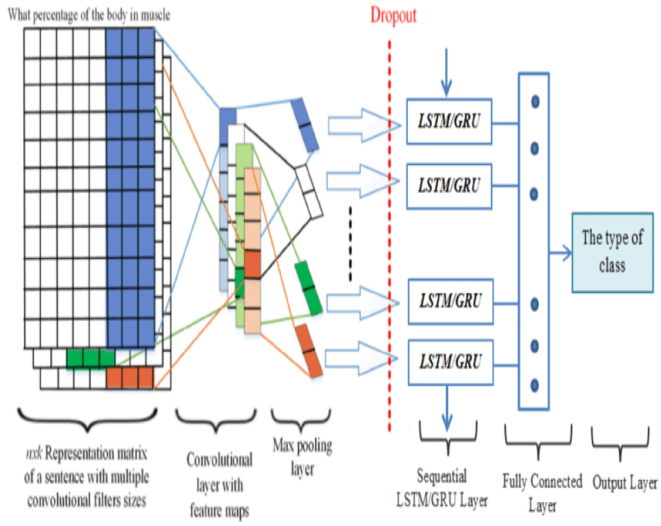

Figure 3 Model Architecture

3.1. Performance Metrics

To assess and compare the performance of our models, we employed several key metrics: Precision, Recall, F1-Score, and Accuracy, each serving a distinct purpose:

Precision: Precision gauges the accuracy of our model's predictions for positive classes. It calculates the ratio of true positives to the sum of true positives and false positives. In mathematical terms, Precision is expressed.

Given the inherent disparities between the convolutional and LSTM layers, we adopted the Keras provided Time Distributed wrapper to apply convolutions while preserving the temporal.

$$\text{Precision} \quad \frac{\text{TPositive}}{\text{TPositive} \; ` \; \text{FPositive}} \qquad (1)$$

Integrity of the data for LSTM processing. The input signal, initially structured as (None, 128, 9), underwent reshaping to (None, 4, 32, 9) to suit the requirements of the Time Distributed 1D convolution layer. This wrapper encapsulated all layers preceding the LSTM components. The flattened feature maps were then channeled into an LSTM layer boasting 128 units and a Rectified Linear Unit (ReLU) activation. This LSTM layer effectively extracted temporal dependencies inherent in sequential data—a crucial aspect, given the sequential nature of signal data. Leveraging the advantages of LSTM, categorized under the recurrent.

Recall: Recall measures our model's effectiveness in identifying actual positive instances. It computes the ratio of true positives to the sum of true positives and false negatives. Mathematically, Recall is defined as

$$\text{Recall} \quad \frac{\text{TPositive}}{\text{TPositive} \; ` \; F\text{Negative}} \qquad (2)$$

F1-Score: The F1-Score quantifies the trade-off between Precision and Recall. It calculates the harmonic mean of Precision and Recall to provide a balanced evaluation metric. The formula for the F1-Score is neural network (RNN) domain, we highlighted its superiority over conventional deep neural networks, as discussed in [1].

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 02 Issue: 04 April 2024
Page No: 889 - 894
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2024.0125

$$F1 \quad \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \qquad (3)$$

Accuracy: Accuracy serves as a straightforward metric representing the ratio of correctly predicted observations to the total number of observations. It takes into account true positives, true negatives, false positives, and false negatives. The accuracy formula is

$$Accuracy \quad \frac{T\ Positive, T\ Negative}{T\ Positive\ T\ Negative\ F\ Positive\ T\ Negative} \qquad (4)$$

These metrics collectively allow us to comprehensively evaluate our model's performance, addressing aspects like precision, recall, balance, and overall correctness in making predictions.

3.2. Performance Analysis

In our comprehensive model evaluation, we conducted an extensive analysis of various ensemble, deep learning, and traditional supervised learning models, uncovering notable performance metrics that offer valuable insights into their capabilities. K-Nearest Neighbors (KNN) demonstrated robust performance, achieving an impressive accuracy of 85%. It excelled with a commendable recall rate of 84.86%, a high precision rate of 85%, and an exceptional F1 score of 84.93%. Conversely, XGBoost displayed remarkable accuracy at 86%, accompanied by outstanding recall and precision rates of 86% and 85.96%, resulting in an impressive F1 score of 84.98%. Random Forest exhibited a commendable accuracy of 86.3%, complemented by a recall rate of 86.2% and a precision rate of 86.1%, yielding an F1 score of 85.98%. Support Vector Machine (SVM) provided solid performance with an accuracy of 80%, supported by precision and recall rates of 80% and 81%, respectively, resulting in an F1 score of 80.5. Long Short-Term Memory (LSTM) outshone all other models, achieving an outstanding 87% accuracy. It was accompanied by a precision rate of 87.1%, a recall rate of 86.95%, and an impressive F1 score of 87.02%. Additionally, the Extra Trees Classifier achieved an accuracy of 87.13%, further solidifying its efficacy. Significantly, our proposed CNN-LSTM-GRU model emerged as the standout performer, boasting a remarkable accuracy of 91.8%. This was coupled with a precision rate of 91.7%, a recall rate of 90.25%, and an exceptional F1 score of 90.47%. These findings collectively highlight the exceptional performance and potential of our ensemble learning model in activity recognition, reaffirming its status as a powerful tool for this application. Looking ahead, we envision exploring more complex human activities and incorporating additional physical attributes to further enhance our system's capabilities. Figure 4 & 5 shows the Accuracy and Loss.

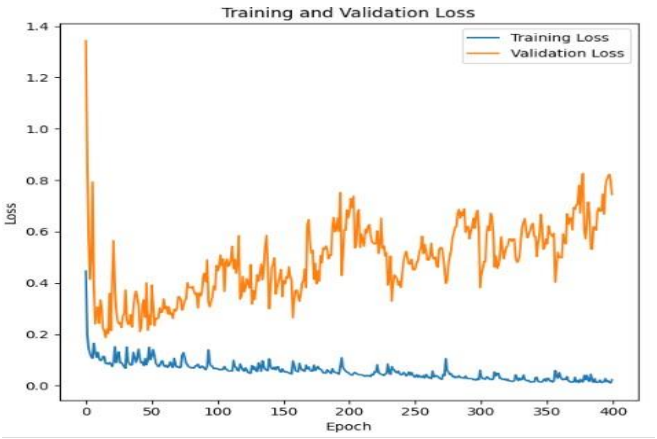

Figure 4 Loss

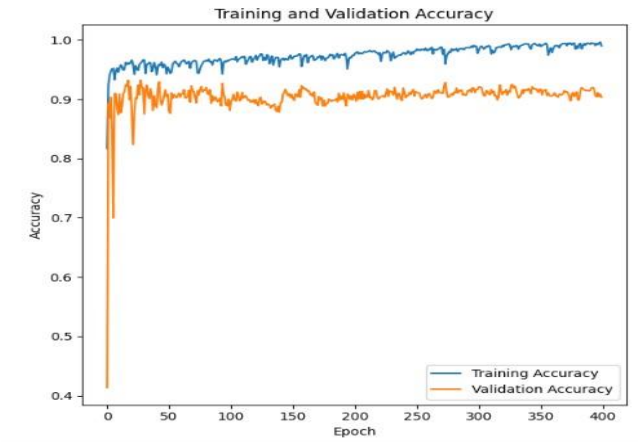

Figure 5 Accuracy

Conclusion

In this study, we introduced a CNN-LSTM approach for human activity recognition, aiming to enhance accuracy by combining the robust feature extraction capabilities of a CNN network with the temporal analysis strengths of an LSTM model used in time series forecasting and classification. The spatial and

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 02 Issue: 04 April 2024
Page No: 889 - 894
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2024.0125

temporal depth of our CNN-LSTM model yielded. Superior performance compared to other deep learning methods utilizing raw signal data. Evaluation on a publicly available dataset (UCI HAR) demonstrated its superiority, achieving over and nearly 2% lower Soft ax loss. While we did not explicitly assess runtime metrics in this paper, our experiments hinted at the efficiency of our proposed approach compared to other models. For future research, we plan to further develop and systematically evaluate the model with varying hyper parameters, including learning rate, batch size, and regularization. The extension of the model to more intricate activities will be explored to address additional challenges in deep learning and human activity recognition, involving assessments on diverse datasets. Additionally, we aim to benchmark our approach against state-of-the-art results for the UCI dataset and other publicly available datasets.

References

[1]. Mutegeki, R. and Han, D. S., "A cnn-lstm approach to human activity recognition," in 2020 International Conference on Artificial Intelligence in Information and Communication (ICAIIC), 2020, pp. 362–366.

[2]. Choudhury, N. A. and Soni, B., "An adaptive batch size-based-cnn-lstm framework for human activity recognition in uncontrolled environment,"IEEE Transactions on Industrial Informatics, vol. 19, no. 10, pp. 10 379–10 387, 2023.

[3]. Mekruksavanich, S. and Jitpattanakul, A., "Smartwatch-based human activity recognition using hybrid lstm network," in 2020 IEEE SENSORS, 2020, pp. 1–4.

[4]. Choudhury, N. A. and Soni, B., "An efficient cnn-lstm approach for smartphone sensor-based human activity recognition system," in 2022 5th International Conference on Computational Intelligence and Networks (CINE), 2022, pp. 01–06.

[5]. Zhang, S., Li, Y., Zhang, S., Shahabi, F., Xia, S., Deng, Y., and Alshurafa, N., "Deep learning in human activity recognition with wearable sensors: A review on advances," Sensors, vol. 22, no. 4, 2022. [Online]. Available: https://www.mdpi.com/1424-8220/22/4/1476

[6]. Khatun, M. A., Yousuf, M. A., Ahmed, S., Uddin, M. Z., Alyami, S. A., Al-Ashhab, S., Akhdar, H. F., Khan, A., Azad, A., and Moni, M. A., "Deep cnn-lstm with self-attention model for human activity recognition using wearable sensor," IEEE Journal of Translational Engineering in Health and Medicine, vol. 10, pp. 1–16, 2022.

[7]. Zhu, J., Chen, H., and Ye, W., "A hybrid cnn–lstm network for the classification of human activities based on micro-doppler radar," IEEE Access, vol. 8, pp. 24 713–24 720, 2020.

[8]. Mekruksavanich, S. and Jitpattanakul, A., "A multichannel cnn-lstm network for daily activity recognition using smartwatch sensor data," in 2021 Joint International Conference on Digital Arts, Media and Technol- ogy with ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunication Engineering, 2021, pp. 277–280.

[9]. Tan, T.-H., Wu, J.-Y., Liu, S.-H., and Gochoo, M., "Human activity recognition using an ensemble learning algorithm with smartphone sensor data," Electronics, vol. 11, no. 3, 2022. [Online]. Available: https://www.mdpi.com/2079-9292/11/3/322

[10]. Challa, S. K., Kumar, A., and Semwal, V. B., "A multibranch cnn-bilstm model for human activity recognition using wearable sensor data," The Visual Computer, vol. 38, no. 12, pp. 4095–4109, 2022.

[11]. Ranasinghe, S., Al Machot, F., and Mayr, H. C., "A review on appli- cations of activity recognition systems with regard to performance and evaluation," International Journal of Distributed Sensor Networks, vol. 12, no. 8, p. 1550147716665520, 2016.