

Towards Multi-View Hand Pose Recognition Using a Fusion of Image Embeddings and Leap 2 Landmarks

Sergio Esteban-Romero¹^a, Romeo Lanzino²^b, Marco Raoul Marini²^c and Manuel Gil-Martín¹^d

¹*Grupo de Tecnología del Habla y Aprendizaje Automático, ETSI Telecomunicación, Universidad Politécnica de Madrid, Av. Complutense 30, 28040, Madrid, Spain*

²*VisionLab, Department of Computer Science, Sapienza University of Rome, Via Salaria 113, Rome 00198, Italy*

Keywords: Multi-View Hand Pose Recognition, Leap Motion Controller 2, Multimodal Data, Multimodal Fusion, Deep Learning.

Abstract: This paper presents a novel approach for multi-view hand pose recognition through image embeddings and hand landmarks. The method integrates raw image data with structural hand landmarks derived from the Leap Motion Controller 2. A Vision Transformer (ViT) pretrained model was used to extract visual features from dual-view grayscale images, which were fused with the corresponding Leap 2 hand landmarks, creating a multimodal representation that encapsulates both visual and landmark data for each sample. These fused embeddings were then classified using a multi-layer perceptron to distinguish among 17 distinct hand poses from the Multi-view Leap2 Hand Pose Dataset, which includes data from 21 subjects. Using a Leave-One-Subject-Out Cross-Validation (LOSO-CV) strategy, we demonstrate that this fusion approach offers a robust recognition performance (F1 Score of $79.33 \pm 0.09\%$), particularly in scenarios where hand occlusions or challenging angles may limit the utility of single-modality data.

1 INTRODUCTION


Non-verbal communication is a crucial component in human interactions, playing a crucial role in expressing emotions, attitudes and intentions beyond words and it represents approximately 65% of human messages (Shin et al., 2024). Besides its human nature, it is also a key element for human-computer interaction (HCI) to make more accessible systems that leverage communicating with electronic devices or with other humans, e.g. via sign language (Miah et al., 2024). In this context, enhancing user interaction in immersive technologies like virtual reality and augmented reality could enable more intuitive and accessible experiences.


The proper recognition of hand poses is challenging due to the numerous joints of the hands, which enable a wide variety of positions and because of occlusions, consequence of the viewpoint, among others (Lee et al., 2024). Consequently, such inconveniences


are also present in some of the datasets that have been traditionally used to design hand pose recognition systems.


The existing solutions to capture dynamics of the human body can be divided into device-based and vision-based systems (Rahim et al., 2020). Device-based solutions often employ wristbands or gloves to track the position of key points that are representative enough of the hand (Lee et al., 2024)(Wang et al., 2023). Analogously, vision-based systems, such as MediaPipe (Zhang et al., 2020) (Chen et al., 2022) aim to capture the most representative positions of the hand by replicating the functioning of device-based systems using only images.

In this work, we present a baseline result using the Multi-view Leap2 Hand Pose Dataset (ML2HP Dataset) (Gil-Martín et al., 2024) on the hand pose recognition task. The dataset includes real images recorded from two different angles to mitigate hand occlusion phenomena alongside landmark coordinates, velocities, orientations, and finger widths relative to the hand. To the best of our knowledge, it is the first result achieved using this dataset. To obtain the baseline result, we first processed the images using a pre-trained Vision Transformer (ViT) (Dosovitskiy

^a <https://orcid.org/0009-0008-6336-7877>

^b <https://orcid.org/0000-0003-2939-3007>

^c <https://orcid.org/0000-0002-2540-2570>

^d <https://orcid.org/0000-0002-4285-6224>

et al., 2020) to be used as input features alongside the raw landmarks. Finally, a multilayer perceptron (MLP) with 17 outputs, equal to the number of poses in the dataset, provides the final probability distribution.

The remainder of this paper is organized as follows. Section 2 briefly describes works focused on the recognition of poses and gestures using different sources. Section 3 provides a description of the dataset, the required data cleaning procedures, the model architecture used to construct the presented baseline and the followed evaluation methodology. Section 4 provides a discussion of the results obtained when validating our model. Section 5 highlights our conclusions and also points towards future lines of research.

2 RELATED WORKS

Hand Gesture Recognition (HGR) has become a crucial area in human-computer interaction, enabling more natural communication with devices through gestures. In the literature, most solutions include a feature extraction method, which can be either manual or based in Artificial Neural Networks (ANN), and a classifier adapted to decode such information (Tan et al., 2023).

Histograms of Oriented Gradients (HOG) and wavelet transforms have been widely used for HGR over the years because of their ability to capture edge and frequency features (Dalal and Triggs, 2005)(Agarwal et al., 2015). However, these techniques often introduce biases stemming from the expert’s choices during feature extraction, as they rely on handcrafted parameters and may not generalize well to diverse datasets. This can limit their effectiveness compared to more modern, data-driven approaches such as deep learning (Tan et al., 2023).

Later, the feature extraction problem has also been addressed using machine learning (ML) strategies to extract features such as convolutional neural networks (Tao et al., 2018) (CNN) or Principal Component Analysis (PCA) (Oliveira et al., 2017) with a specific focus on sign language. Moreover, improvements in the field were achieved for Chinese, Arabic, and Japanese using Deep Learning methods (Yuan et al., 2021)(Aly and Aly, 2020).

Other lines of research explore more complex architectures based on a multi-stage deep learning solution that achieves state-of-the-art results in various HGR datasets such as the creative senz3D dataset (Creative Senz3D) or the Kinetic and Leap Motion Gestures dataset (Kinetic and Leap Motion Gestures)

which comprises RGB images and depth maps.

The latest advances in vision-based RGB systems use Vision Transformers (ViT) (Dosovitskiy et al., 2021) to leverage the capabilities of HGR systems. Current challenges in RGB still image-based hand gesture recognition (HGR) include limited model performance in addressing orientation changes, partial occlusions, and accurately capturing depth and spatial details. Furthermore, the scarcity of diverse datasets and the demand for more computationally efficient models further complicate the development of effective solutions (Shin et al., 2024).

Moreover, other works have been focused on landmark-based approaches for hand pose and gesture recognition. These methods detect key points such as finger joints to capture the structure of the hand and feed deep learning architectures to model and classify hand gestures or poses. For example, previous works used MediaPipe landmarks to feed a transformer and perform a sign language recognition task (Luna-Jiménez et al., 2023), or used several libraries to extract landmarks and perform human pose estimation (Chung et al., 2022).

A promising future direction would be to integrate both image-based and landmark-based modalities. While images capture detailed spatial information, landmarks provide a simplified, efficient hand structure representation. Combining these two sources could enhance model accuracy and generalization, addressing challenges like partial occlusions and orientation changes.

3 MATERIALS AND METHODS

In this section, we describe the dataset that has been used to obtain the baseline result presented in this paper, the data cleaning process, the model architecture and the evaluation methodology followed to train and evaluate the proposed system.

3.1 Dataset

The Multi-view Leap2 Hand Pose Dataset (ML2HP Dataset) (Gil-Martín et al., 2024) is a comprehensive and meticulously curated dataset designed to address the challenges of hand pose recognition in multi-view settings. Captured using two Leap Motion Controller 2 devices, the dataset provides a rich source of real-world data that enables accurate and reliable hand pose recognition models, particularly for human-computer interaction applications.

This dataset comprises 714,000 instances, collected from 21 subjects performing 17 distinct hand

poses, such as "Open Palm", "Closed Fist", "Like" and "OK Sign". The subjects' ages range from 22 to 68 years, with a diverse gender distribution, making the dataset suitable for generalization across different demographic groups. Each instance in the dataset includes real images along with 247 hand properties, such as landmark coordinates, palm velocity, finger orientations, and finger widths. The dataset is also balanced across different hand poses, and hand usage (right or left), ensuring robustness in training machine learning models.

A key feature of the ML2HP dataset is its multi-view recording setup, which employs two Leap Motion Controller 2 devices positioned at complementary angles (Horizontal and Vertical viewpoints), corresponding to a dual-camera setup. This dual-camera configuration mitigates occlusion issues, ensuring that hand poses are captured accurately even when parts of the hand are obscured from one camera's view.

This dataset presents instances where the orientation of the hand relative to each device can influence hand pose detection when using a single device. When the hand faced directly toward the horizontal device, the vertical device often struggled with accurate hand pose detection due to occlusion and limited visibility of the fingers, as illustrated in the Open Palm example of Figure 1. Similarly, when the hand was oriented toward the vertical device, the horizontal device faced comparable challenges, resulting in incomplete landmark representation, as seen in the Like hand pose of Figure 2. However, in certain cases where the hand was positioned diagonally, both devices successfully captured the hand pose accurately, as demonstrated in the Four hand pose of the Figure 3.

This way, this dataset composed of multi-view and multimodal (images and landmarks) information provides a particularly valuable set for developing and testing hand-tracking models that can generalize well across different subjects, and hand usage. Moreover, to the best of our knowledge, no baseline performance metrics exist for this dataset, allowing us to explore and establish an initial benchmark for hand pose recognition task using the available data.

3.2 Data Cleaning

When exploring the dataset we encounter some issues coming from the official acquisition program provided to operate Leap Motion Controller 2. In particular, we identified two different types of issues:

- **Missing Values:** in one specific frame out of the 714,000 available, we find missing values in 29

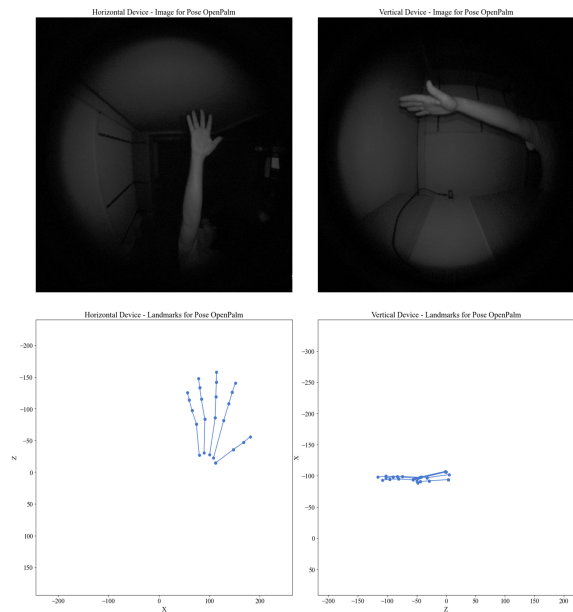


Figure 1: Images and 2D landmark representations for OpenPalm class from both viewpoint devices (Gil-Martín et al., 2024).

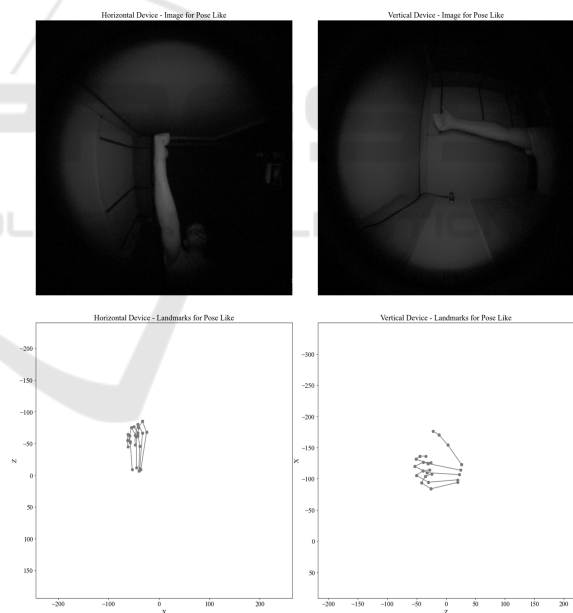


Figure 2: Images and 2D landmark representations for Like class from both viewpoint devices (Gil-Martín et al., 2024).

columns for the horizontal device and 55 for the vertical device recordings. The specific metadata associated with the frame can be found in Table 1. We assigned the mean value of the column to those missing values. No further exploration of which value to assign has been carried out.

- **Non Printable Characters:** some float values include "non-printable" characters which lead to

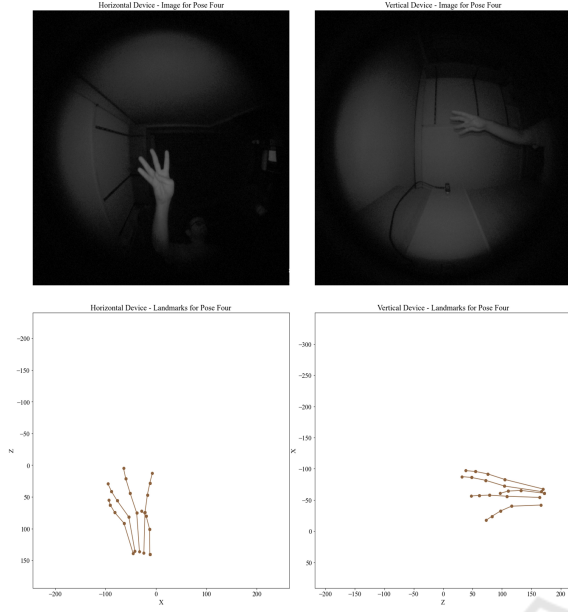


Figure 3: Images and 2D landmark representations for Four class from both viewpoint devices (Gil-Martín et al., 2024).

Table 1: Metadata associated to the frame containing missing values in the dataset.

Property	Value
frame_id	232
subject_id	19
which_hand	Left_Hand
pose	OpenPalm
device	Horizontal

errors in the system as they were recognized as strings. We hypothesize that this issue may also come from the official acquisition system provided or from the usage of different encodings when processing the files. However, we did not conduct further investigation on the matter. To overcome that difficulty, we used regular expressions to isolate those specific values for their convenient cleansing and casting into actual float numbers.

3.3 Model Architecture

Our approach tackles some of the most concerning problems associated to HGR using the ML2HP Dataset. Specifically, we propose a solution that for every sample integrates the information of both viewpoints including the image and the hand landmark information. The architecture is illustrated in Figure 4.

Specifically, we use a pretrained Vision Transformer (ViT) model¹ from Huggingface as the feature

¹<https://huggingface.co/google/vit-base-patch16-384>

Table 2: Dimensions corresponding to each of the variables involved in the proposed model architecture, where C is the number of channels, H and W correspond to the height and width of the image, $N_{vertical}$ and $N_{horizontal}$ correspond to the ViT output for each viewpoint image, $N_{landmarks}$ is the number of available landmarks per viewpoint, and $N_{classes}$ corresponds to the number of hand poses.

Variable	Dim
C	3
H	512
W	512
$N_{vertical}$	768
$N_{horizontal}$	768
$N_{landmarks}$	242
$N_{classes}$	17

extractor. The ViT model is designed to process only images with three channels, which presents a limitation when working with single-channel grayscale images. To address this, we replicated the intensity values of the grayscale images across the remaining two channels, effectively converting them into three-channel images.

Then, we extract the Classification (CLS) token from the model’s final hidden state, which serves as a compact representation of the entire image (Dosovitskiy et al., 2020). This representation is then concatenated with the extracted landmarks, resulting in a multimodal embedding that encapsulates all the relevant information for each sample.

Finally, to classify each sample into one out of the 17 possible classes we use the multimodal embedding as input for a MLP that will be adapted throughout the training process. The layers in the MLP have the following dimensionality: $input_dim \rightarrow 1,024 \rightarrow 512 \rightarrow 256 \rightarrow 128 \rightarrow 17$, where $input_dim = N_{horizontal} + N_{vertical} + 2 * N_{landmarks}$. The specific dimensions of each variable are presented in Table 2. We used CrossEntropyLoss Pytorch implementation as loss function to optimize throughout the training process.

For our model, we used Adam optimizer (Kingma and Ba, 2017) and set the learning rate to 10^{-5} to ensure gradual and stable updates during training. The batch size was set to 256, balancing memory usage and training efficiency. Training was conducted for a maximum of 3 epochs to prevent overfitting and reduce computational time considering the large amount of samples available. No validation of the selected hyperparameters has been performed in this study.

Arguably, there might be some over-representation as well as dependencies between the input features since the landmarks are descriptors of the hands that appear in the images. However, we consider both modalities can complement each other,

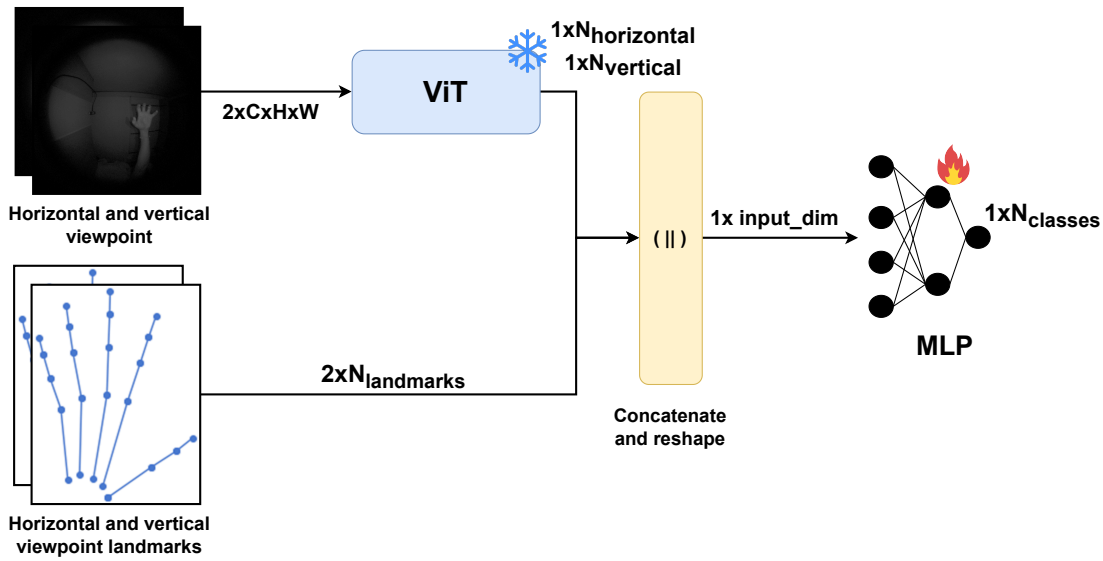


Figure 4: Architecture used to create the baseline result for ML2HP Dataset where a pretrained ViT is used as feature extractor for images. Then, in combination with the raw landmark values, the MLP is fine-tuned to solve the task. The snowflake indicates those parts that remain frozen and the flame those fine-tuned throughout the training process.

specially in those situations where we experience hand occlusions or there is not a clear view of the hand from any viewpoint. Besides this hypothesis, we encourage researchers to further investigate the relevance of each modality as well as the existing correlation between features.

3.4 Evaluation Methodology

To evaluate the system using the whole dataset in a subject-independent scenario, we employed a Leave-One-Subject-Out Cross-Validation (LOSO-CV) approach as the data distribution strategy. In this methodology, data from all subjects except one are used to train the system, while the data from the left-out subject are used to test it. This process is repeated, with each subject being left out in turn, and the results are averaged across all iterations. This approach simulates a realistic scenario where the system is evaluated with recordings from subjects not used in the training phase.

As evaluation metrics, we used accuracy, which is defined as the ratio between the number of correctly classified samples and the total number of samples. For a classification problem with N testing examples and $N_{classes}$ classes, accuracy is defined in Equation 1:

$$\text{Accuracy} = \frac{1}{N} \sum_{i=1}^{N_{classes}} P_{ii} \quad (1)$$

Considering R_i as the sum of all examples in the i -th column of the confusion matrix and S_i as the sum

of all examples in the i -th row, the precision (Equation 2), recall (Equation 3), and F1-score (Equation 4) metrics are defined as follows:

$$\text{Precision} = \frac{1}{N_{classes}} \sum_{i=1}^{N_{classes}} \frac{P_{ii}}{R_i} \quad (2)$$

$$\text{Recall} = \frac{1}{N_{classes}} \sum_{i=1}^{N_{classes}} \frac{P_{ii}}{S_i} \quad (3)$$

$$\text{F1-score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (4)$$

To show statistical significance, we used confidence intervals, which represent plausible values for a specific metric. A significant difference between the results of two experiments is established when their confidence intervals do not overlap. Equation 5 shows the computation of confidence intervals associated with a specific metric value and N samples for a 95% confidence level:

$$\text{CI}(95\%) = \pm 1.96 \cdot \sqrt{\frac{\text{metric} \cdot (100 - \text{metric})}{N}} \quad (5)$$

4 RESULTS AND DISCUSSION

Table 3 shows the results for the LOSO-CV evaluation, with an F1 Score of 79.33 ± 0.09 %. This score reflects a robust performance, indicating a strong balance between precision and recall, which is essential

for reliable classification outcomes. However, considering the substantial sample size of 714,000 examples, there exists still work for improvement in overall performance.

Table 3: Performance metrics with 95% confidence intervals for the LOSO-CV evaluation.

Metric	Value (%)
Accuracy	79.65 ± 0.09
Precision	80.63 ± 0.09
Recall	79.65 ± 0.09
F1 Score	79.33 ± 0.09

The confusion matrix related to the results is shown in Figure 5 and highlights some of the best and worst classified hand poses.

Considering that there are 42,000 examples per class, One hand pose is the best classified, achieving 39,719 correct predictions with minimal misclassification. Spiderman also performs well, with 36,827 correctly identified instances, although it is occasionally confused with Stop (2,886 times). This confusion likely arises from their similar configurations, both involving extended fingers, making them harder to distinguish when occluded or viewed from certain angles. Open Palm is another well-recognized pose, with 36,685 correct classifications, though some instances are misclassified as Tiger (3,359 times), reflecting the slight overlap in appearance when fingers are not fully straightened or curled.

On the other hand, some poses show poor performance, with quite misclassifications. OK is one of the worst classified, with 28,770 correct predictions. It is frequently confused with OpenPalm (3,735 times), likely because both involve the extension of several fingers, and the circular thumb-index gesture in OK can sometimes appear flattened or ambiguous from certain angles. Rock also struggles, achieving only 28,910 correct classifications. It is often confused with Tiger (6,574 times) and Spiderman (2,279 times), as these poses share similar elements, such as the partial extension of specific fingers, which can be difficult to distinguish under certain viewpoints.

These results indicate that subtle differences in finger arrangements and slight variations in curvature contribute to misclassifications. For example, the similarity between Rock and Spiderman, with their partially extended fingers, highlights the difficulty of accurately distinguishing between such poses. Likewise, OpenPalm being misclassified as Tiger suggests that some pose may lack enough distinct visual cues when viewed from certain perspectives.

These findings suggest future research directions, including the need to explore the impact of camera perspective. Identifying which camera—horizontal or

vertical—provides the clearest view of each instance may improve recognition by leveraging the most informative viewpoint.

5 CONCLUSIONS

In order to establish a baseline for future research, we evaluated our proposed method on the Multi-view Leap2 Hand Pose Dataset using a LOSO-CV strategy in order to provide an accurate system capable to generalize across different individuals. To the best of our knowledge, this is the first result achieved using this dataset. The proposed architecture used ViT to extract features from images, demonstrating the advantages of a multimodal approach that combines image data with hand landmark information. This integration offers a robust performance, even when in some cases hand pose may be occluded. The system offers a F1 Score of 79.33 ± 0.09 %, which indicates strong classification performance across the dataset. In particular, the confusion matrix reveals specific poses that are frequently misclassified, such as Tiger and Open Palm, suggesting a need for enhanced strategies to distinguish between similar poses.

Future work could explore alternative partitioning strategies for the ML2HP dataset, such as separating training and testing based on distinct camera orientations or hand dominance. Additionally, detecting which camera (horizontal or vertical) the hand is primarily oriented towards for each instance and evaluating performance using data from only that viewpoint would provide insights into how much information can be effectively extracted from a single perspective. In addition, it could be possible to investigate the impact of using only one modality—either the image data or the hand landmarks—rather than both. This would help determine whether the visual features alone or the landmark information is enough for accurate hand pose recognition in certain scenarios. Exploring these configurations would provide a clearer understanding of the individual contributions of each modality and help develop more efficient models that optimize either visual or landmark-based recognition. Furthermore, applying the proposed model to new datasets will provide deeper insights into its functionality and improve the understanding of its practical performance.

ACKNOWLEDGEMENTS

Sergio Esteban-Romero research was supported by the Spanish Ministry of Education (FPI grant

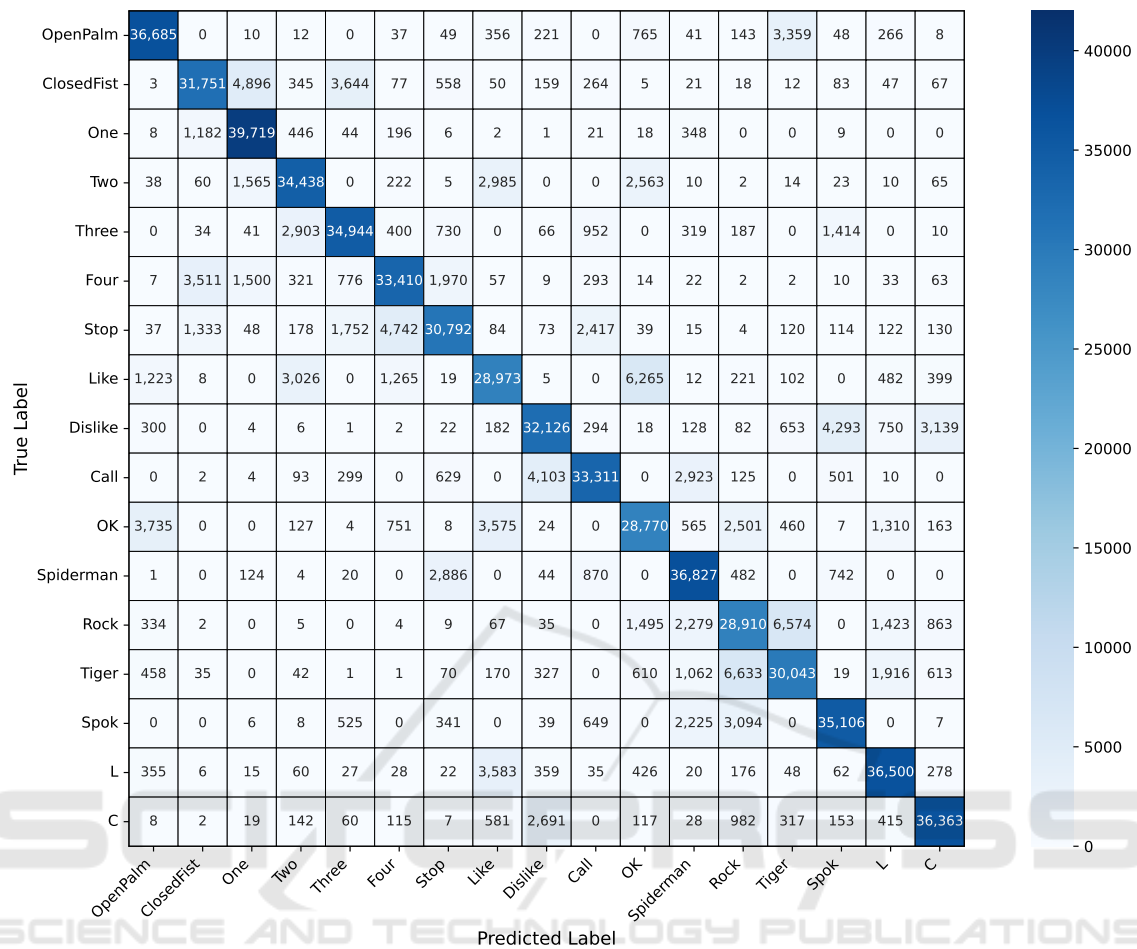


Figure 5: Confusion matrix for the baseline system using a LOSO-CV.

PRE2022-105516). This work was funded by Project ASTOUND (101071191 — HORIZON-EIC-2021-PATHFINDERCHALLENGES-01) of the European Commission and by the Spanish Ministry of Science and Innovation through the projects TRUSTBOOST (PID2023-150584OB-C21 and PID2023-150584OB-C22), GOMINOLA (PID2020-118112RB-C21 and PID2020-118112RB-C22) and BeWord (PID2021-126061OB-C43), funded by MCIN/AEI/10.13039/501100011033 and by the European Union “NextGenerationEU/PRTR”.

REFERENCES

- Agarwal, R., Raman, B., and Mittal, A. (2015). Hand gesture recognition using discrete wavelet transform and support vector machine. In *2015 2nd International Conference on Signal Processing and Integrated Networks (SPIN)*, pages 489–493.
- Aly, S. and Aly, W. (2020). Deeparslr: A novel signer-independent deep learning framework for isolated arabic sign language gestures recognition. *IEEE Access*, 8:83199–83212.
- Chen, R.-C., Manongga, W. E., and Dewi, C. (2022). Recursive feature elimination for improving learning points on hand-sign recognition. *Future Internet*, 14(12).
- Chung, J.-L., Ong, L.-Y., and Leow, M.-C. (2022). Comparative analysis of skeleton-based human pose estimation. *Future Internet*, 14(12).
- Dalal, N. and Triggs, B. (2005). Histograms of oriented gradients for human detection. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 886–893 vol. 1.
- Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houtsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *CoRR*, abs/2010.11929.
- Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houtsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *CoRR*, abs/2010.11929.

- N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale.
- Gil-Martín, M., Marini, M. R., San-Segundo, R., and Cinque, L. (2024). Dual leap motion controller 2: A robust dataset for multi-view hand pose recognition. *Sci. Data*, 11(1).
- Kingma, D. P. and Ba, J. (2017). Adam: A method for stochastic optimization.
- Lee, C.-J., Zhang, R., Agarwal, D., Yu, T. C., Gunda, V., Lopez, O., Kim, J., Yin, S., Dong, B., Li, K., et al. (2024). Echowrist: Continuous hand pose tracking and hand-object interaction recognition using low-power active acoustic sensing on a wristband. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–21.
- Luna-Jiménez, C., Gil-Martín, M., Kleinlein, R., San-Segundo, R., and Fernández-Martínez, F. (2023). Interpreting sign language recognition using transformers and mediapipe landmarks. In *Proceedings of the 25th International Conference on Multimodal Interaction, ICMI '23*, page 373–377, New York, NY, USA. Association for Computing Machinery.
- Miah, A. S. M., Hasan, M. A. M., Tomioka, Y., and Shin, J. (2024). Hand gesture recognition for multi-culture sign language using graph and general deep learning network. *IEEE Open Journal of the Computer Society*.
- Oliveira, M., Chatbri, H., Ferstl, Y., Farouk, M., Little, S., O'Connor, N. E., and Sutherland, A. (2017). A dataset for irish sign language recognition.
- Rahim, M. A., Miah, A. S. M., Sayeed, A., and Shin, J. (2020). Hand gesture recognition based on optimal segmentation in human-computer interaction. In *2020 3rd IEEE International Conference on Knowledge Innovation and Invention (ICKII)*, pages 163–166. IEEE.
- Shin, J., Miah, A. S. M., Kabir, M. H., Rahim, M. A., and Al Shiam, A. (2024). A methodological and structural review of hand gesture recognition across diverse data modalities. *IEEE Access*.
- Tan, C. K., Lim, K. M., Chang, R. K. Y., Lee, C. P., and Alqahtani, A. (2023). Hgr-vit: Hand gesture recognition with vision transformer. *Sensors*, 23(12).
- Tao, W., Leu, M. C., and Yin, Z. (2018). American sign language alphabet recognition using convolutional neural networks with multiview augmentation and inference fusion. *Engineering Applications of Artificial Intelligence*, 76:202–213.
- Wang, H., Ru, B., Miao, X., Gao, Q., Habib, M., Liu, L., and Qiu, S. (2023). Mems devices-based hand gesture recognition via wearable computing. *Micromachines*, 14(5).
- Yuan, G., Liu, X., Yan, Q., Qiao, S., Wang, Z., and Yuan, L. (2021). Hand gesture recognition using deep feature fusion network based on wearable sensors. *IEEE Sensors Journal*, 21(1):539–547.
- Zhang, F., Bazarevsky, V., Vakunov, A., Tkachenka, A., Sung, G., Chang, C.-L., and Grundmann, M. (2020). Mediapipe hands: On-device real-time hand tracking.