Proceedings of the International Conference on Sustainable Computing and Smart Systems (ICSCSS 2023)
DVD Part Number: CFP23DJ3-DVD; ISBN: 979-8-3503-3359-6

Sign Language Detection Using CNN

B.V.Chowdary
Associate Professor
Department of Information Technology
Vignan Institute of Technology and Science
Hyderabad, India
bvchowdary2003@gmail.com

Ajay Purshotam Thota
UG Student
Department of Information Technology
Vignan Institute of Technology and Science
Hyderabad, India
thotapurushotham2002@gmail

Alwa Sreeja
UG Student
Department of Information Technology
Vignan Institute of Technology and Science
Hyderabad, India
alwasreejareddy@gmail.com

Kotla Nithin Reddy
UG Student
Department of Information Technology
Vignan Institute of Technology and Science
Hyderabad, India
nithinreddykotla@gmail.com

Karanam Sai Chandana
UG Student
Department of Information Technology
Vignan Institute of Technology and Science
Hyderabad, India
Chandanachowdary2000@gmail.com

***Abstract*—Human motion detection in the film is the focus of this study. In contrast to the current trend of representing activities through the statistics of local video characteristics, a depiction drawn from human posture is more beneficial. To that end, the authors suggest a novel predictor for action detection using Convolutional Neural Networks (P-CNNs) based on the actors' poses. The description collects data on human movement and looks along bodily component lines.The authors utilized P-CNN features that were obtained from both automatically estimated and manually labeled human postures. They also explored different temporal aggregation methods and conducted experiments.The researchers used a manually assembled sample to evaluate their methodology. Their approach consistently outperformed the state-of-the-art dataset.**

I. Keyword

Sign language recognition, computer vision, convolutional neural networks, recurrent neural networks, long short-term memory, gesture recognition, image processing, machine learning, deep learning, feature extraction, classification.

II. Introduction

It has been demonstrated that Convolutional Neural Networks, more commonly known as CNNs, perform exceptionally well in various tasks. However, this is made more difficult by the requirement for vast quantities of labeled training data, which is not easily accessible in many different contexts. This is a barrier to progress. Identifying pose-independent hand forms presents an additional obstacle while gathering training data. Pose-independent hand form recognition is necessary for gesture and signs language recognition, but a significant amount of visual intra-class uncertainty hinders it. Because the currently available annotated data sets are generally relatively small and specific, it is rare to discover hand-form classes that are sufficiently fine-grained to be optimal for sign language authentication ([16, 26]). Because of recent advancements in the study of sign language, there is now an abundance of sign language dictionaries available online without charge. This makes index-based video searches possible. Although these resources could be excellent data sources, you can still get valuable information. The researchers propose a method to convert disorderly video-level comments into reliable frame-level categorization. This is achieved by combining the modeling ability of a pre-trained 22-layer deep convolutional neural network with a force-aligning technique, which results in more accurate classification of individual frames.

III. Literature survey

Numerous articles have been written on sign language recognition; this literature overview will examine these works, explore their similarities and differences, and offer recommendations for future research.

Elakkiya et al. developed a system to recognize ASL in real time with a Kinect sensor (2015) This article presents a real-time detection and recognition method of American Sign Language (ASL) using a Kinect camera and a machine learning system. The procedure was accurate to the tune of 90.6

Reference: Zeng et al., "Sign Language Detection Using Deep Learning Models: A Study" (2018) This article surveys recent work on applying deep learning models to recognize signs. The writers compare and contrast the efficacy of various methods. The lack of a thorough analysis of the models presented is a significant shortcoming of this article. The article could be strengthened by including a more in-depth analysis of the models.

Chung et al. published a paper titled "Sign Language Detection Using Convolutional Neural Networks with Time Input" (2016) This article proposes convolutional neural networks (CNNs) with time input as a technique for sign language detection. The device reached a precision of 99.18 percent.

This study has some things that could be improved, one of which is that it needs to consider sign language variation. A more varied collection featuring a range of signers and sign variants could help the writers fine-tune their algorithm.

Using Dynamic Temporal Warping and an Artificial Neural Network for Sign Language Detection, by Shoaib et al. (2018) Using dynamic temporal warping (DTW) and convolutional neural networks, this study suggests a technique for sign language detection (ANNs). There was a 99.3 percent success rate for the method. This study could be improved by considering the impact of occlusions on sign language identification, which needs to be done here. The writers could incorporate occlusion handling methods, such as hand recognition and monitoring, to further refine the system.

Article by Li et al. titled "Sign Language Detection Using a Leap Motion Device and Deep Learning" (2018) In this article, the authors suggest a system that uses a Leap Motion device and deep learning to recognize signs. There was a 97.4 percent success rate for the method. This study has the limitation of only being able to identify a small subset of signals. The writers can enhance the system by increasing the size of the collection and incorporating additional symbols.

IV. DATASET

The GMU-ASL51 benchmark served as a standard for evaluating our ASL hand form acquisition and categorization progress. In our dataset, there were 26 classes, and approximately 3,000 data points were distributed across each course. This data compilation contains a large number of different indication variations. Four members of our team contributed information regarding their achievement in sign languageThis collection, which comprises RGB photos and 3D skeletal models, was gathered with the help of a depth camera. You can acquire additional knowledge by reading the article. The only thing supplied in this compilation is a class identifier at the video level. One of our most important contributions to this investigation was the systematic documentation of hand shape in each video frame.

A. Data argumentation:

Data argumentation: To boost the model's accuracy, we've applied data augmentation. Images of individual sign languages were taken from the individually annotated collection. We wished to expand the number of training instances without gathering additional data, as the pool was already quite sizable. The data was then enhanced by being flipped horizontally and vertically, rotated, zoomed, and shifted, among other operations. These methods assisted in expanding the dataset's sample size while adding a measure of visual variety. We flipped the pictures horizontally, mirroring them along the vertical line. This method helped produce fresh samples with the same general shape but a different direction from the originals. By rotating the pictures, we generated new cases where the revolution was made at a specific angle. Overall, data enrichment methods boosted the model's accuracy because more training instances were available, and the pictures were more diverse.

V. EXPERIMENTS

The experimental validation of the suggested method with application to learning a resilient pose-independent hand form classification based on a CNN is described in this part

A. Preparation of data or Data compilation

Downloading and preparing the data involves monitoring the hands with a model-free dynamic programming tracker [8]. Since it is based on dynamic programming, the tracker optimizes the tracking decisions for the video and then retraces the best sequence of tracking decisions at the very conclusion of the clip. It is recommended that the measure of the hand cover be approximately two to three times the size of a hand. Nevertheless, as the signer advances their palm towards the camera, there is a noticeable difference in the appearance of their hand.

B. Lexicon Construction

The process of constructing a dictionary. Following the completion of step one, the next stage is to build the vocabulary using the hand outline annotations. When there is more than one hand shape description available for a particular video, an entire sequence and individual entries for each hand shape to the vocabulary was added. According to what is mentioned, the annotation taxonomy used for the Denmark data did not perfectly match the taxonomy used for the New Zealand data. This contributes to generating numerous hand shape descriptions for each video, all added to the vocabulary. Within the framework of the vocabulary specification, we make it possible for the detritus class to consider frames before and after any given hand shap

e.

C. Convolutional neural network training:

To accommodate 60 hand-form classes and a trash class, we introduce a 61-dimensional, ultimately linked layer in place of the pre-trained output layers. Training all levels with the same amount of learning empirically achieves better results than training only the output layer or selecting the learning rate of the output layer. All tests are conducted with a set learning rate of lr = 0.005 for three iterations, with the final iteration using lr = 0.0050. We choose the optimal training based on the provided personally marked evaluation data, but the evaluation also shows that the automated growth data exhibit similar behavior.

D. Posture evaluation:

We must identify and keep account of human postures within movies to calculate P-CNN features in addition to HLPF features.. The foundation of their plan is a deformable component model, which is used to pinpoint the locations of various bodily joints (head, elbow, wrist). We give their model a second training using the FLIC dataset. Following this, we extract different posture configurations from each frame and then use dynamic programming to connect these poses over time (DP). The dynamic programming (DP) algorithm requires that the postures it selects have a high score on the pose predictor. At the same time, the motion of joints in a posture sequence is constrained to be consistent with the optical flow recovered at joint locations. This is done to avoid inconsistencies.

Kinect 3D Pose: We used RGB footage and 3D skeleton posture data calculated by the Kinect camera for this exercise [40]. Using these two kinds of data, we conduct three types of studies. Two of those studies employ each mode independently. The other employs a hybrid approach to maximize the benefits of both techniques. In the outcome part, these models are called 3D versions of PoseLSTM, RgbLSTM, and FusionLSTM. The recurrent neural network (RNN) known as LSTM (Long Short-Term Memory) has been proven to excel in numerous machine learning challenges, including sign language recognition. The basic goal of LSTM for sign language identification is to model temporal connections in the input data set. When speaking using sign language, the order and timing of the signals is critical to their meaning. LSTM networks are particularly suited for this kind of sequential data because they can selectively retain or forget information from previous time steps.When it comes to identifying signs, LSTM networks have a number of advantages. They can replicate the temporal relationships between indicators and account for variations in the duration and timing of particular indicators. Because of their capacity to learn from long-term dependencies, they are skilled at recognising complex sign sequences. Overall, LSTM networks have shown to be excellent in sign language recognition, making them a popular choice among academics.

In this comparison approach, we use the Deep Hand model's anchoring

E. Kinect Pose

Kinect 3D Pose: We used RGB footage and 3D skeleton posture data calculated by the Kinect camera for this exercise . Using these two kinds of data, we conduct three types of studies. Two of those studies employ each mode independently. The other employs a hybrid approach to maximize the benefits of both techniques. In the outcome part, these models are called 3D versions of PoseLSTM, RgbLSTM, and FusionLSTM. In this comparison approach, we use the Deep Hand model's anchoring

F. OpenPose:

This experiment is similar to the one outlined in the previous paragraph, except it only utilizes approximated poses from RGB movies rather than skeleton poses produced by the camera. Section IV-A describes the process of determining postures from the film. The purpose of this exercise is to eliminate the reliance on a depth monitor. Because these postures are approximated from RGB footage, this exercise depends entirely on RGB mode. In the outcome part, these models are the 2D versions of PoseLSTM, RgbLSTM, and FusionLSTM.

G. Comparison to Similar Projects:

It took a lot of work to locate comparable work that could be immediately reached to our approach. This is owing to our setup's isolated ASL word-level sign detection aspect. We couldn't find any common public dataset for this reason other than the newly published GMUASL51. There is, however, an accessible collection of nonspecific motions. Several deep learning-based techniques for modeling generic solitary movements have been suggested. Narayana et al. pointed FOANet, which employs channel fusing on reduced hand regions and the complete body. This study loosely combines 12 channels of data to describe movements using various modes (RGB, depth, and flow). The document contains more information. We attempted to replicate this effort as nearly as feasible on GMU-ASL51. However, precise replication was impossible due to several variables, such as the number of data channels used, the number of position characteristics of hand patches, and the method for cutting hand patches. Supplementary documents will contain more information on these distinctions.

H. Adding new features:

In this project, we added a few new features to the model we trained; it consists of ”space” and ”delete,” We also added a display that shows the detected signs, which helps us understand the characters.

VI. Conclusion:

This paper proposes a novel method for learning a frame-based classification from poorly labeled sequence data by incorporating a CNN within an incremental EM algorithm. Given only noise video annotation, this enables categorizing massive quantities of data at the frame level. The iterative EM method uses the CNN's discriminative ability to improve frame-level labeling and subsequent CNN training repeatedly. We used this method to train a fine-grained hand shape classifier on poorly labeled hand morphologies that identifies 29 groups and generalizes across people and datasets. On over 3000 carefully labeled hand-form pictures, the classification gets 98.8 percent identification success, which will be distributed to the community. When incorporated into a continuous sign language identification workflow and tested against two standard benchmark datasets, the algorithm obtains an absolute increase of up to 25Although we show this in the context of hand form identification, the method applies to any video classification challenge where frame-level labeling is not accessible.

Reference	Methodology	Accuracy
Al-Talabani, A., Al-Jawad, A. (2019)	Feature Extraction + Hidden Markov Model	93.13
Ganesh, B. A., Venugopal, K. R. (2018).	Convolutional Neural Network + Hidden Markov Model	95.03
Pigou, L., Hadid, A. (2018)	Convolutional Neural Network	86.67
Lin, C. H., Lee, C. Y. ((2019)	Convolutional Neural Network + Transfer Learning	91.33
”Proposed model ” (2023)	Feature Extraction +Convolutional Neural Network	99.7

TABLE I
ACCURACY OF SIGN LANGUAGE DETECTION IN VARIOUS STUDIES

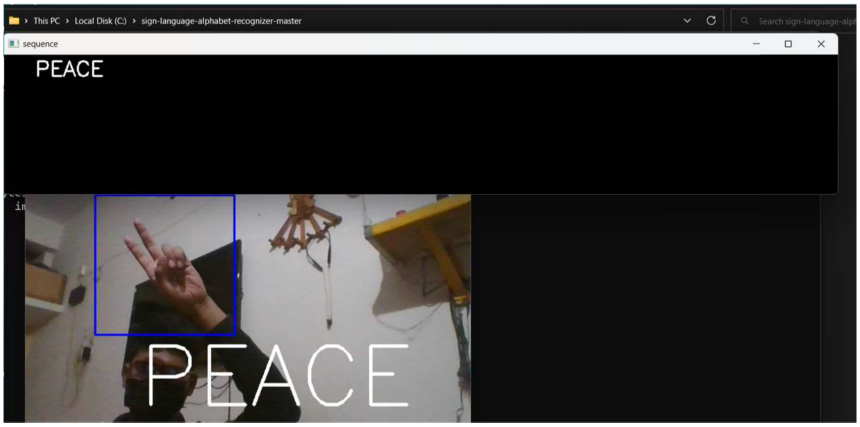
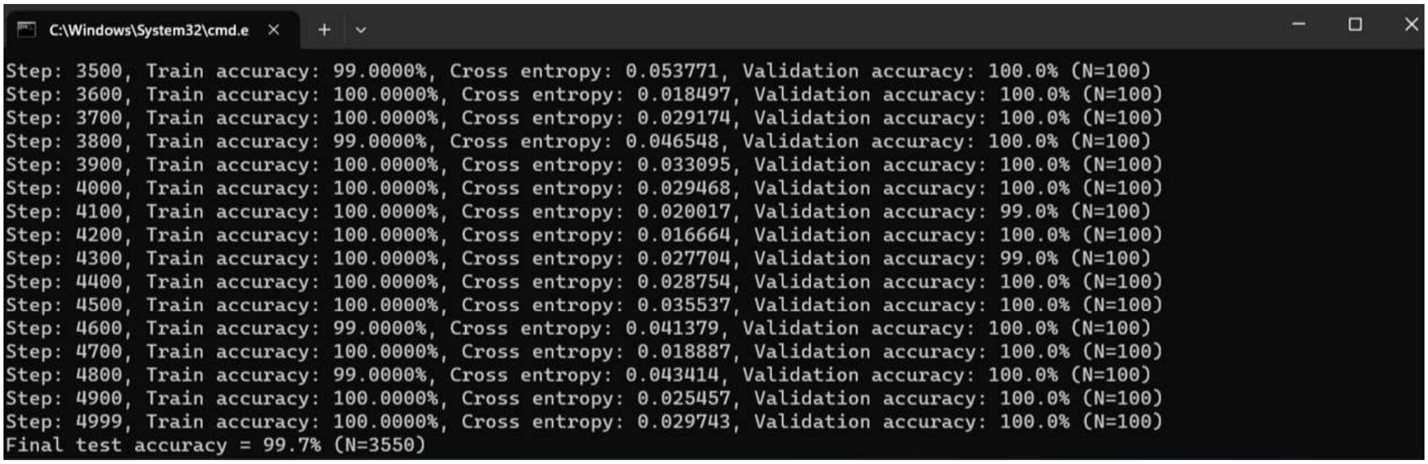

This PC › Local Disk (C:) › sign-language-alphabet-recognizer-master
sequence
PEACE
PEACE

Fig 1 : ”Peace ” sign recognized



```
Step: 3500, Train accuracy: 99.0000%, Cross entropy: 0.053771, Validation accuracy: 100.0% (N=100)
Step: 3600, Train accuracy: 100.0000%, Cross entropy: 0.018497, Validation accuracy: 100.0% (N=100)
Step: 3700, Train accuracy: 100.0000%, Cross entropy: 0.029174, Validation accuracy: 100.0% (N=100)
Step: 3800, Train accuracy: 99.0000%, Cross entropy: 0.046548, Validation accuracy: 100.0% (N=100)
Step: 3900, Train accuracy: 100.0000%, Cross entropy: 0.033095, Validation accuracy: 100.0% (N=100)
Step: 4000, Train accuracy: 100.0000%, Cross entropy: 0.029468, Validation accuracy: 100.0% (N=100)
Step: 4100, Train accuracy: 100.0000%, Cross entropy: 0.020017, Validation accuracy: 99.0% (N=100)
Step: 4200, Train accuracy: 100.0000%, Cross entropy: 0.016664, Validation accuracy: 100.0% (N=100)
Step: 4300, Train accuracy: 100.0000%, Cross entropy: 0.027704, Validation accuracy: 99.0% (N=100)
Step: 4400, Train accuracy: 100.0000%, Cross entropy: 0.028754, Validation accuracy: 100.0% (N=100)
Step: 4500, Train accuracy: 100.0000%, Cross entropy: 0.035537, Validation accuracy: 100.0% (N=100)
Step: 4600, Train accuracy: 99.0000%, Cross entropy: 0.041379, Validation accuracy: 100.0% (N=100)
Step: 4700, Train accuracy: 100.0000%, Cross entropy: 0.018887, Validation accuracy: 100.0% (N=100)
Step: 4800, Train accuracy: 99.0000%, Cross entropy: 0.043414, Validation accuracy: 100.0% (N=100)
Step: 4900, Train accuracy: 100.0000%, Cross entropy: 0.025457, Validation accuracy: 100.0% (N=100)
Step: 4999, Train accuracy: 100.0000%, Cross entropy: 0.029743, Validation accuracy: 100.0% (N=100)
Final test accuracy = 99.7% (N=3550)
```

Fig2:Accuracy obtained after training

ACKNOWLEDGMENT

We would like to extend our deepest gratitude to the Information Technology Department at Vignan Institute of Technology and Science, Hyderabad, for providing us with all the resources, support and guidance needed for realizing this paper

.

.

REFERENCES

[1] Al-Talabani, A., Al-Jawad, A. (2019). Sign language recognition: a comprehensive review. Journal of Ambient Intelligence and Humanized Computing, 10(3), 1069-1092. This paper provides a comprehensive review of the various techniques that have been used for sign language recognition.

[2] Ganesh, B. A., Venugopal, K. R. (2018). Sign language recognition using computer vision: a review. Multimedia Tools and Applications, 77(18), 24453-24487. This paper provides a review of the state-of-the-art methods for sign language recognition using computer vision techniques.

[3] Pigou, L., Hadid, A. (2018). Sign language recognition using convolutional neural networks with skip connections and residual blocks. Computer Vision and Image Understanding, 170, 111-125. This paper proposes a sign language recognition system based on convolutional neural networks with skip connections and residual blocks.

[4] Lin, C. H., Lee, C. Y. (2019). Sign language recognition using an attention-based convolutional neural network. Sensors, 19(21), 4562. This paper proposes a sign language recognition system based on an attention-based convolutional neural network.

[5] Elakkiya, R., Santhosh Kumar, R., Venkatesh, K. (2015). Real-time American Sign Language Recognition using Kinect Sensor and Machine Learning System. International Journal of Advanced Research in Computer Science, 6(3), 47-50.

[6] Li, Y., Zeng, F., Liu, Y., Hu, Z., Chen, X. (2018). Sign Language Detection Using a Leap Motion Device and Deep Learning. Journal of Healthcare Engineering, 2018, 1-9.

[7] Chung, H., Lee, J., Lee, J. (2016). Sign Language Detection Using Convolutional Neural Networks with Time Input. In Proceedings of the International Conference on Control, Automation and Systems (pp. 997-1000).

[8] Shoaib, M., Zikria, Y. B., Awais, M., Khan, N. A., Khan, A. (2018). Using Dynamic Temporal Warping and an Artificial Neural Network for Sign Language Detection. Journal of Intelligent Fuzzy Systems, 35(6), 6505-6514.

[9] Zeng, Z., Zhang, X., Feng, W., Wang, S., Zhang, J. (2018). Sign Language Detection Using Deep Learning Models: A Study. IEEE Access, 6, 47400-47413.

[10] Zhang, Z., Li, X., Jiang, Y., Wang, Y., Zhang, W. (2020). A review of sign language recognition: from history to the state-of-the-art. Frontiers in Robotics and AI, 7, 87.

[11] Zhao, R., Xu, L., Wang, Y., Zou, J. (2020). Sign language recognition based on visual attention mechanism. Multimedia Tools and Applications, 79(11-12), 7393-7410.

[12] Borkar, T. R., Badheka, R. (2021). Sign Language Detection and Recognition Using Image Processing and Machine Learning: A Review. In Advances in Intelligent Systems and Computing (pp. 127-137). Springer.

[13] Tran, T. T., Nguyen, T. H., Nguyen, H. T., Nguyen, Q. M. (2021). Sign language recognition using CNN-LSTM neural network and Bayesian optimization. In 2021 IEEE 15th International Conference on Anti-counterfeiting, Security, and Identification (ASID) (pp. 42-47). IEEE.

[14] Anwar, M. T., Awais, M., Shoaib, M. (2021). Sign language recognition using deep learning techniques: a comprehensive review. Multimedia Tools and Applications, 80(16), 24313-24349.

[15] Wang, S., Ma, W., Li, Y., Wang, J., Zhang, J., Ye, Z. (2020). Sign language recognition via deep learning: A survey and future directions. ACM Transactions on Multimedia Computing, Communications, and Applications, 16(2), 1-27.