A Comprehensive Review of Deep Neural Networks for Dynamic Hand Gestures Recognition

Simon Wu
Department of System Design Engineering
University of Waterloo
Waterloo, Canada
s23wu@uwaterloo.ca

Shuao Wu
Department of System Design Engineering
University of Waterloo
Waterloo, Canada
s22wu@uwaterloo.ca

Shan Xue
Department of Electrical and Computer Engineering
University of Waterloo
Waterloo, Canada
s2xue@uwaterloo.ca

Shichen You
Department of System Design Engineering
University of Waterloo
Waterloo, Canada
syou@uwaterloo.ca

***Abstract*—Dynamic hand gestures recognition is an important research field of human-computer interaction, with application ranging from sign language to virtual reality. In this review paper, several techniques for dynamic hand gestures recognition are reviewed, with a special emphasis on convolutional neural networks (CNNs). We delve into different CNN architectures, including 3DCNN, lightweight CNN, fine-tuned CNN, CNN with long short-term memory (LSTM), and CNN with recurrent neural network (RNN). The performance of these various techniques in recognizing dynamic hand gestures is then compared and evaluated. However, we also identify several gaps between the methods mentioned in the literature and their practical applications. These gaps include data availability and quality, computational resources, transferability, and real-time processing.**

***Index Terms*—Dynamic hand gestures recognition, 3DCNN, CNNs, RNN, LSTM, Review**

I. Introduction

Communication between humans can be conveyed through both verbal and nonverbal languages. Nonverbal language being expressed through body language, hand movements and facial expressions. In particular, hand gestures play an important role in nonverbal communication and increasingly being integrated into human life. For instance, hand gestures are more used in sign language to provide communication between people. Also, hand gestures have been applied to human-computer interaction (HCI) systems in recent decades, with a wide range of applications such as video games, remote surgery, and virtual reality. With the development of image processing and visual communications, more and more researchers are dedicating efforts to studying hand gestures recognition. However, there are several challenges in hand gestures recognition such as being sensitive to variations in size, speed and lighting, performing poorly against complex backgrounds, and accurately detecting the gesturing phase [1]. Researchers provide multiple ways to categorize hand gestures. One such way is to classify hand gestures based on observable features or based on the interpretation. Under observable approach, gestures can be classified as either static hand postures or dynamic hand gestures by considering temporal relationships. Static hand postures basically rely on the shape and flex angles of the fingers, while dynamic hand gestures rely on static hand postures and also considers the hand trajectories and orientations. There are numerous techniques for hand gesture recognition. Those techniques can be classified as (i) Hidden Markov Models and other statistical methods, (ii) Artificial Neural Networks (ANNs) and other learning based methods, (iii) Eigenspace based methods, (iv) curve fitting, and (v) dynamic time warping. In this survey paper, we mainly focus on hand gestures recognition in the field of ANNs, with a special emphasis placed on the convolutional neural networks (CNNs). The rest of the survey paper is organized as follows. In the section of related work, we delved into CNNs with different architecture. These CNNs are the extended version of standard CNNs. In this section, we reviewed 3DCNN, lightweight CNN, fine-tuned CNN, CNN with long short-term memory (LSTM), and CNN with recurrent neural network (RNN). In the third section, we compared and evaluated the different CNNs from the literature which apply to hand gestures recognition techniques. In the final section, we represented the conclusion by summarizing the key findings and insights that have been gained throughout our survey.

II. Related Work

A. 3DCNN

Köpüklü et al. [2] proposed a system for real-time hand gesture detection consisting of a detector and a classifier. The detector's primary function is to activate the classifier when it detects a gesture. It is designed to discriminate between gesture and no gesture classes. To increase robustness, the detector runs on fewer frames than the classifier, and it is trained using a weighted cross-entropy loss to reduce the possibility of false positives. Any design that performs well for classification can be utilized as the classifier; in this work, two 3D CNN architectures, C3D and ResNext-101, were employed. To avoid overfitting, the models were pretrained on the Jester dataset and fine-tuned on the EgoGesture and nvGesture datasets.

Al-Hammadi et al. [3] [4] proposed a 3DCNN for real-time hand gesture classification. They used linear sampling to convert the input video to RGB frames and designed a 6-block architecture with varying kernel sizes. The 3DCNN utilized ReLU activation function and SoftMax layers. To globalize the regional properties learned by the 3DCNN model, the authors created three different fusion strategies: multilayer perceptron (MLP) neural network, long short-term memory (LSTM) network, and stacked autoencoder, which were evaluated for their efficacy comparing with the previous 3DCNNs.

Liu et al. [5] developed a dynamic recognition model, CBAM-C3D, using a 3D convolutional neural network and a convolutional block attention module for both RGB and depth images. The attention mechanism, CBAM, aids in focusing on relevant elements in the input data, reducing overfitting, and increasing accuracy [6] [7]. 3D-CBAM extends the CBAM technique to 3D data by adding attention modules to each convolutional block, including a spatial and channel attention module [12]. CBAM-C3D contains two 3D-CBAM layers followed by max-pooling layers, with three of these combinations, and three fully connected layers. To overcome inconsistent action duration challenges, an optimized inter-frame difference method is used to simplify data processing and extract key frames from a representative set of images obtained by unifying the scale of video data.

B. Lightweigh CNN

Yang et al. [8] proposed a lightweight deep neural network (MDHandNet) for hand gesture and sign language recognition using micro-Doppler images. Micro-Doppler imaging is a technique that uses radar to capture the Doppler signature of a moving object, such as a human hand, and extract unique features related to hand motion. The authors leverage micro-Doppler imaging to capture the movement of the hand and develop a deep learning model that can recognize hand gestures and sign language. The network comprises three parts: backbone subnet, hierarchical feature processing subnet, and output subnet. The backbone subnet extracts features from various size receptive fields to gather maximum feature information. The hierarchical feature processing subnet further processes the features to meet high-performance requirements for classification tasks. Finally, the output subnet combines the features and maps them into category probabilities to classify HG/SL language actions.

C. Fine-tuned CNN

Sahoo et al. [9] used pre-trained CNN models, AlexNet and VGG-16, this work aims to classify hand-gesture images. AlexNet consists of five convolution layers, three pooling layers, and three fully connected layers with ReLU activation function and a dropout layer. VGG-16, with more convolutional layers, is used for recognition of static hand gestures. Fine-tuning of pre-trained CNN is employed by transferring weights to a network learning on target dataset and updating weights and biases after each iteration. The last fully connected layer of the pre-trained AlexNet is changed to 34 nodes for the HUST-ASL dataset.

D. CNN + LSTM

Zhang et al. [10] proposed a Short-Term Sampling Neural Network (STSNNs) which aims to recognize dynamic hand gestures using a sliding window technique combined with neural networks. Hand gesture recognition requires capturing both spatial and temporal information from video inputs. This approach divides consecutive frames into groups and takes a random sample from each group, ensuring a fixed number of samples for each video sample regardless of its duration. Representative samples are created by combining RGB frames and optical flow frames, and a ConvNet is used to capture features from each sample. All ConvNets share the same parameters to reduce training parameters. An LSTM module is used to learn long-range temporal features from the feature sequence, and the SoftMax function is used to calculate predicted class probabilities.

E. CNN + RNN

Lai and Yanushkevich [11] proposed an approach for recognizing hand gestures involving using both depth and skeleton points and consists of two main components: a depth-based CNN+RNN and a skeleton-based RNN. The system extracts temporal patterns through RNN and classification using an MLP. The fusion of both networks is expected to yield higher performance by selecting the best features from both the skeleton and depth-based spatial information. The three main fusion techniques considered are feature-level fusion, score-level fusion, and decision-level fusion. The feature-level fusion can be performed at any layer before the MLP, while score-level fusion is performed after or between the fully connected and softmax layers. The decision-level fusion is not examined in this paper. Concatenation, averaging, and maximum are the types of fusion techniques considered, with concatenation performed at the feature-level and averaging and maximum applied to the score-level fusion.

III. Technology Comparison

3DCNN is one of the major techniques in dynamic hand gesture recognition because it leverages both the spatial content and the temporal relationships between frames, it also can be combined with other techniques to build a more powerful model to recognize hand gestures in real-time. The 3DCNN model proposed by Köpüklü et al. [2] consists of a detector before the classifier. The detector can distinguish between gesture and non-gesture from the sequence of input images; thus, the classifier will be activated only when the gestures are detected, which is an efficient way to decrease the computational complexity. Moreover, to avoid the misclassification caused by the hand getting out of view, the authors suggest a method where the previous detector predictions' raw SoftMax probabilities are stored in a queue and filtered to generate the final detector decisions. Since dynamic hand gesture recognition should consider misclassification, no detection, and multi-detection in one gesture, they use Levenshtein accuracy, which is more comprehensive, for the model measurement. Since the classification performance is the most relevant part of their model, they chose C3D and ResNet-101 as the classifier, which leads to a high computational complexity of their model.

Al-Hammadi et al. [3] [4] also use 3DCNN for real-time hand gesture detection, but with fusion strategies to extract features. To ensure the performance of their model, Al-Hamadi et al. used three different datasets and pre-processed these training and testing data to reduce the computation cost and training convergence of the model. Also, they used PCA to reduce the learned features that can minimize the number of variables but maximize the information about the distribution of original data. Thus, the model becomes less computational complexity and maintains a great performance. Also, there are two different data modes: signer-dependent and signer-independent. The model has a great performance in both modes means that the information from signers except the gesture does not affect the model's performance.

Liu et al. [5] also do some effort on reducing the input data to reduce the cost of computation by designing a method to extract the keyframe based on inter-frame differences. To improve the performance, they also enhance the feature extraction on spatial and time dimensions. However, compared with the previous 3DCNN model, this CBAM-C3D model's classification accuracy is not high enough. Also, as the authors mentioned, their model has a huge amount of parameters and slow network prediction, and poor performance in real-time hand gesture recognition.

Most of the 3DCNN model has a problem with the model size and computational complexity, Yang et al. [8] focused on the lightweight DNN and proposed MDHandNet with fewer parameters and lower computational complexity but maintain the classification accuracy in dynamic hand gesture recognition. Different from most vision-based hand gesture recognition, Yang et al. use micro-Doppler images so that the performance of the model isn't affected by the light in the environment. It is an advantage of their model, but also a disadvantage of their model that using this model needs a radar sensor to capture the micro-Doppler images, which is not a general method for dynamic hand gesture recognition. To increase the accuracy of the recognition, fine-tuned CNN is also a useful method.

Sahoo et al. [9] fine-tuned two pre-trained CNNs: AlexNet and VGG-16, so they did not spend time on the training model. Then, they combined two score vectors as one by using the fusion method. This model outperforms CNN, FFCN, AlexNet with SVM, and VGG-16 with SVM in static hand gesture recognition, but it still needs to improve its performance in dynamic hand gesture recognition in the future.

Zhang, Wang, and Lan [10] proposed STSNN which is a CNN with LSTM to capture the long-term information for the input data. They are concerned about the cost of computing and training in their work. To reduce the computing cost, they sampled the frames from videos. They also reduce the training cost and improve the model performance by using transfer learning. The chosen pre-trained model is Inception V2, which has a good balance of efficiency and accuracy. Another novel point in their work is that they zoom out the original video to make a new dataset. It helps confirm the robustness of a trained model and offers additional training data samples. Moreover, since different sampling representatives from various groups can provide a wide range of input combinations, random sampling in each group and stacking optical flow frames atop RGB frames may also produce data variety. Although their model has a good performance in hand gesture recognition from video, one of the problems is that each input video contains only one gesture. The model uses trimmed video as input data and has a great performance does not mean the model can still have satisfiable performance with untrimmed video.

Lai and Yanushkevich [11] applied CNN and RNN in dynamic hand gesture recognition. In their work, they use both depth and skeleton information, which can provide more robust and accurate gesture recognition compared to using only one type of information. RNN is designed to work with sequential data that is suitable for dynamic hand gesture recognition since the input data is video. However, RNNs can be computationally expensive to train and can require significant amounts of memory to store the network weights and activations. Also, vanishing and exploding gradient are a major problem of RNN. This can make it difficult to train RNNs over long sequences and get worse when applying this model in real-time hand gesture recognition. Moreover, the proposed approach uses a relatively small dataset (DHG-14/28) for training and testing, which may limit the generalizability of the results and may not fully represent the complexity and diversity of real-world scenarios.

IV. Gap & Future

A. Gap between current study and practical application

So far, we have reviewed a variety of technologies related to gesture recognition, including 3DCNN, CNNs with different architectures and adding auxiliary equipment and other methods. They either provide a new training model structure or a more efficient gesture representation method. Compared with earlier methods, their results have corresponding improvements in different aspects. Nevertheless, we still noticed some gaps between the methods mentioned in these papers and practical applications.

1) Data availability and quality: First and foremost is the availability and quality of data, which is generally lacking in the papers we reviewed. In these papers, quite a lot of experimental data is recorded by less than 5 people repeatedly making gestures, which will naturally lead to bias in dataset.

At the same time, we believe that the duplicate data generated by a single data source will interfere with the validity of the test dataset. The researchers did not describe in the paper how they determined to avoid the overfitting problem.

2) Computational resources: Another challenge of deploying some methods for practical problems is the requirement for large-scale computational resources, especially for deep learning models such as 3d-CNN and combination of CNN and RNN that have a large number of parameters. Training and testing these models can be computationally expensive and time-consuming, which can limit their practical applicability. This is quite clear from the heat maps of reviewed papers, most of high accuracy results come from the structure with more neurons. Although we have reviewed a paper on the lightweight CNN model before, the experiment in that paper is based on Micro Doppler format data, which provides one more dimensional feature than traditional RGB image frames. And may require additional equipment.

3) Transferability: The above discussion has also led us to consider transferability. As we have already known, the lightweight CNN model based on Micro-doppler image compared with the general CNN or RNN model has lower computing power requirements, but it may not generalize well to other datasets. Even for those methods that operate on RGB format data, due to the nature of deep-learning models that trained on one dataset may not have the same performance on the other dataset. Not to mention that for sign language, expressions are different in different languages, for example, the papers we reviewed included gesture expressions in English, Chinese and Arabic, which further increases the difficulty of model transferability.

4) Real-time processing: Finally, the discussion of real-time processing is missing from the papers we reviewed. Real-time processing is essential for many practical applications, such as robotics, autonomous vehicles, or human-computer interaction. However, deep learning models can be computationally intensive and may not be suitable for real-time processing especially on low-power devices.

B. Future works

Based on the above discussion about gaps and the potential of Deep learning in the field of gesture recognition, we think that the following aspects can be improved in the future to improve their performance and applicability.

1) Attention mechanisms: According to our study, we think that attention mechanisms can still have a lot of room for improvement in this field. In the case of a limited dataset, introducing attention mechanisms into the model can help the neural network selectively focus on the most relevant parts of the input data. It can help the model focus on the most relevant parts of the input data and ignore the irrelevant or noisy parts. In our review of related methods combining CNNs and RNNs, we found similar structures and felt that this should be more widely used.

2) Integration with RNNs: As mentioned earlier, the combination of CNNs and RNNs has shown promising results in action recognition tasks. This approach could also be applied to other gesture recognition models, where the temporal dependencies between gestures could be modeled using an RNN, while the spatial features could be extracted using a CNN. This method can significantly improve the accuracy of the results and make the structure of the model more meaningful.

3) Incorporation of depth information: In this study, we learned about RGBD format as well as micro-Doppler format data. Compared with the traditional RGB video frame, the data in those formats provides an extra depth of information. This depth information can reduce the influence of light, skin color, etc. on the general model. By incorporating this depth information into CNN models, it may be possible to improve the accuracy and robustness of gesture recognition systems, especially in challenging lighting conditions.

4) Real-time gesture recognition: Real-time gesture recognition is essential for many applications. Therefore, developing lightweight and efficient CNN models that can perform real-time gesture recognition on low-power devices could be an important future trend.

C. Conclusion

Overall, while these technologies have shown great potential for practical applications, there are still challenges that need to be addressed, such as data availability and quality, computational resources and transferability. Addressing these challenges will be critical for the successful adoption of these technologies in real-world settings. The future trend of Deep learning in the field of gesture recognition could focus on improving accuracy, robustness, and efficiency by integrating various techniques.

References

[1] S. Ji, W. Xu, M. Yang, and K. Yu, "3d convolutional neural networks for human action recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 1, pp. 221–231, 2013.

[2] O. Köpüklü, A. Gunduz, N. Kose, and G. Rigoll, "Real-time hand gesture detection and classification using convolutional neural networks," in *2019 14th IEEE International Conference on Automatic Face Gesture Recognition (FG 2019)*, 2019, pp. 1–8.

[3] M. Al-Hammadi, G. Muhammad, W. Abdul, M. Alsulaiman, M. A. Bencherif, and M. A. Mekhtiche, "Hand gesture recognition for sign language using 3dcnn," *IEEE Access*, vol. 8, pp. 79 491–79 509, 2020.

[4] M. Al-Hammadi, G. Muhammad, W. Abdul, M. Alsulaiman, M. A. Bencherif, T. S. Alrayes, H. Mathkour, and M. A. Mekhtiche, "Deep learning-based approach for sign language gesture recognition with efficient hand gesture representation," *IEEE Access*, vol. 8, pp. 192 527–192 542, 2020.

[5] Y. Liu, D. Jiang, H. Duan, Y. Sun, G. Li, B. Tao, J. Yun, Y. Liu, and B. Chen, "Dynamic gesture recognition algorithm based on 3d convolutional neural network," *Computational intelligence and neuroscience*, vol. 2021, p. 4828102, 2021.

[6] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[7] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri, "A closer look at spatiotemporal convolutions for action recognition," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6450–6459.

[8] Y. Yang, J. Li, B. Li, and Y. Zhang, "Mdhandnet: a lightweight deep neural network for hand gesture/sign language recognition based on micro-doppler images," vol. 25, 2022.

[9] J. P. Sahoo, A. J. Prakash, P. Pławiak, and S. Samantray, "Real-time hand gesture recognition using fine-tuned convolutional neural network," *Sensors*, vol. 22, no. 3, p. 706, Jan 2022. [Online]. Available: http://dx.doi.org/10.3390/s22030706

[10] W. Zhang, J. Wang, and F. Lan, "Dynamic hand gesture recognition based on short-term sampling neural networks," *IEEE/CAA Journal of Automatica Sinica*, vol. 8, no. 1, pp. 110–120, 2021.

[11] K. Lai and S. N. Yanushkevich, "Cnn+rnn depth and skeleton based dynamic hand gesture recognition," in *2018 24th International Conference on Pattern Recognition (ICPR)*, 2018, pp. 3451–3456.