 doi:10.20944/preprints202206.0368.v1

Article

A 3DCNN-LSTM multi-class temporal segmentation for hand gesture recognition

Letizia Gionfrida[1,2,*], Wan M. R. Rusli[1], Angela E. Kedgley[1], and Anil A. Bharath[1]

[1] Department of Bioengineering, Imperial College London, London, SW7 2AZ, UK
[2] John A. Paulson School of Engineering and Applied Sciences, Harvard University, Boston, MA, USA.
* Correspondence: gionfrida@seas.harvard.edu

Abstract: This paper introduces a multi-class hand gesture recognition model developed to identify a set of defined hand gesture sequences in two-dimensional RGB video recordings. The work presents an action detection classifier that looks at both appearance and spatiotemporal parameters of consecutive frames. The classifier utilizes a convolutional-based network combined with a long-short-term memory unit. To leverage the need for a large-scale dataset, the model uses an available dataset to then adopt a technique known as transfer learning to fine-tune the model on the hand gestures of relevance. Validation curves performed over a batch size of 64 indicate an accuracy of 93.95% (± 0.37) with a mean Jaccard index of 0.812 (± 0.105) for 22 participants. The presented model illustrates the possibility of training a model with a small set of data (113,410 fully labelled frames). The proposed pipeline embraces a small-sized architecture that could facilitate its adoption.

Keywords: hand gesture classification; transfer learning; three-dimensional convolutional; LSTM network

1. Introduction

Hand gestures are a critically important form of non-verbal communication. The interpretation of hand gestures with wearable sensors [1], [2], or cameras [3], [4] aims to transform the hand gestures into meaningful instructions; this interaction is also known as hand gesture recognition. The field of hand gesture recognition has seen significant improvements over the past few years [5] and, most recently, bundled with the latest advancements in computer vision, has encouraged the development of new technologies to support rehabilitation [6], [7], robot control, and home automation [8].

Computer vision techniques rely on convolutional neural networks (CNNs) to extract two-dimensional (appearance-based) and three-dimensional (motion-based) array features. CNNs are generally used in image recognition to process pixel data. They take raw pixel data as input, train the designed architecture, and automatically extract features. These models have been divided into static (two-dimensional) and dynamic (three-dimensional) based on the model's output features. Several investigations [9]–[11] have implemented two-dimensional static appearance-based hand gesture recognition models (also known as two-dimensional CNN models) intending to develop a computationally inexpensive classifier to extract stable shapes of the human hand. However, these models do not consider the spatio-temporal parameters that occur from sequential frames of a video recording, and appearance alone cannot accurately identify the gesture signature [12]. Therefore, new approaches, known as three-dimensional dynamic hand gesture recognition, have emerged to fill this gap.

Three-dimensional dynamic hand gesture recognition models also rely on CNNs, act like conventional two-dimensional CNNs, and have spatial-temporal filters. Since their introduction in 2015 [13], these models have been primarily embraced for hand gesture recognition [13]–[15], presenting excellent characteristics in recognising hand actions from both appearance and spatio-temporal features. However, they require more parameters

 doi:10.20944/preprints202206.0368.v1

than two-dimensional CNNs, meaning vast datasets are needed, and making them more challenging to train [16]. Furthermore, these approaches have additional drawbacks that include cost, the logistical challenges of dealing with complex and lengthy datasets, and the requisite quality of captured images needed for appropriate training. To overcome these drawbacks, previous research has leveraged a technique known as transfer learning [17].

Transfer learning is a methodology where architecture is implemented and trained on a generic activity and is then adopted for a specific different but linked activity (**Error! Reference source not found.**). This technique is often employed to tackle the issue of a deficiency of training data [18], [19]. The usual objective of transfer learning techniques is to learn visual features from the initial assignment [19]. This technique can train and acquire a forthcoming linked task from fewer data samples. Transfer learning is adopted when a novel, minor dataset is smaller than the dataset used to train the pre-trained architecture.

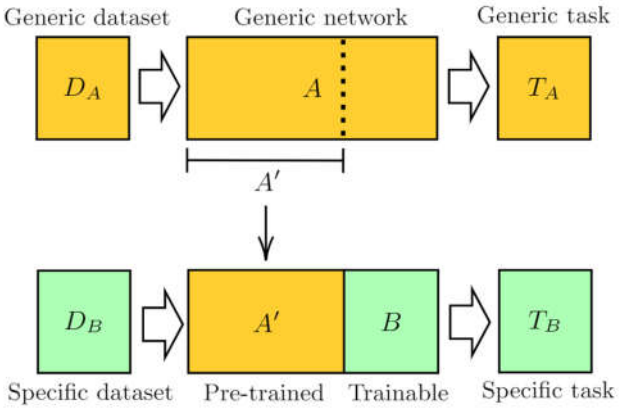

Figure 1. Schematic of the approach to transfer learning, whereby architecture implemented and trained on one activity and is adopted for a different but linked activity.

Another hurdle in dynamic gesture recognition for three-dimensional CNNs is recognising specific actions when dealing with continuous video streams [21]. Identifying human activities within video sequences is difficult because of the vast irregularity of hand actions on a time scale, unclear frame quantity, distribution, and limits of gesture signatures [22]–[24]. Furthermore, hand motions are often intricate and articulated and, when performed in an uncontrolled environment, can lead to occlusion that can limit the tracking. However, the ability to track and segment hand gestures in the real world can answer the need of applying these models to more realistic and generalisable tasks.

Manual segmentation of continuous video recordings is considered the most adopted technique when training hand gesture recognition [25]. However, the process is lengthy, and often a large proportion of frames is left unlabelled, causing indexing issues in the training of novel classification methods. The ability to automatically detect action in video recordings has an essential function for different applications that require end-to-end process automation. But, while much work has been produced on increasing the accuracy of hand gesture recognition models and enhancing the strength of these approaches [3], [5], [26], just a few attempts have been presented for temporal segmentation [27], [28].

Attempts at temporal segmentation have focused on motion trajectory [29] and skeletal tracking [30] from depth cameras. However, these systems were sensitive to the backgrounds and lighting conditions. A different approach, presented by Camgoz et al., suggested windowing the continuous video stream for segmentation [31]. However, the length of the sliding volume was fixed, often cutting part of the critical features of the gestures. Moreover, appearance and hand motion information complement a temporal segmentation classifier [28]. Still, Camgoz et al. also used only time-series data detected

 doi:10.20944/preprints202206.0368.v1

from hand motion, with no appearance information [31]. In contrast, Wang presented a segmentation method that contained both action and appearance-based information, and used both RGB and depth capture modalities [28].

Increasingly, enormous datasets of human movement are publicly available, as researchers seek to pool resources and work more openly. The 20BN Jester is a state-of-the-art dataset and the largest of human hand gestures collected from monocular RGB cameras. It contains a total of 148,092 videos corresponding to 5,331,312 frames [15]. Each video is, on average, three seconds, and the dataset contains a total of 27 classes.

This aim of this paper is to present the training of a CNN using a small set of data and the development of a narrow architecture that can run efficiently for continuous hand gesture recognition. The key objectives of this paper include:

1) To implement and to test the accuracy of a three-dimensional CNN model combined with a long-shot term memory (LSTM) unit to reliably classify and segment continuous video recordings and improve current manual-based segmentation when deploying models capable of executing tasks smoothly in real-world scenarios.

2) To evaluate the performance of transfer learning in implementing an architecture that is trained on a larger scale dataset, and then fine-tuned with a small-scale dataset.

3) To lay the foundations for a small-scale and reliable model, paving the way to a broader and optimised application that can be used to automatically detect where to run the keypoint hand tracking network.

2. Materials and Methods

Experimental set-up

Twentytwo volunteers (twelve female, ten male) participated in this experiment. All the participants were healthy, presenting with no hand pathology, no loss in mobility, and no experience of upper limb joint surgery or fracture in the six months preceding the data collection. All participants were informed, both verbally and in writing, of their right to withdraw from the study at any time. Written informed consent was obtained from each participant. The protocol was approved by the Imperial College Research Ethics Committee (ICREC). The entire pipeline adopted in the study is illustrated in Figure 2.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 June 2022 doi:10.20944/preprints202206.0368.v1

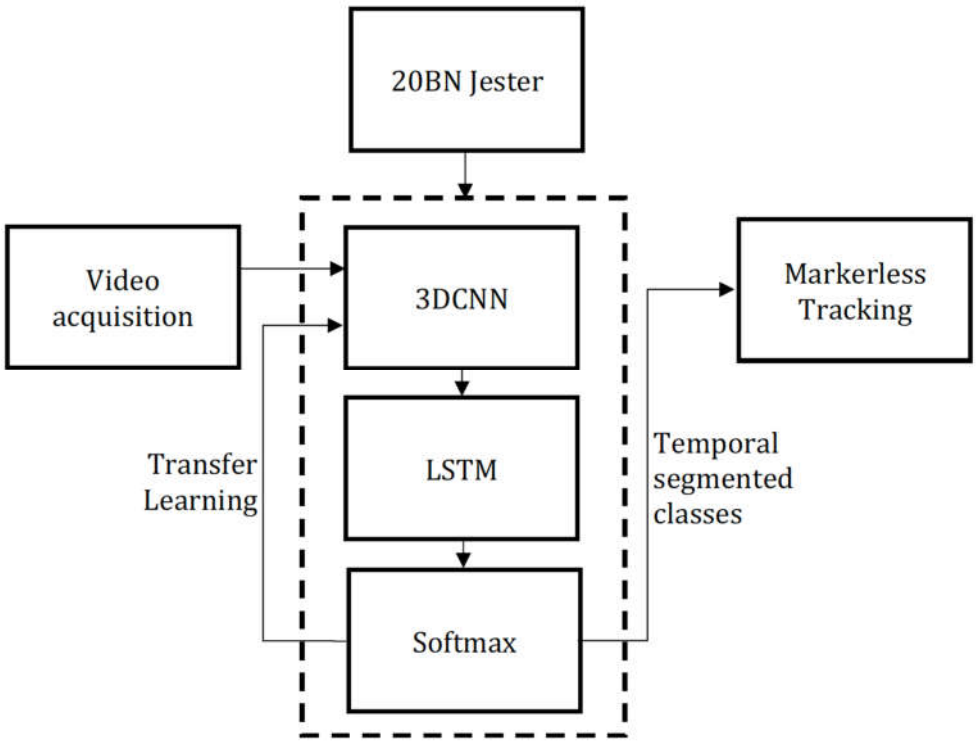

Figure 2. Flowchart of the experimental setup for the hand gesture recognition investigation. The pipeline uses transfer learning, pre-training the architecture on the 20BN Jester dataset [15], a three-dimensional convolutional neural network (3DCNN), a long short term memory (LSTM) unit and the output function (Softmax).

Data collection

Participants were asked to record one video sequence during online video meetings. There was a timed PowerPoint to make the video acquisition consistent, to support participants on the activities to be performed during the recordings, and to inform participants on the way to position themselves relative to the device for the recordings.

To perform the hand gestures, participants were asked to use a standard device camera to capture the required hand exercises using any laptop, smartphone, desktop computer. A standard camera was defined as a camera developed from 2012 onwards that was able to capture video recordings at a rate of thirty frames per second. To assess if the data were captured from an acceptable browser and operating system, participants were asked to check that the specifications of the recording system.

The hand activities performed by participants in this part of the investigation included abduction and adduction, metacarpophalangeal joint flexion, and thumb opposition. Each was performed four times with both the left and right hands. During these exercises, participants were asked to hold static poses for five seconds. Four classes of gestures were defined based on the trials (**Error! Reference source not found.**).

Figure 3. Illustration showing hand gestures classified during each trial: no gesture, abduction and adduction (Abd and add), metacarpophalangeal (MCP) flexion and thumb opposition.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 June 2022 doi:10.20944/preprints202206.0368.v1

The hand gesture sequences were captured from continuous video recordings of 250 seconds. The continuous video sequences were then manually segmented and labelled. Examples representing the data collected from twelve representative participants are illustrated in Figure 4.

Figure 4. Examples of anonymized frames of the videos from twelve representative participants. The images show the variance in the people's appearance and background scenes.

In addition to the captured data, the 20BN Jester dataset acquired by Materzynska et al. [15] was used. The classes of interest in this study, "no gesture", "abduction and adduction", "MCP flexion", and "thumb opposition", were not present in the Jester dataset. Therefore, out of the 27 classess of the 20BN Jester dataset, five hand activities were considered. These hand tasks of the 20BN Jester were count-to-five, swiping down and left, thumb-up, and thumb-down. These activities were selected to include different image frames of isolated digits and the palm with all the digits for both the left and right hands.

Pre-processing

The captured frames were normailsed to ensure that each input to the three-dimensional CNN had the same distribution, and each class had the same number of frames. This was particularly important as, although the timing of the participants' actions was marked by the PowerPoint presentation, individuals could execute a hand gesture at different speeds. Ideally, a three-dimensional CNN input should always be balanced, making the model converge faster. If the input frames were not normalised, the weights could have had different calibrations across features, making the cost function converge ineffectively.

The frame length was set to be equal for all the acquisitions for which the hand gestures were at the centre of the video [9], [33]. Following the structure of the 20BN Jester dataset, the normalisation was applied to impose a fixed length, set to be 32 frames. If the number of frames was higher or lower, a down-sampling or a padding function was applied, respectively, to generate fixed-length videos. Given the S_n sequence of RGB frames, the L_S length of the sequence, and the L_F fixed length, the padding and down sampling techniques were defined as:

$$S_n = \begin{cases} padding(S_n), & L_s < L_F \\ (S_n), & L_s = L_F \\ downsampling(S_n), & L_s > L_F \end{cases} \quad (5.1)$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 June 2022 doi:10.20944/preprints202206.0368.v1

Following normalisation, the images were resampled to be 64 × 64 pixels to expedite classification. The labels were assigned manually, and the videos were manually trimmed for input into the segmentation classifier. Finally, for training and validation, the dataset was split into training, validation and testing sets, with a 70:20:10 ratio.

Of the data from the video collected, a total of 2,812 short video sequences of healthy volunteers performing three different hand activities were used for testing and validation, including 1,968 (≈ 70% of the dataset) were used for training and 845 (≈ 30% of the dataset) were used for validation and testing. Each short video sequence contained 32 frames, for 89,984 frames in total. A total of 5,155 short video sequences were collected, of which 3,609 (≈ 70% of the dataset) were used for training and 1,546 (≈ 30% of the dataset) were used for validation and testing. Each short video sequence contained 32 frames for a total of 113,410 frames for training and 6,784 for validation.

Postprocessing

After the data pre-processing, the architecture was implemented based on an existing model originally introduced by Tran et al. [34], known as C3D. Specifically, a modified version of the C3D network, similar to the multimodal RGB-D-based network by Hakim et al. [12], was considered. Furthermore, to make sure that the three-dimensional CNN model was able to learn longer sequences, another unit, able to acquire long-term temporal features, was combined with the three-dimensional CNN, an LSTM unit. The final architecture (Figure 5) consisted of a three-dimensional CNN layer with three convolutional layers, a Rectified Linear Unit (ReLU) as activation function in the hidden layers used to avoid vanishing gradient, one LSTM layer, a flatten layer, a fully connected dense layer and an activation function, also known as the Softmax layer.

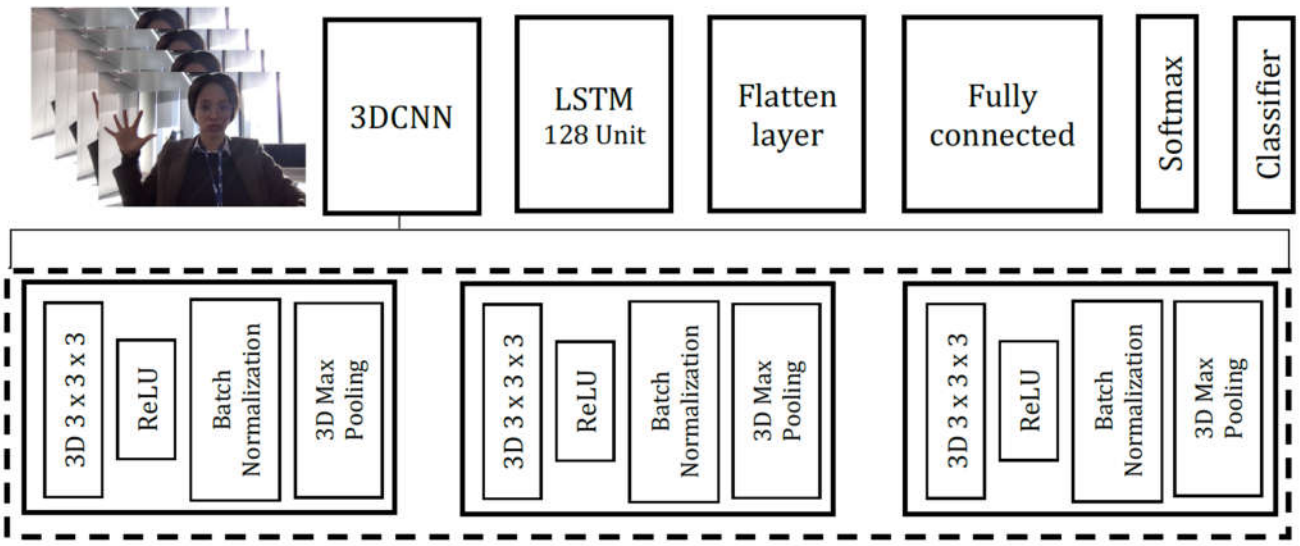

Figure 5. Three-dimensional convolutional neural network (3DCNN) with long short-term memory (LSTM) for dynamic hand gesture recognition. The video sequence is fed into the 3DCNN to operate 1D & 3D convolutions for time & space dimensions. The hidden layers (dashed box) with Rectified Linear Units (ReLU) as activation function the 3DCNN has limitations to learning long-term information and therefore, the vector goes into an LSTM. The tensor in output is then flattened into a single dimension, passed into a fully connected layer and finally, the activation function (Softmax) predicts the classes.

The multi-dimensional input tensors were flattened into a single dimension. The flattened layer is often employed in the presence of multi-dimensional output. This layer aims to produce a linear output that can be conveyed onto a dense layer. A dense layer (also called fully connected) joined every input neuron to every output neuron in the preceding layer. Finally, the Softmax function produced a vector that denoted the list of probability

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 June 2022 doi:10.20944/preprints202206.0368.v1

classes of possible results. Based on the output from the Softmax, the frames were then segmented into those where the activities occurred and those where there was no gesture. The class "no gesture" was provided in case no activity was performed, but also for frames without a hand, when participants placed the hand down following a performed activity.

The baseline model was pre-trained on the selected five classes of the 20BN Jester dataset. Starting from the pre-trained architecture, a technique known as transfer-learning [18] was then used to fine-tune the model to the activities performed in this study. The technique took the parameters from the previously trained model, froze the last layers to avoid the weights in the last (frozen) layers being updated, and then new trainable layers were added, together with new data to fine-tune the model.

A total of four tests was performed. During the first two tests, transfer learning was used with three convolutional layers. Then, to increase performance, an additional convolutional layer and an increased sample size were considered. The first two tests were evaluated over mini-batches of 13 epochs, following the segmentation classifier proposed by Wang [28]. The last two tests were evaluated over a batch size of 64 epochs, a training batch size also presented in Wang's investigation [28]. A 12GB NVIDIA Tesla K80 graphics processing unit provided by Google Colaboratory was used for training of the 20BN Jester dataset for the baseline model, TensorFlow [35] was used to deploy the model, and the training took approximately nine and a half hours. For the first and the second tests, the training times were respectively one and a half hours and two and a half hours, while for the last two tests, they were two and four hours.

The metric used to evaluate the performances of the model was the Jaccard index or intersection over union value [36]. The index is often used for segmentation classifiers and was computed to analogise a set of predicted labels with a set of the corresponding true labels. Letting A and B be the set of frames predicted and ground truth manually labelled, respectively, the index is defined as:

$$JACCARD = \frac{|A \cap B|}{|A \cup B|} \tag{5.2}$$

The Jaccard index varies from zero to one, the larger is the index, the higher is the accuracy of the temporal segmentation classifier. Training and validation accuracies were testeds for 13 and 64 epochs for a small sub-portion of 12 participants and for 22 participants to evaluate how variaty in population sizes can improve training and validation perfromances.

3. Results

Training and validation accuracies for 13 and 64 epochs for 12 and 22 participants show limited levels of accuracy (below 70%) reached for 13 epochs and increased level of accuracy (93.95%) reached for 64 epochs (Figure 6). In the training and validation curves illustrated for 64 epochs, the training performed on 22 participants outperforms the training on 12 participants. Overfitting was observed during training after 50 epochs, in both cases (12 and 22 participants), suggesting that additional training would not result in the model having improved learning.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 June 2022 doi:10.20944/preprints202206.0368.v1

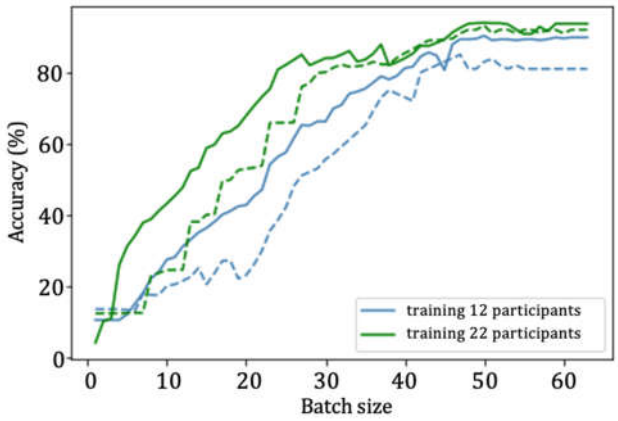

Figure 6. Results of the training (solid line) and validation (dotted line) for a training batch (batch size) of 64 epochs for 12 and 22 participants.

A representative output from the Softmax function (Error! Reference source not found.) of the temporal segmentation for a continuous video recording for the three-dimensional CNN hand gesture classifier trained for 64 epochs and 22 participants illustrates the agreement with manual segmentation (ground truth).

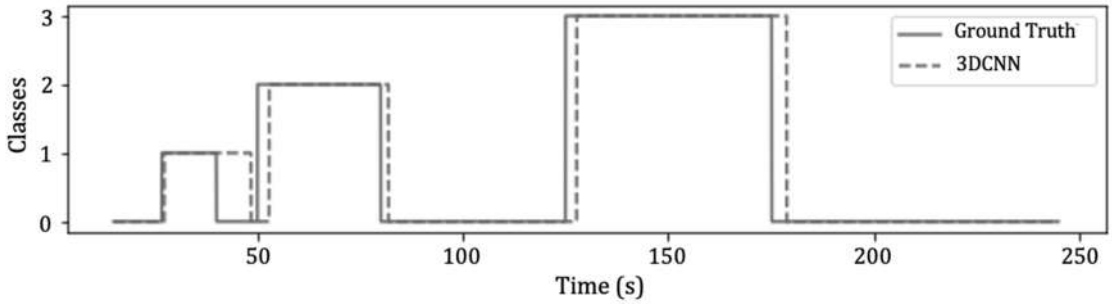

Figure 7. An example of the temporal segmentation and classification in output from Softmax function of the three-dimensional convolutional neural network for 64 epochs and 22 participants (dashed lines) compared against the ground truth manually segmented for Participant 1 for the labels "no gesture" (class=0), "abduction and adduction" (class=1), "metacarpophalangeal (MCP) flexion" (class=2), and "thumb opposition"(class=3).

The training runs, executed for batch-sized 64, had an initial mean Jaccard index that reached 0.794 (± 0.44), increasing to 0.812 (± 0.105) for the enlarged sample size of 22 participants (Table 1).

Table 1. Comparison of the three-dimensional convolutional neural network for 12 and 22 participants using the mean Jaccard index $\bar{J}_s$ and the accuracy percentage (%).

Dataset	Number of frames	Mean Jaccard Index $\bar{J}_s$	Accuracy (%)
12 participants	89,984	0.794	83%
22 participants	113,410	0.812	93.95%

The validation accuracy was 83% (± 0.05), increasing to an accuracy level of 93.95% (± 0.37) when additional participants were included. The "no gesture" label agreed with the manually segmented ground truth 96.47% of the time for all participants. The "abduction and adduction" class agreed with the ground truth 92.5% of the time for all participants. The "MCP flexion" label agreed with the manually obtained labels 95.7% of the time for all participants. Finally, the "thumb opposition" class was in agreement with the ground truth 90.93% of the time for all participants.

 doi:10.20944/preprints202206.0368.v1

4. Discussion

This work illustrates a CNN that automatically classifies and segments videos containing specific hand exercises including no gesture, abduction and adduction, MCP flexion, and thumb opposition. The segmentation of continuous video recordings was based upon a classifier that identified when the label "no gesture" was present. The presented pipeline addressed the challenge of hand gesture recognition from long video sequences captured using a monocular RGB camera.

The implementation of the three-dimensional CNN was based on a model known as C3D, proposed by Tran et al. [34] and made of an high-resolution and a low-resolution sub-architecture, both trained individually. Even if the C3D model presented good performance, the cost of training two different models is high, so a modified version, which incorporated the two networks into one, was used in this work. This modified C3D, however, could only detect short temporal characteristics from short video sequences, while the aim of this work was to introduce a network that detects short-term temporal features from long video sequences. Therefore, the final CNN was combined with an LSTM unit, capable of learning the long-term dependencies in long video sequences.

Previous studies that combined three-dimensional CNN with LSTM units for hand activity recognition used both RGB and depth modalities to extract the motion signature [12], [28], while the three-dimensional architecture implemented in this work was only based on an RGB sequence, showing a similar level of accuracy (93.95%) can be reached also from a single acquisition modality. Furthermore, the proposed network outperformed the 82% accuracy presented by Hakim et al. [12]. The overfitting observed after 64 epochs was similar to that of other investigations that used dual modalities [25, 26]. The use of transfer learning to reach an acceptable (above 80%) level of accuracy enables the possibility of scaling this approach to include different hand gesture activities, showing how the model can be trained effectively on a small dataset to create an effective small-size segmentation classifier.

The mean Jaccard recorded in Wang's study was 0.6127 for the RGB modality [28], while in this investigation the mean Jaccard reached 0.794, outperforming the value presented in Wang's investigations. However, Wang's accuracy was based on the Montalbano Gesture Dataset [37], containing different hand activities from those implemented in this investigation. Therefore, further investigations would be needed to compare the performances of this network using this metric. Furthermore, no inconsistency was shown across the segmented video recordings for action and participants, meaning that segmentation accuracy was not based on specific actions or on specific participants.

To adopt and scale this application in real-work scenarios, if multiple classes are considered, future directions could include testing this approach for real-time application using a finite state machine system that can decrease the classes under inspection and increase the accuracy for real-time application. To further improve the model's performance for real-time applications, the input image size or the number of layers could be increased. On top of the 20BN Jester dataset, an additional dataset could be used to enhance the model's performance. The Jester dataset was developed by actors and did not provide numerous occlusion cases. Regardless, in realistic circumstances, occlusion exists. Recordings captured in unconstrained scenarios may incorporate additional types of interference, such as blurry hand gestures if the participants or the camera moves suddenly during the acquisition. Rescuing identifiable cues of image interference for a real-time hand recognition model would be an attractive research direction. Furthermore, while the supervised-based transfer learning produced expected outcomes, the approach presented in this work could be transported to unsupervised learning and could support the automated labelling and segmentation of long video recordings, increasing the models' generalizability.

Adapting current gesture recognition techniques to the specific mobility exercises would have benefits that go beyond this single application. A real-time device that requires minimal manual processing could process and identify multiple gestures as soon

as an image frame is received. This approach could be deployed into online hand gesture recognition studies for advanced assistance systems, surveillance, aided robotics, and clinical applications. For instance, the pipeline illustrated here could be integrated into remote monitoring clinical solutions, presenting the training of a model that uses a smaller dataset implemented on a small architecture that can run efficiently to solve the classification problem for hand temporal segmentation. This would pave the way to a broader application in hand tracking models, incorporating other hand activities categories and obtaining a more generalizable approach, that would include different hand exercise programmes and different hand conditions.

Author Contributions: The authors confirm contribution to the paper as follows: study conception and design: LG, AAB, AEK; data collection: LG, WMRR; data analysis and interpretation of results: LG; draft manuscript preparation: LG, AAB, AEK.

Funding: The dataset analysed during the current study was acquired through funding from the Wellcome Trust as part of the Medical Engineering Solutions in Osteoarthritis Centre of Excellence at Imperial College London.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Acknowledgments: The authors acknowledge Google Colaboratory from Google Research for providing the deep learning hardware (12GB NVIDIA Tesla K80 graphics processing unit).

Conflicts of Interest: The author declares no conflict of interest.

References

[1] Y. Li, H. Di, Y. Xin, and X. Jiang, 'Optical fiber data glove for hand posture capture', *Optik*, vol. 233, p. 166603, May 2021, doi: 10.1016/j.ijleo.2021.166603.

[2] L. Dipietro, A. M. Sabatini, and P. Dario, 'Evaluation of an instrumented glove for hand-movement acquisition', *J Rehabil Res Dev*, vol. 40, no. 2, pp. 179–189, Apr. 2003.

[3] R. F. Pinto, C. D. Borges, A. Almeida, and I. C. Paula, 'Static hand gesture recognition based on convolutional neural networks', *Journal of Electrical and Computer Engineering*, vol. 2019, 2019.

[4] W. Wu, M. Shi, T. Wu, D. Zhao, S. Zhang, and J. Li, 'Real-time Hand Gesture Recognition Based on Deep Learning in Complex Environments', in *2019 Chinese Control And Decision Conference (CCDC)*, Nanchang, China, Jun. 2019, pp. 5950–5955. doi: 10.1109/CCDC.2019.8833328.

[5] J. S. Sonkusare, N. B. Chopade, R. Sor, and S. L. Tade, 'A Review on Hand Gesture Recognition System', in *2015 International Conference on Computing Communication Control and Automation*, Feb. 2015, pp. 790–794. doi: 10.1109/ICCUBEA.2015.158.

[6] T. Primya, G. Kanagaraj, K. Muthulakshmi, J. Chitra, and A. Gowthami, 'Gesture recognition smart glove for speech impaired people', *Materials Today: Proceedings*, Feb. 2021, doi: 10.1016/j.matpr.2020.12.872.

[7] Z. Halim and G. Abbas, 'A Kinect-Based Sign Language Hand Gesture Recognition System for Hearing- and Speech-Impaired: A Pilot Study of Pakistani Sign Language', *Assistive Technology*, vol. 27, no. 1, pp. 34–43, Jan. 2015, doi: 10.1080/10400435.2014.952845.

[8] V. Metsis, P. Jangyodsuk, V. Athitsos, M. Iversen, and F. Makedon, 'Computer aided rehabilitation for patients with rheumatoid arthritis', 2013, pp. 97–102.

[9] V. Adithya and R. Rajesh, 'A deep convolutional neural network approach for static hand gesture recognition', *Procedia Computer Science*, vol. 171, pp. 2353–2361, 2020.

[10] C. J. L. Flores, A. G. Cutipa, and R. L. Enciso, 'Application of convolutional neural networks for static hand gestures recognition under different invariant features', 2017, pp. 1–4.

[11] R. F. Pinto, C. D. Borges, A. Almeida, and I. C. Paula, 'Static hand gesture recognition based on convolutional neural networks', *Journal of Electrical and Computer Engineering*, vol. 2019, 2019.

[12] N. L. Hakim, T. K. Shih, S. P. Kasthuri Arachchi, W. Aditya, Y.-C. Chen, and C.-Y. Lin, 'Dynamic hand gesture recognition using 3DCNN and LSTM with FSM context-aware model', *Sensors*, vol. 19, no. 24, p. 5429, 2019.

[13] P. Molchanov, S. Gupta, K. Kim, and J. Kautz, 'Hand gesture recognition with 3D convolutional neural networks', 2015, pp. 1–7.

[14] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, 'Large-scale video classification with convolutional neural networks', 2014, pp. 1725–1732.

[15] J. Materzynska, G. Berger, I. Bax, and R. Memisevic, 'The Jester Dataset: A Large-Scale Video Dataset of Human Gestures', in *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, Seoul, Korea (South), Oct. 2019, pp. 2874–2882. doi: 10.1109/ICCVW.2019.00349.

[16] J. Carreira and A. Zisserman, 'Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset', in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, Jul. 2017, pp. 4724–4733. doi: 10.1109/CVPR.2017.502.

[17] C. Tan, F. Sun, T. Kong, W. Zhang, C. Yang, and C. Liu, 'A survey on deep transfer learning', 2018, pp. 270–279.

[18] S. Tammina, 'Transfer learning using vgg-16 with deep convolutional neural network for classifying images', *International Journal of Scientific and Research Publications (IJSRP)*, vol. 9, no. 10, pp. 143–150, 2019.

[19] H.-C. Shin *et al.*, 'Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning', *IEEE transactions on medical imaging*, vol. 35, no. 5, pp. 1285–1298, 2016.

[20] A. Mari, T. R. Bromley, J. Izaac, M. Schuld, and N. Killoran, 'Transfer learning in hybrid classical-quantum neural networks', *Quantum*, vol. 4, p. 340, Oct. 2020, doi: 10.22331/q-2020-10-09-340.

[21] F. Jiang, S. Zhang, S. Wu, Y. Gao, and D. Zhao, 'Multi-layered gesture recognition with Kinect.', *J. Mach. Learn. Res.*, vol. 16, no. 1, pp. 227–254, 2015.

[22] I. Rodríguez-Moreno, J. M. Martínez-Otzeta, B. Sierra, I. Rodriguez, and E. Jauregi, 'Video activity recognition: State-of-the-art', *Sensors*, vol. 19, no. 14, p. 3160, 2019.

[23] M. Vrigkas, C. Nikou, and I. A. Kakadiaris, 'A review of human activity recognition methods', *Frontiers in Robotics and AI*, vol. 2, p. 28, 2015.

[24] R. Mahmoud, S. Belgacem, and M. N. Omri, 'Deep signature-based isolated and large scale continuous gesture recognition approach', *Journal of King Saud University-Computer and Information Sciences*, 2020.

[25] M. Panwar and P. Singh Mehra, 'Hand gesture recognition for human computer interaction', in *2011 International Conference on Image Information Processing*, Nov. 2011, pp. 1–7. doi: 10.1109/ICIIP.2011.6108940.

[26] M. Al-Hammadi, G. Muhammad, W. Abdul, M. Alsulaiman, and M. S. Hossain, 'Hand Gesture Recognition Using 3D-CNN Model', *IEEE Consumer Electronics Magazine*, vol. 9, no. 1, pp. 95–101, Jan. 2020, doi: 10.1109/MCE.2019.2941464.

[27] G. Zhu, L. Zhang, P. Shen, J. Song, S. A. A. Shah, and M. Bennamoun, 'Continuous gesture segmentation and recognition using 3DCNN and convolutional LSTM', *IEEE Transactions on Multimedia*, vol. 21, no. 4, pp. 1011–1021, 2018.

[28] H. Wang, 'Two Stage Continuous Gesture Recognition Based on Deep Learning', *Electronics*, vol. 10, no. 5, p. 534, 2021.

[29] X. Peng, L. Wang, Z. Cai, and Y. Qiao, 'Action and gesture temporal spotting with super vector representation', 2014, pp. 518–527.

[30] X. Chai, Z. Liu, F. Yin, Z. Liu, and X. Chen, 'Two streams recurrent neural networks for large-scale continuous gesture recognition', 2016, pp. 31–36.

 doi:10.20944/preprints202206.0368.v1

[31] N. C. Camgoz, S. Hadfield, O. Koller, and R. Bowden, 'Using Convolutional 3D Neural Networks for User-independent continuous gesture recognition', in *2016 23rd International Conference on Pattern Recognition (ICPR)*, Dec. 2016, pp. 49–54. doi: 10.1109/ICPR.2016.7899606.

[32] M. Bertalmio, G. Sapiro, V. Caselles, and C. Ballester, 'Image inpainting', 2000, pp. 417–424.

[33] V. A. Shanthakumar, C. Peng, J. Hansberger, L. Cao, S. Meacham, and V. Blakely, 'Design and evaluation of a hand gesture recognition approach for real-time interactions', *Multimed Tools Appl*, vol. 79, no. 25, pp. 17707–17730, Jul. 2020, doi: 10.1007/s11042-019-08520-1.

[34] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, 'Learning spatiotemporal features with 3d convolutional networks', 2015, pp. 4489–4497.

[35] M. Abadi *et al.*, 'Tensorflow: A system for large-scale machine learning', 2016, pp. 265–283.

[36] A. A. Taha and A. Hanbury, 'Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool', *BMC medical imaging*, vol. 15, no. 1, pp. 1–28, 2015.

[37] S. Escalera, V. Athitsos, and I. Guyon, 'Challenges in multi-modal gesture recognition', *Gesture recognition*, pp. 1–60, 2017.