Peer-reviewed version available at *Sensors* **2020**, *20*, 5151; doi:10.3390/s20185151

Article

British Sign Language Recognition via Late Fusion of Computer Vision and Leap Motion with Transfer Learning to American Sign Language

Jordan J. Bird[1], Anikó Ekárt[2], and Diego R. Faria[1]

[1] ARVIS Lab - Aston Robotics Vision and Intelligent Systems (https://arvis-lab.io/), Aston University, Birmingham, United Kingdom; {birdj1, d.faria}@aston.ac.uk

[2] School of Engineering and Applied Science, Aston University, Birmingham, United Kingdom; a.ekart@aston.ac.uk

Version August 13, 2020 submitted to Sensors

Abstract: In this work, we show that a late fusion approach to multi-modality in sign language recognition improves the overall ability of the model in comparison to the singular approaches of Computer Vision (88.14%) and Leap Motion data classification (72.73%). With a large synchronous dataset of 18 BSL gestures collected from multiple subjects, two deep neural networks are benchmarked and compared to derive a best topology for each. The Vision model is implemented by a CNN and optimised MLP and the Leap Motion model is implemented by an evolutionary optimised deep MLP topology search. Next, the two best networks are fused for synchronised processing which results in a better overall result (94.44%) since complementary features are learnt in addition to the original task. The hypothesis is further supported by application of the three models to a set of completely unseen data where a multi-modality approach achieves the best results relative to the single sensor method. When transfer learning with the weights trained via BSL, all three models outperform standard random weight distribution when classifying ASL, and the best model overall for ASL classification was the transfer learning multi-modality approach which scored 82.55% accuracy.

Keywords: Sign Language Recognition, Multi-modality, Late Fusion

1. Introduction

Sign language is the ability to converse mainly by use of the hands, as well as in some cases the body, face, and head. Recognition and understanding of Sign Language is thus an entirely visuo-temporal process performed by human beings. In the United Kingdom alone, there are 145,000 deaf adults and children who use British Sign Language (BSL) [1]. Of those people, 15,000 report BSL as their main language of communication [2] which implies a difficulty of communication with those who cannot interpret the language. Unfortunately, when another person cannot interpret sign language (of who are the vast majority), a serious language barrier is present due to disability.

In addition to the individuals who act as interpreters for those who can only converse in Sign Language, or who only feel comfortable doing so, this work aims to improve autonomous classification techniques towards dictation of Sign Language in real-time. The philosophy behind this work is based on a simple argument, *if a building were to have a ramp in addition to stairs for easier access of the disabled, then why should a computer system not be present in order to aid with those hard of hearing or deaf?* In this work, we initially benchmark two popular methods of sign language recognition with Computer Vision and a Leap Motion 3D hand tracking camera after gathering a large dataset of gestures. Following these initial experiments, we then present a multi-modality approach which

Peer-reviewed version available at Sensors 2020, 20, 5151; doi:10.3390/s20185151

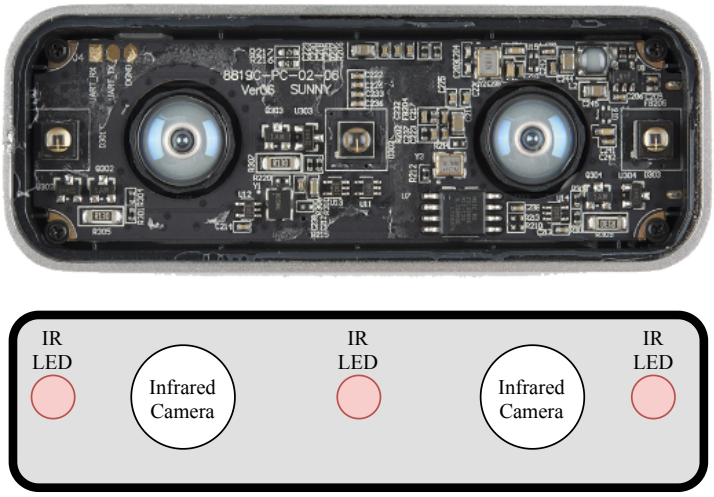

Figure 1. Photograph and labelled sketch of the stereoscopic infrared camera array within a Leap Motion Controller, illuminated by three infrared LEDs.

fuses the two forms of data in order to achieve better results for two main reasons; firstly, mistakes and anomalous data received by either sensor has the chance to be mitigated by the other, and secondly, a deep neural network can learn to extract useful complimentary data from each sensor as well as the standard approach of extracting information towards the class itself. The driving force behind improving the ability of these two sensors is mainly cost, in that the solution presented is of extremely minimal cost and with further improvement beyond the 18 gestures explored in this study, could easily be implemented within public places such as restaurants, schools, and libraries etc. in order to improve the lives of disabled individuals and enable communication with those they otherwise could not communicate with.

In this work, the approaches of single modality learning and classification are compared to multi-modality late fusion. The main scientific contributions presented by this work are as follows:

1. Collection of a large BSL dataset from five subjects and a medium-sized ASL dataset from two subjects.
2. Tuning of classification models for Computer Vision (processing layer prior to output), Leap Motion Classification (evolutionary topology search), and multi-modality late fusion of the two via concatenation to a neural layer. Findings show that multi-modality is the strongest approach for BSL classification compared to the two single-modality inputs as well as state of the art statistical learning techniques.
3. Transfer learning from BSL to improve ASL classification. Findings show that weight transfer to the multi-modality model is the strongest approach for ASL classification.

The remainder of this work is as follows, Section 2 explores the current state of the art for Sign Language Classification. Section 3 details the method followed for these experiments, which includes data collection, data pre-processing and the machine learning pipeline followed. The results for all of the experiments are presented in Section 4 including indirect comparison to other state of the art works in the field, before conclusions are drawn and future work is suggested in Section 5.

2. Related Work

Sign Language Recognition is a collaboration of multiple fields of research which can involve pattern matching, computer vision, natural language processing, and linguistics [3]. Classically, sign language recognition was usually performed by temporal models trained on sequences of video. Many works from the late 1990's through to the mid-2000's found best results when applying varying forms of Hidden Markov Models to videos [4–7]. More recently though, given affordable sensors that provide

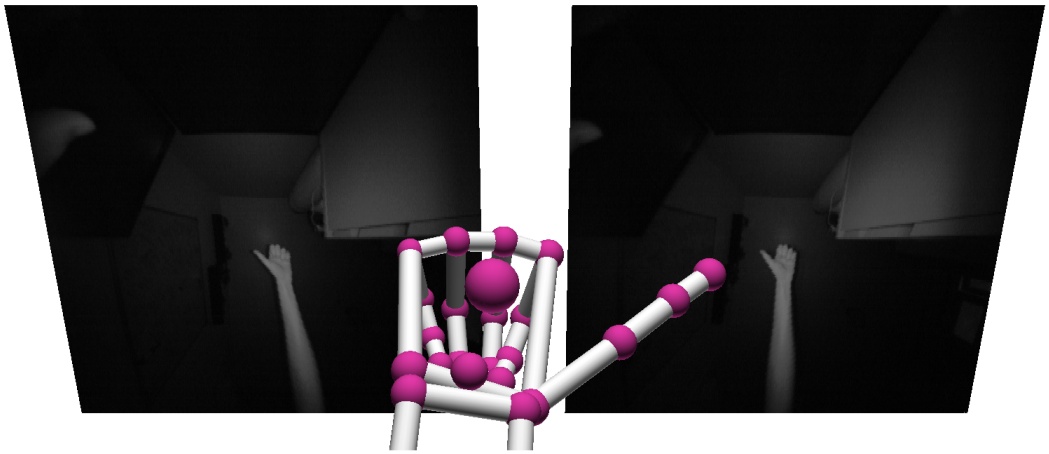

Figure 2. Screenshot of the view from Leap's two infrared cameras and the detected hand reproduced in 3D. Note that this study uses a front-facing view rather than up-facing as shown in the screenshot.

more useful information than a video clip, studies have focused upon introducing this information towards stronger and more robust real-time classification of non-verbal languages. Sign language recognition with depth-sensing cameras such as Kinect and Leap Motion is an exciting area within the field due to the possibility of accessing accurate 3D information from the hand through stereoscopy similar to human depth perception via images from two eyeballs. Kinect allows researchers to access RGBD channels via a single colour camera and a single infrared depth-sensing camera. A Microsoft Kinect camera was used to gather data in [8], and features were extracted using a Support Vector Machine from depth and motion profiles. Researchers in [9] found that generating synchronised colour-coded joint distance topographic descriptor and joint angle topographical descriptor and used as input to a two-steam CNN produced effective results; the CNNs in this study were concatenated by late fusion similar to the multi-modality method in this study and results were around 92% for a 20-class dataset.

The Leap Motion Controller, a sketch of which can be observed in Figure 1, is a device that combines stereoscopy and depth-sensing in order to accurately locate the individual bones and joints of the human hand. An example of the view of the two cameras translated to a 3D representation of the hand can be seen in Figure 2. The device measures 3.5x1.2x0.5 inches and is thus a more portable option compared to the Microsoft Kinect. Features recorded from the 26 letters of the alphabet in American Sign Language were observed to be classified at 79.83% accuracy by a Support Vector Machine algorithm [10]. Similarly to the aforementioned work, researchers found that a different dataset also consisting of 26 ASL letters were classifiable at 93.81% accuracy with a Deep Neural Network [11]. Another example achieved 96.15% with a deep learning approach on a limited set of 520 samples (20 per letter) [12]. Data fusion via Coupled Hidden Markov Models was performed in [13] between Leap Motion and Kinect, which achieved 90.8% accuracy on a set of 25 Indian Sign Language gestures.

In much of the state-of-the-art work in Sign Language recognition, a single modality approach is followed, with multi-modality experiments being some of the latest studies in the field. Additionally, studies often fail to apply trained models to unseen data, ergo towards real-time classification (the ultimate goal of SL recognition). With this in mind, Wang et al. proposed that sign language recognition systems are often affected by noise, which may negatively impact real-time recognition abilities [14]. In this work, we benchmark two single-modality approaches as well as a multi-modality late fusion approach of the two both during training, and on unseen data towards benchmarking a more realistic real-time ability. Additionally, we also show that it is possible to perform transfer learning between two ethnologues with the proposed approaches, for British and American Sign Languages.

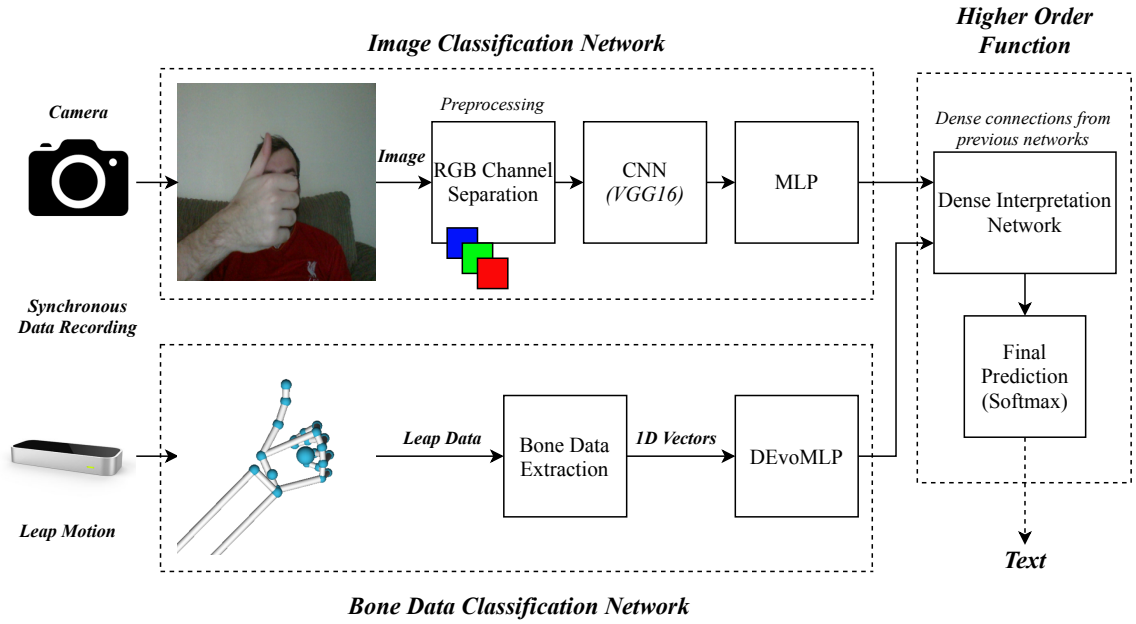

Figure 3. An overall diagram of the three benchmarking experiments. Above shows the process of image classification and below shows Leap Motion data classification for the same problem of sign language recognition. The higher order function network shows the late fusion of the two to form a multi-modality solution.

The inspiration for the network topology and method of fusion in this work comes from [15] (albeit applied to scene recognition in this instance), similarly, this work fuses two differing synchronous data types via late-fusion by benchmarking network topologies at each step. In the aforementioned work though, weights of the networks were frozen for late fusion layer training (derived from benchmarking the two separate models). In this experiment, all weights are able to train from the start of the late fusion network from scratch, and thus, the networks can extract complimentary features from each form of data for classification in addition to the usual method of extracting features for direct classification and prediction.

3. Proposed Approach: Multi-modality Late Fusion of Deep Networks

Within this section, the proposed approach for the late fusion experiments are described. The experiments that this section mainly refers to can be observed in Figure 3 which outlines the image classification, Leap Motion classification, and multi-modality late fusion networks.

3.1. Dataset Collection and Pre-processing

Five subjects contributed to a dataset of British Sign Language. 18 differing gestures were recorded at a frequency of 0.2s each using a laptop, an image was captured using the laptop's webcam and Leap Motion data is recorded from the device situated above the camera facing the subject. This allowed for 'face-to-face' communication, since the subject was asked to communicate as if across from another human being. The 'task-giver' was situated behind the laptop and stopped data recording if the subject made an error while performing the gesture.

From the Leap Motion sensor, data was recorded for each of the thumb, index, middle, ring, and pinky fingers within the frame (labelled 'left' or 'right'). The names of the fingers and bones can be observed in the labelled diagram in Figure 4. For each hand, the start and end positions, 3D angles between start and end positions, and velocities of the arm, palm, and finger bones (metacarpal, proximal, intermediate and distal bones) were recorded in order to numerically represent the gesture

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 August 2020 doi:10.20944/preprints202008.0209.v1

Peer-reviewed version available at Sensors 2020, 20, 5151; doi:10.3390/s20185151

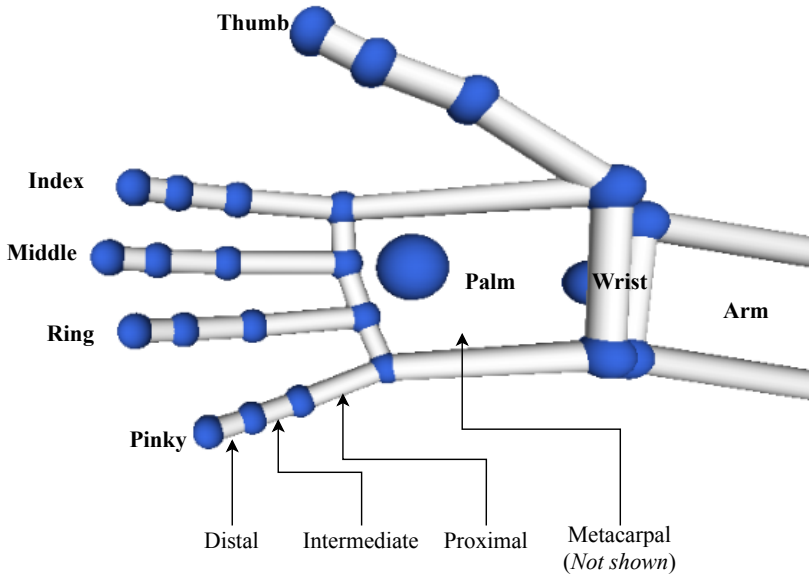

Figure 4. Labelled diagram of the bone data detected by the Leap Motion sensor. Metacarpal bones are not rendered by the LMC Visualiser.

being performed. The pitch, yaw, and roll of the hands were also recorded. If one of the two hands were not detected then its values were recorded as '0' (eg. a left handed action will also feature a vector of zeroes for the right hand). If the sensor did not detect either hand, data collection was automatically paused until the hands were detected in order to prevent empty frames. Thus, every 0.2 seconds, a numerical vector is output to describe the action of both hands. The 3D angle θ between two points a and b in space is calculated by the following:

$$\theta = arccos\left(\frac{ab}{|\,a\,||\,b\,|}\right), \tag{1}$$

where $|\,a\,|$ and $|\,b\,|$ are:

$$\begin{aligned} |\,a\,| &= \sqrt{{a_x}^2 + {a_y}^2 + {a_z}^2} \\ |\,b\,| &= \sqrt{{b_x}^2 + {b_y}^2 + {b_z}^2}, \end{aligned} \tag{2}$$

with regards to the x, y, and z co-ordinates of each point in space. The start and end points of each bone in the hand from the LMC are treated as the two points.

The 18 British Sign Language[1] gestures recorded were selected due to them being common useful words or phrases in language. A mixture of one and two-handed gestures were chosen. Each gesture was recorded twice where subjects switched dominant hands.

The useful gestures for general conversation were:

1. Hello/Goodbye
2. You/Yourself
3. Me/Myself
4. Name
5. Sorry
6. Good

[1] Visual examples of the BSL gestures can be viewed at https://www.british-sign.co.uk/british-sign-language/dictionary/

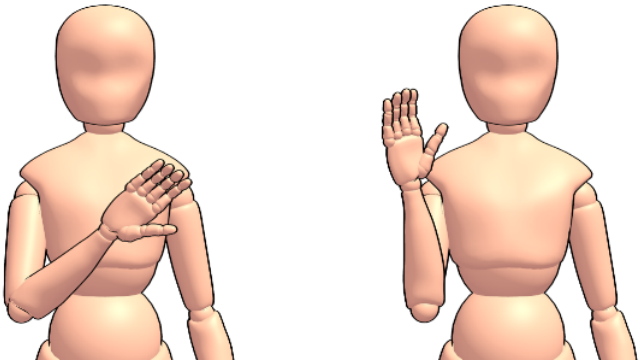

Figure 5. The sign for 'Hello' in British Sign Language

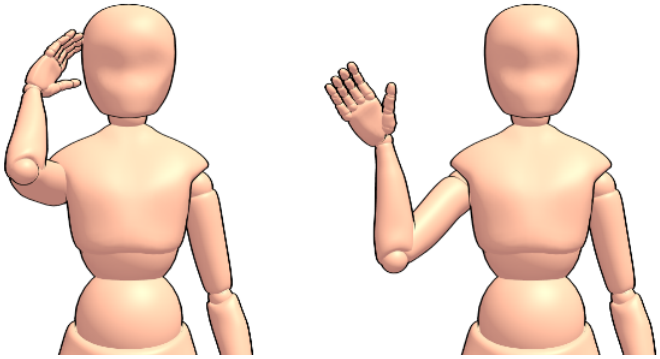

Figure 6. The sign for 'Hello' in American Sign Language

7. Bad
8. Excuse Me
9. Thanks/Thank you
10. Time

The gestures for useful entities were:

1. Airport
2. Bus
3. Car
4. Aeroplane
5. Taxi
6. Restaurant
7. Drink
8. Food

Following this, a smaller set of the same 18 gestures but in American Sign Language[2] are collected from two subjects for thirty seconds each (15 per hand) towards the transfer learning experiment. 'Airport' and 'Aeroplane/Airplane' in ASL are similar, and so, 'Airport' and 'Jet Plane' are recorded instead. Figures 5 and 6 show a comparison of how one signs 'hello' in British and American sign languages; though the gestures differ, the hand is waved and as such it is likely that useful knowledge can be transferred between the two languages.

[2] Visual examples of the ASL gestures can be viewed at https://www.handspeak.com/

Peer-reviewed version available at Sensors 2020, 20, 5151; doi:10.3390/s20185151

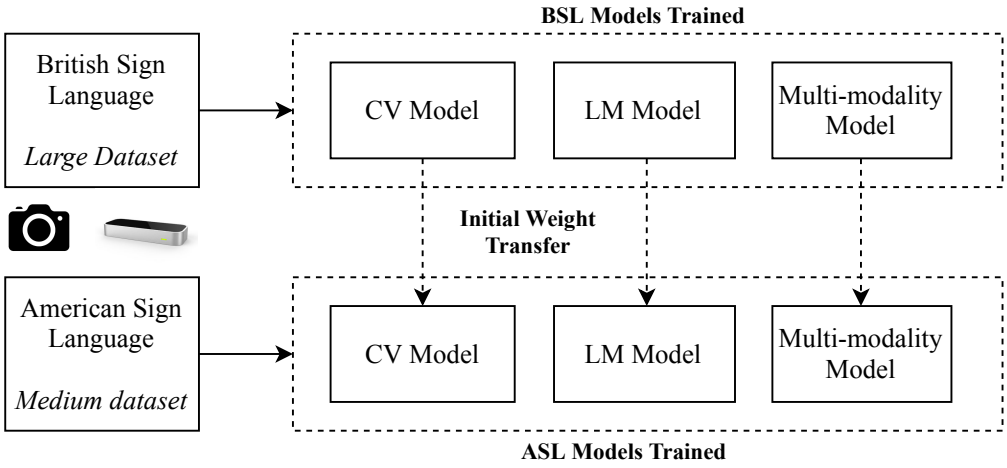

Figure 7. Transfer Learning Experiments which train on BSL and produce initial starting weight distributions for the ASL models

3.2. Deep Learning Approaches

For the image classification network, VGG16 [16] convolutional layers are used as a starting point and the three 4096 neuron hidden layers are removed. The convolutional layers are followed by $2, 4, 8, ..., 4096$ ReLu neuron layers in each of the ten benchmarking experiments to ascertain a best-performing interpretation layer. For the Leap Motion data classification problem, an evolutionary search is performed [17] to also ascertain a best-performing neural network topology; the search is set to a population of 20 for 15 generations, since during manual exploration, stabilisation of a final best result tends to occur at around generation 11. The evolutionary search is run three times in order to mitigate the risk of a local maxima being carried forward to the latter experiments.

With the best CNN and Leap Motion ANN networks derived, a third set of experiments is then run. The best topologies (with softmax layers removed) are fused into a single layer of ReLu neurons in the range $2, 4, 8, ..., 4096$.

All experiments are benchmarked with randomised 10-fold cross validation, and training time is uncapped to a number of epochs and rather executed until no improvement of accuracy occurs after 25 epochs. Thus, the results presented are the maximum results attainable by the network within this boundary of *early stopping*.

Following the experiments on BSL, initial preliminary experiments for Transfer Learning between languages is performed. Figure 7 shows the outline for the transfer experiments, in which the learnt weights from the three BSL models are transferred to their ASL counterparts as initial starting weight distributions and ultimately compared to the usual method of beginning with a random distribution. This experiment is performed in order to benchmark whether there is useful knowledge to be transferred between each of the model pairs.

3.3. Experimental Hardware

The deep learning experiments in this study were performed on an Nvidia GTX 980Ti which has 2816 1190MHz CUDA cores and 6GB of GDDR5 memory. Given the memory constraints, images are resized to 128x128 although they were initially captured in larger resolutions. All deep learning experiments were written in Python for the Keras [18] library and TensorFlow [19] backend.

The statistical models trained in this study were performed with a Coffee Lake Intel Core i7 at a clock speed of 3.7GHz. All statistical learning experiments were written in Python for the SciKit-Learn library [20].

Peer-reviewed version available at Sensors 2020, 20, 5151; doi:10.3390/s20185151

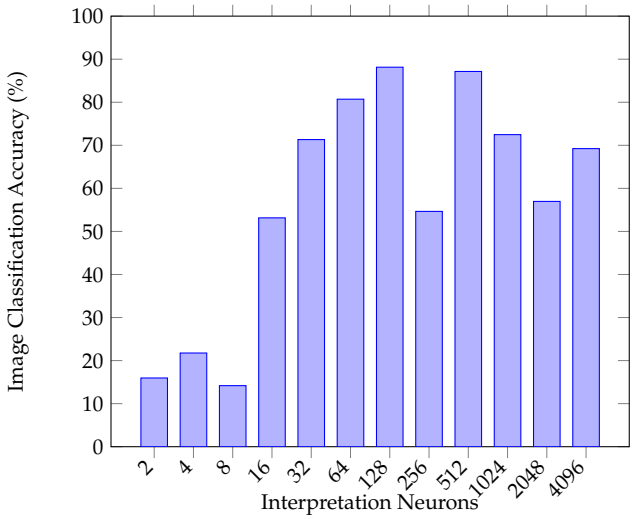

Figure 8. Mean Image 10-fold Classification Accuracy corresponding to interpretation neuron numbers.

Table 1. Final results of the three Evolutionary Searches sorted by 10-fold validation Accuracy along with the total number of connections within the network.

Hidden Neurons	Connections	Accuracy
171, 292, 387	243,090	**72.73%**
57, 329, 313	151,760	70.17%
309, 423, 277	385,116	69.29%

4. Experimental Results

4.1. Fine Tuning of VGG16 Weights and Interpretation Topology

Figure 8 shows the results for tuning of the VGG network for image classification. Each result is given as the classification ability when a layer of neurons are introduced beyond the CNN operations and prior to output. The best result was a layer of 128 neurons prior to output which resulted in a classification accuracy of 88.14%. Most of the results were relatively strong except for 2-8 neurons and, interestingly, layers of 256 and 2048 neurons. Thus, the CNN followed by 128 neurons forms the first branch of the multi-modality system for image processing alongside the best Leap Motion network (in the next section). The SoftMax output layer is removed for purposes of concatenation, and the 128 neuron layer feeds into the interpretation layer prior to output.

4.2. Evolutionary Search of Leap Motion DNN Topology

The evolutionary search algorithm [17] is applied three times for a population of 20 through 15 generations. The maximum number of neurons was 1024, and the maximum number of layers was 5. After an initial random initialisation of solutions, the algorithm performs roulette selection for each solution and generates an offspring (where number of layers, number of neurons per layer are bred). At the start of each new generation, the worst performing solutions outside of the population size 20 range are deleted, and the process runs again. The final best result is reported at the end of the simulation. Table 1 shows the best results for three runs of the Leap Motion classification networks. Of the three, the best model was a deep neural network of 171, 292, 387 neurons which resulted in a classification accuracy of 72.73%. Interestingly, the most complex model found was actually the worst performing of the best three results selected. This forms the second branch of the multi-modality network for Leap Motion classification in order to compliment the image processing network. Similarly to the image processing and network, the SoftMax output layer is removed and the final layer of 387

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 August 2020 doi:10.20944/preprints202008.0209.v1

Peer-reviewed version available at Sensors 2020, 20, 5151; doi:10.3390/s20185151

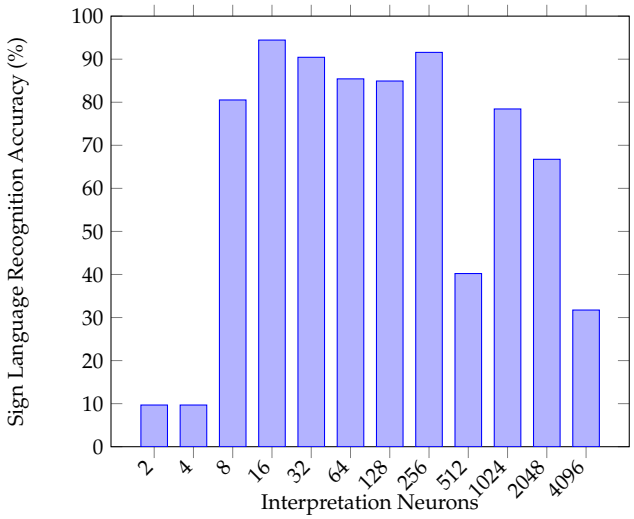

Figure 9. Multi-modality 10-fold Classification Accuracy corresponding to interpretation neuron numbers towards benchmarking the late-fusion network.

Table 2. Sign Language Recognition scores of the three models trained on the dataset

Model	Sign Language Recognition Ability
Computer Vision	88.14%
Leap Motion	72.73%
Multi-modality	**94.44%**

neurons for Leap Motion data classification is connected to the dense interpretation network layer along with the 128 hidden neurons of the image network.

4.3. Fine Tuning the Final Model

Figure 9 shows the results of fine-tuning the best number of interpretation neurons within the late fusion layer, the best set of hyperparameters found to fuse the two prior networks was a layer of 16 neurons which achieved an overall mean classification ability of 94.44%. This best-performing layer of 16 neurons receives input from the Image and Leap Motion classification networks and is connected to a final SoftMax output. Given the nature of backpropagation, the learning process enables the two input networks to perform as they were prior (that is, to extract features and classify data) but a new task is also then possible; to extract features and useful information from either data format which may compliment the other, for example, for correction of common errors, or for contributing to confidence behind a decision.

4.4. Comparison and Analysis of Models

Table 2 shows a comparison of the final three tuned model performances for recognition of British Sign Language through the classification of photographic images (Computer Vision) and bone data (Leap Motion) compared to the multi-modality approach that fuses the two networks together. The maximum classification accuracy of the CV model achieved 88.14%, the Leap Motion model achieved 72.73% but the fusion of the two allowed for a large increase towards 94.44% accuracy. A further comparison to other statistical approaches can be observed in Table 3; although the DNN approach is relatively weak compared to all statistical models except for Gaussian Naive Bayes, it contributes to the Multi-modality approach by extracting features complimentary to the CNN prior to late fusion as well as the task of classification - this, in turn, leads to the multi-modality approach attaining the best overall result. The best statistical model, the Random Forest, was outperformed by the CNN by 1.07% and the Multi-modality approach by 7.37%. Performance aside, it must be noted that the

Table 3. Comparison of other statistical models and the approaches presented in this work.

Model	Input Sensor(s)	Sign Language Recognition Ability
MM(DNN, CNN)	LMC, Camera	**94.44%**
CNN	Camera	88.14%
RF	LMC	87.07%
SMO SVM	LMC	86.78%
QDA	LMC	85.46%
LDA	LMC	81.31%
LR	LMC	80.97%
Bayesian Net	LMC	73.48%
DNN	LMC	72.73%
Gaussian NB	LMC	34.91%

Table 4. Results of the three trained models applied to unseen data

Approach	Correct/Incorrect	Classification Accuracy
Computer Vision	1250/1800	69.44%
Leap Motion	752/1800	41.78%
Multi-modality	**1377/1800**	**76.5%**

statistical approaches are far less computationally complex than deep learning approaches; should the host machine for the task not have access to a GPU with CUDA abilities, a single-modality statistical approach is likely the most realistic candidate. Should the host machine, on the other hand, have access to a physical or cloud-based GPU or TPU, then it would be possible to enable the most superior model which was the deep learning multi-modality approach.

Table 4 shows the final comparison of all three models when tasked with predicting the class labels of unseen data objects, 100 per class (18 classes). The error matrix for the best model, which was the multi-modality approach at 76.5% accuracy can be observed in Figure 10. Interestingly, most classes were classified with high confidence with the exception of three main outliers; 'thanks' was misclassified as 'bus' in almost all cases, 'restaurant' was misclassified as a multitude of other classes, and 'food' was often mistaken for 'drink' although this did not occur vice-versa. Outside of the anomalous classes which must be improved in the future with more training examples, the multi-modality model was able to confidently classify the majority of all other phrases.

Table 5 shows a comparison of state of the art approaches so Sign Language recognition. The training accuracy found in this work is given as comparison since other works report such metric, but it is worth noting that this work showed that classification of unseen data is often lower than the training process. For example, the multi-modality approach score of 94.44% was reduced to 76.5% when being applied to completely unseen data.

4.5. Transfer Learning from BSL to ASL

Table 6 shows the results for transfer learning from BSL to ASL. Interestingly, with the medium sized ASL dataset and no transfer learning, the Multi-modality approach is worse than both the Computer Vision and Leap Motion models singularly. This, and considering that the best model overall for ASL classification was the Multi-modality model with weight transfer from the BSL model, suggests that data scarcity poses an issue for multi-modality models for this problem.

The results show that transfer learning improves the abilities of the Leap Motion and Multi-modality classification approaches to sign language recognition. With this in mind, availability of trained weights may be useful to improve the classification of other datasets regardless of whether or not they are in the same sign language. Overall, the best model for ASL classification was the multi-modality model when weights are transferred from the BSL model. This approach scored 82.55% classification ability on the ASL dataset. The results suggest that useful knowledge can be transferred

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 August 2020 doi:10.20944/preprints202008.0209.v1

Peer-reviewed version available at Sensors 2020, 20, 5151; doi:10.3390/s20185151

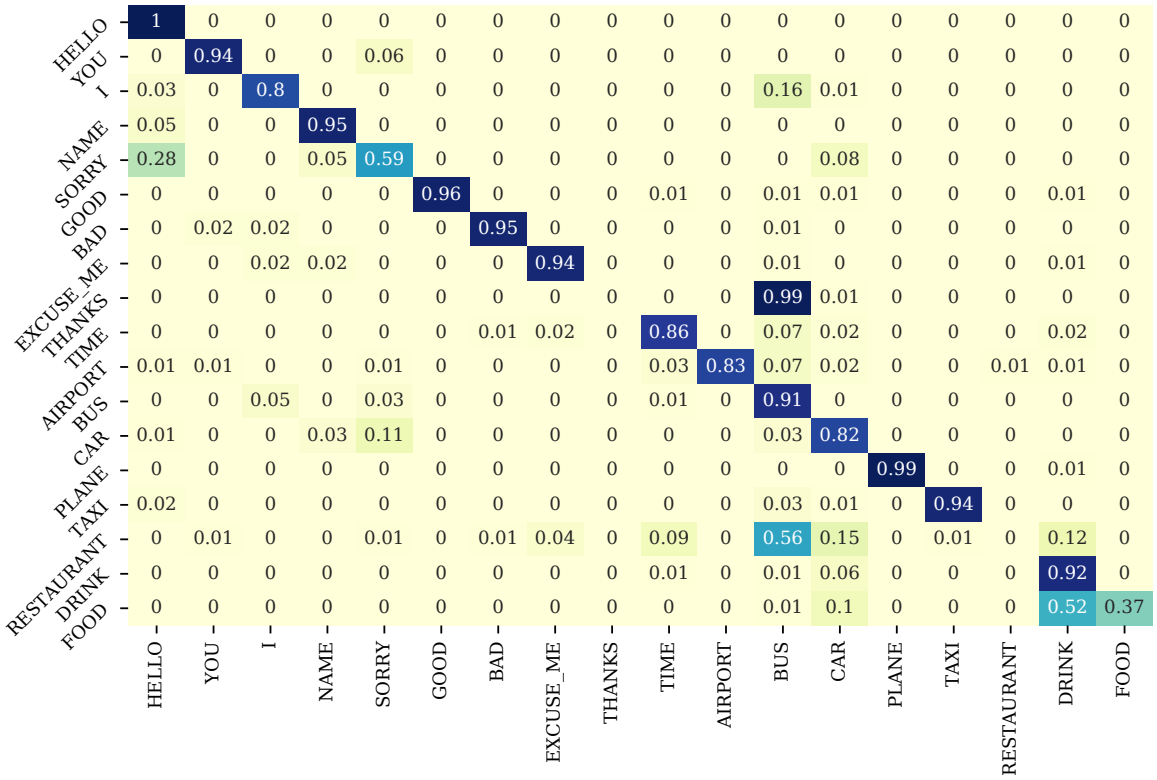

Figure 10. Confusion matrix for the best model (multi-modality, 76.5%) on the set of unseen data (not present during training).

Table 5. Other state of the art works in autonomous Sign Language Recognition, indirectly compared due to operation on different datasets and with different sensors. Note: it was observed in this study that classification of unseen data is often lower than results found during training, but many works do not benchmark this activity

Study	Sensor	Input	Approach	Classes	Score (%)
Huang et al. [21]	Kinect	Skeleton	DNN	26	97.8
Filho et al. [22]	Kinect	Depth	KNN	200	96.31
Morales et al. [23]	Kinect	Depth	HMM	20	96.2
Hisham et al. [24]	LMC	Point Cloud	DTW	28	95
Kumar et al. [25]	LMC	Point Cloud	HMM, BLSTM	50	94.55
Quesada et al. [26]	RealSense	Skeleton	SVM	26	92.31
Kumar et al. [9]	MoCap	Skeleton	2-CNN	20	92.14
Yang [27]	Kinect	Depth	HCRF	24	90.4
Cao Dong et al. [28]	Kinect	Depth	RF	24	90
Elons et al. [29]	LMC	Point Cloud	MLP	50	88
Kumar et al. [30]	Kinect	Skeleton	HMM	30	83.77
Chansri et al. [31]	Kinect	RGB, Depth	HOG, ANN	42	80.05
Chuan et al. [10]	LMC	Point Cloud	SVM	26	79.83
Quesada et al. [26]	LMC	Skeleton	SVM	26	74.59
Chuan et al. [10]	LMC	Point Cloud	KNN	26	72.78
This study	LMC, RGB	Hand feats, RGB	CNN-MLP-LF	18	94.44

Table 6. Results of pre-training and classification abilities of ASL models, with and without weight transfer from the BSL models

Model	Non-transfer from BSL		Transfer from BSL	
	Epoch 0	*Final Ability*	*Epoch 0*	*Final Ability*
Computer Vision	2.98	80.68	13.28	81.82
Leap Motion	5.12	67.82	7.77	70.95
Multi-modality	5.12	65.4	21.31	**82.55**

between sign languages for image classification, Leap Motion classification, and late-fusion of the two towards multi-modality classification.

Though future work is needed for further explore the transfer learning hypotheses, the results in these initial experiments suggest the possibility of success when transferring knowledge between models and ultimately improving their recognition performances.

5. Conclusions and Future Work

This work has presented multiple experiments for the singular sensor and multi-modality approaches to British and American Sign Language. The results from the experiments suggest that a multi-modality approach outperforms the two singular sensors during both training and for classification of unseen data. This work also presented a preliminary Transfer Learning experiment from the large BSL dataset to a medium-sized ASL dataset, in which the best model for classification of ASL was found to be the multi-modality model when weights are transferred from the BSL model. All of the network topologies in this work that were trained, compared, and ultimately fused together towards multi-modality were benchmarked and studied for the first time. Accurate classification of Sign Language, especially unseen data, enables the ability to perform the task autonomously and thus provide a digital method to interpretation of non-spoken language within a situation where interpretation is required but unavailable. In order to realise this possibility, future work is needed. The hypotheses in these experiments were argued through a set of 18 common gestures in both British and American Sign Languages. In future, additional classes are required to allow for interpretation of conversations rather than the symbolic communication enabled by this study. In addition, since multi-modality classification proved effective, further tuning of hyperparameters could enable better results, and other methods of data fusion could be explored in addition to the late fusion approach used in this work. Transfer learning could be explored with other forms of non-spoken language such as, for example, Indo-Pakistani SL which has an ethnologue of 1.5 million people and Brazilian SL with an ethnologue of 200,000 people.

A cause for concern that was noted in this work was the reduction of ability when unseen data is considered, which is often the case in machine learning exercises. Such experiments and metrics (ability on unseen dataset, per-class abilities) are rarely performed and noted in the State of the Art works within sign language recognition. Since the main goal of autonomous sign language recognition is to provide a users with a system which can aid those who otherwise may not have access to a method of translation and communication, it is important to consider how such a system would perform when using trained models to classify data that was not present in the training set. That is, real-time classification of data during usage of the system and subsequently the trained classification models. In this work, high training results were found for both modalities and multi-modality, deriving abilities that are competitive when indirectly compared to the state of the art works in the field. When the best performing 94.44% classification ability model (multi-modality) was applied to unseen data, it achieved 76.5% accuracy mainly due to confusion within the 'thanks' and 'restaurant' classes. Likewise, the Computer Vision model reduced from 88.14% to 69.44% and the Leap Motion model reduced from 72.73% to 41.78% when comparing training accuracy and unseen data classification ability. Future

work is needed to enable the models a better ability to generalise towards real-time classification abilities that closely resemble their abilities observed during training.

Author Contributions

Conceptualization, Jordan J. Bird; Investigation, Jordan J. Bird; Methodology, Jordan J. Bird; Software, Jordan J. Bird; Supervision, Aniko Ekart and Diego R. Faria; Writing – original draft, Jordan J. Bird; Writing – review editing, Jordan J. Bird, Aniko Ekart and Diego R. Faria.

1. Ipsos, M.; others. GP patient survey–national summary report. *London: NHS England* **2016**.
2. ONS. 2011 Census: Key statistics for England and Wales, March 2011, 2012.
3. Wadhawan, A.; Kumar, P. Sign Language Recognition Systems: A Decade Systematic Literature Review. *Archives of Computational Methods in Engineering* **2019**, pp. 1–29.
4. Starner, T.; Pentland, A. Real-time american sign language recognition from video using hidden markov models. In *Motion-based recognition*; Springer, 1997; pp. 227–243.
5. Assan, M.; Grobel, K. Video-based sign language recognition using hidden markov models. International Gesture Workshop. Springer, 1997, pp. 97–109.
6. Vogler, C.; Metaxas, D. Parallel hidden markov models for american sign language recognition. Proceedings of the Seventh IEEE International Conference on Computer Vision. IEEE, 1999, Vol. 1, pp. 116–122.
7. Haberdar, H.; Albayrak, S. Real time isolated turkish sign language recognition from video using hidden markov models with global features. International Symposium on Computer and Information Sciences. Springer, 2005, pp. 677–687.
8. Agarwal, A.; Thakur, M.K. Sign language recognition using Microsoft Kinect. 2013 Sixth International Conference on Contemporary Computing (IC3). IEEE, 2013, pp. 181–185.
9. Kumar, E.K.; Kishore, P.; Kumar, M.T.K.; Kumar, D.A. 3D sign language recognition with joint distance and angular coded color topographical descriptor on a 2–stream CNN. *Neurocomputing* **2020**, *372*, 40–54.
10. Chuan, C.H.; Regina, E.; Guardino, C. American sign language recognition using leap motion sensor. 2014 13th International Conference on Machine Learning and Applications. IEEE, 2014, pp. 541–544.
11. Chong, T.W.; Lee, B.G. American sign language recognition using leap motion controller with machine learning approach. *Sensors* **2018**, *18*, 3554.
12. Naglot, D.; Kulkarni, M. Real time sign language recognition using the leap motion controller. 2016 International Conference on Inventive Computation Technologies (ICICT). IEEE, 2016, Vol. 3, pp. 1–5.
13. Kumar, P.; Gauba, H.; Roy, P.P.; Dogra, D.P. Coupled HMM-based multi-sensor data fusion for sign language recognition. *Pattern Recognition Letters* **2017**, *86*, 1–8.
14. Wang, X.; Jiang, F.; Yao, H. DTW/ISODATA algorithm and Multilayer architecture in Sign Language Recognition with large vocabulary. 2008 International Conference on Intelligent Information Hiding and Multimedia Signal Processing. IEEE, 2008, pp. 1399–1402.
15. Bird, J.J.; Faria, D.R.; Premebida, C.; Ekárt, A.; Vogiatzis, G. Look and Listen: A Multi-modality Late Fusion Approach to Scene Classification for Autonomous Machines. *arXiv preprint arXiv:2007.10175* **2020**.
16. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. International Conference on Learning Representations, 2015.
17. Bird, J.J.; Ekart, A.; Buckingham, C.D.; Faria, D.R. Evolutionary optimisation of fully connected artificial neural network topology. Intelligent Computing-Proceedings of the Computing Conference. Springer, 2019, pp. 751–762.
18. Chollet, F.; others. Keras. https://keras.io, 2015.
19. Abadi, M.; Agarwal, A.; Barham, P.; Brevdo, E.; Chen, Z.; Citro, C.; Corrado, G.S.; Davis, A.; Dean, J.; Devin, M.; Ghemawat, S.; Goodfellow, I.; Harp, A.; Irving, G.; Isard, M.; Jia, Y.; Jozefowicz, R.; Kaiser, L.; Kudlur, M.; Levenberg, J.; Mané, D.; Monga, R.; Moore, S.; Murray, D.; Olah, C.; Schuster, M.; Shlens, J.; Steiner, B.; Sutskever, I.; Talwar, K.; Tucker, P.; Vanhoucke, V.; Vasudevan, V.; Viégas, F.; Vinyals, O.; Warden, P.; Wattenberg, M.; Wicke, M.; Yu, Y.; Zheng, X. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems, 2015. Software available from tensorflow.org.

20. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, E. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **2011**, *12*, 2825–2830.
21. Huang, J.; Zhou, W.; Li, H.; Li, W. Sign language recognition using real-sense. 2015 IEEE China Summit and International Conference on Signal and Information Processing (ChinaSIP). IEEE, 2015, pp. 166–170.
22. Costa Filho, C.F.F.; Souza, R.S.d.; Santos, J.R.d.; Santos, B.L.d.; Costa, M.G.F. A fully automatic method for recognizing hand configurations of Brazilian sign language. *Research on Biomedical Engineering* **2017**, *33*, 78–89.
23. Caballero Morales, S.O.; Trujillo Romero, F. 3D modeling of the mexican sign language for a speech-to-sign language system. *Computación y Sistemas* **2013**, *17*, 593–608.
24. Hisham, B.; Hamouda, A. Arabic Static and Dynamic Gestures Recognition Using Leap Motion. *J. Comput. Sci.* **2017**, *13*, 337–354.
25. Kumar, P.; Gauba, H.; Roy, P.P.; Dogra, D.P. A multimodal framework for sensor based sign language recognition. *Neurocomputing* **2017**, *259*, 21–38.
26. Quesada, L.; López, G.; Guerrero, L. Automatic recognition of the American sign language fingerspelling alphabet to assist people living with speech or hearing impairments. *Journal of Ambient Intelligence and Humanized Computing* **2017**, *8*, 625–635.
27. Yang, H.D. Sign language recognition with the kinect sensor based on conditional random fields. *Sensors* **2015**, *15*, 135–147.
28. Dong, C.; Leu, M.C.; Yin, Z. American sign language alphabet recognition using microsoft kinect. Proceedings of the IEEE conference on computer vision and pattern recognition workshops, 2015, pp. 44–52.
29. Elons, A.; Ahmed, M.; Shedid, H.; Tolba, M. Arabic sign language recognition using leap motion sensor. 2014 9th International Conference on Computer Engineering & Systems (ICCES). IEEE, 2014, pp. 368–373.
30. Kumar, P.; Saini, R.; Roy, P.P.; Dogra, D.P. A position and rotation invariant framework for sign language recognition (SLR) using Kinect. *Multimedia Tools and Applications* **2018**, *77*, 8823–8846.
31. Chansri, C.; Srinonchat, J. Hand gesture recognition for Thai sign language in complex background using fusion of depth and color video. *Procedia Computer Science* **2016**, *86*, 257–260.