Identifying American Sign Language Fingerspelling Images Using TensorFlow

Ayisat Akintoye, Jules Donaghey, Kathryn Hays, Brian Nguyen

Applied Cognition and Neuroscience Program

University of Texas at Dallas

# **Abstract**

Sign language is the formalized nonverbal communication utilized by deaf, hard of hearing, nonverbal, and abled people. Communication via sign language is often met with a language barrier that can be augmented using technology, like machine learning and artificial intelligence. This project sought to identify fingerspelling hand gestures based on American Sign Language (ASL) using a convolutional neural network. Variance in size, shape, and position of hands and fingers pose a potential difficulty in the automated identification of signs. Static images were presented, as well as familiar and novel images. There was an average accuracy value of 80% over time and increasing the values of the training dataset in increments showed that higher accuracy was achieved with greater portions of the dataset. This project has the potential to work as a baseline for a more complex neural network for ASL to text interpretation.

# **Introduction**

Technology is one of the main driving forces in developing novel communication methods. Communication with the deaf community is often met with a language barrier that can be augmented using technology in a way similar to multilingual machine translation programs. Using machines to interpret languages, while allowing for communication between parties that would require real-time translation, has been a focus of many researchers (Bretzner, Laptev & Lindeberg, 2002; Freeman & Roth, 1994; Garcia & Viesca, 2016; Liang & Ouhyoung, 1998; Murakami & Taguchi, 1991; Starner, Weaver, & Pentland, 1998; Wang et al., 2012; Yang et al., 2002). A common goal would be to provide a solution in a precise manner with little error through mobile applications commonly accessible in developed countries.

Research concerning communication in the deaf community has led to the creation of hand recognition software and machine learning algorithms to interpret sign languages. Various methods of recognition has been used including motion tracking gloves (Hernández-Rebollar et al., 2002; Kim, Jang, & Bien, 1996; Liang & Ouhyoung, 1998) or the use of a neuromorphic cameras to track movements (Ameen & Vadera, 2017; Liu & Fujimura, 2004; Rivera-Acosta et al., 2017; Xu, 2012; Zhang & Tian, 2015; Zheng et al., 2017).

Only a few studies have implemented imaging with web cameras to conduct the same tests using only image processing algorithms (Chaudary, 2017; Garcia & Viesca, 2016; Starner, Weaver, & Pentland, 1998; Xu, 2012). The main problems concerning building this type of interpretation system include: 1) hand recognition accuracy, 2) accuracy of the results, and 3) efficiency of the algorithm. Previous studies have established that accuracy of hand recognition can be increased by processing the images that are being tested by the algorithm, with images consisting of contrasting colors and enlarged pixels are easier for algorithms to differentiate based upon feature extraction techniques (Ameen & Vadera, 2017; Bretzner, Laptev & Lindeberg, 2002; Freeman & Roth, 1994; Hernández-Rebollar et al., 2002; Liu & Fujimura, 2004; ; LeCun, et al., 2015; Rivera-Acosta et al., 2017; Zheng et al., 2017). Efficiency is heavily dependent on the programming implemented and the datasets being used; the more intricate the network is, the more efficient the program becomes. Higher accuracies of the results are often obtained from retraining the networks and increasing the networks’ knowledge with various stockpiled images to increase its feature identification abilities (Hinton et al., 2006).

There are similar studies that use neural networks and computational modeling approaches to the translation of ASL, commonly fingerspelling, with the end goal of real-time recognition (Bretzner, Laptev & Lindeberg, 2002; Chaudary, 2017; Garcia & Viesca, 2016; Kindiroglu et al., 2012; Murakami & Taguchi, 1991; Starner, Weaver, & Pentland, 1998). Our project sought to identify fingerspelling hand gestures based on American Sign Language (ASL) using a convolutional neural network. We created a simplified version of existing programs; an algorithm which can identify images containing signed communication. As with most languages, the lexicon of American Sign Language (ASL) is vast and so, instead of words, we are interested in the identification of fingerspelling gestures. Fingerspelling is a component of sign language wherein individual letters are formed by the fingers to spell out words. By narrowing our focus to fingerspelling, we do not have to consider or include dialectal variations of many signs, as well as fewer instances of coarticulation.

This project is based on the concept of transfer learning, a machine learning technique where models solve problems by using knowledge learned from solving different but related problems. This technique utilizes the models that were trained on larger datasets and integrate them with final classification layers fitted for new more specific data. As a result, transfer learning demands less time and data. Our training was followed with testing of the network with a new dataset containing the same signed letters with new images followed by analysis of the algorithms’ accuracy. Our goal is to ascertain the reliability of the convolutional neural network when introduced to novel material.

# **Methods**

We used a machine learning algorithm capable of recognizing a number of fingerspelling images and interpreting them. A dataset consisting of 24 static fingerspelling images was compiled, excluding the dynamic signs for the letters “j” and “z”. The algorithm used was adapted from programs found using Tensorflow and machine learning algorithms using Python programming similar to those used in research from other studies. Once training with the original ‘familiar’ dataset yielded positive results with an acceptable percentage accuracy, testing with a novel dataset was conducted, consisting of 24 new images containing the same fingerspelling signs as the familiar dataset with a different set of hands.



*Figure 1*. Path of study.

## Dataset Compilation and Processing

Our dataset came from the ASL FingerSpelling Dataset from the University of  
Surrey’s Center for Vision, Speech and Signal Processing. Only color images were used. Dataset A is comprised of 24 static signs of ASL captured from 5 users in different sessions with similar lighting and backgrounds. It contains 65,000 images with their height-to-width ratios varying significantly with an average of approximately 150x150 pixels. Due to the similarity between images, we grouped them into one class, picking random letters from the fingerspelling images. Data from four out of five users was used to train the network, and the remaining was used for testing.

The Tensorflow model requires square 299x299 RGB images, therefore, all images were resized. We scaled the pixel values so that they stay in the range [0, 255] and normalize the data with the ‘input\_mean and input\_std’ flags; first, subtract ‘input\_mean’ from each pixel value, then divide it by ‘input\_std’. An ASL-specific image manipulation technique was employed to display images of ASL signs that have been horizontally flipped to assess performance on both left and right hands.

## Convolutional Neural Network

A convolutional neural Network (CNN) will be used to classify ASL letters. CNNs can translate results into a probability by applying a softmax function. The probability of classified as k-th category for n-th data xn is:

A screenshot of a cell phone

Description generated with very high confidence

TensorFlow has a ‘tf.nn.softmax’ function which calculates the probability directly from the matrix F:

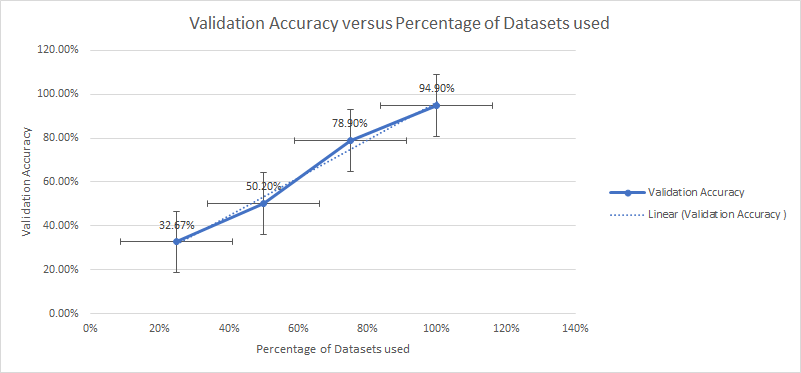
A close up of text on a white background

Description generated with very high confidence

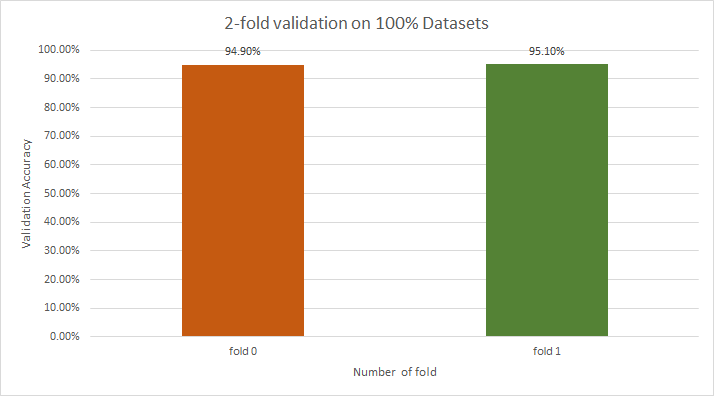
TensorFlow’s image recognition model Inception-v3 is pre-trained on the ImageNet Large Visual Recognition Challenge using data from 2012. We compared the effectiveness of a variety of pre-trained weights. The results were analyzed using TensorBoard, which created graphs and histograms of the accuracy and values for the weights and biases.

# **Results**

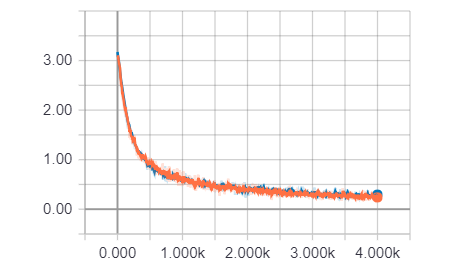
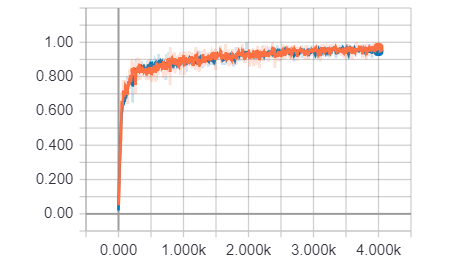
There was an average accuracy value of 80% over time. When the training dataset was split and ran in portions (first, 25% then 50%, then 75%) we found that higher accuracy results were achieved with greater values of the dataset. The highest accuracy of 94.9% was achieved in the very last layer train. Twofold cross-validation of the dataset yielded similar results. Accuracy increased over time and cross-entropy values decreased exponentially indicating low errors, but it never reached zero, which would have indicated a perfect model. The weights and biases were similarly distributed for both the training and test sessions.

**

*Figure 2.* Accuracy of the network over different percentages of data; with more data, higher accuracy percentages were achieved.



*Figure 3.* Two-fold cross validation yielded the similar results.



*Figure 4.* Accuracy (left) and cross-entropy (right) graphs produced by TensorBoard. The training (orange) and test (blue) runs were similar.

# **Discussion**

Research into sign recognition either make use of motion tracking gloves to avoid the problem of depth perception, or focus on the biological and cognitive processes that involve visual processing (Allison, Puce & McCarthy, 2000, Arbib, 2005, Emmorey, Thompson & Colvin, 2009; Muir & Richardson, 2005). Our study focused on using a simple device, a webcam, to develop the images used for processing. This allowed us to create a highly accessible classification algorithm with a moderately high level of accuracy. It is common for cellular devices to come equipped with cameras, and so a more advanced version of this algorithm could be created for smartphone applications to accompany the other translation services already available. It is already possible to conduct real-time detection (PC with webcam) on iOS (using Xcode) and Android (using Android Studio). The next step after training a more advanced version of this model would be to incorporate latent semantics analysis to provide contextual narrowing of sign interpretation. This could serve as a form of next level processing similar to technology currently used for predictive text, similar to Garcia & Viesca (2016) or Cooper, Holt & Bowden (2011).

Another interesting advancement in this research deals with incorporation of gestures into user interfaces allowing for more natural communication and interactions between artificial intelligence and humans (Rautaray & Agrawal, 2015; Wu & Huang, 1999). Gestures deliver more natural, creative and intuitive methods for communicating with our computers and advancements would drive research in gesture taxonomies, recognition techniques, and software frameworks. Research advancements in the field of image recognition are very wide in scope; our project serves as a baseline for studies in all aspects of the field.

# **References**

Allison, T., Puce, A., & McCarthy, G. (2000). Social perception from visual cues: role of the STS region. Trends in cognitive sciences, 4(7), 267-278.

Ameen, S., & Vadera, S. (2017). A convolutional neural network to classify American Sign Language fingerspelling from depth and colour images. *Expert Systems*, *34*(3).

Arbib, M. A. (2005). From monkey-like action recognition to human language: An evolutionary framework for neurolinguistics. Behavioral and brain sciences, 28(2), 105-124.

Bretzner, L., Laptev, I., & Lindeberg, T. (2002, May). Hand gesture recognition using multi-scale colour features, hierarchical models and particle filtering. In Automatic Face and Gesture Recognition, 2002. Proceedings. Fifth IEEE International Conference on (pp. 423-428). IEEE.

Chaudhary, B. (2017). Real-time translation of sign language into text. Data Science Retreat. https://github.com/BelalC/sign2text, apr.

Cooper, H., Holt, B., & Bowden, R. (2011). Sign language recognition. In Visual Analysis of Humans (pp. 539-562). Springer, London.

Emmorey, K., Thompson, R., & Colvin, R. (2008). Eye gaze during comprehension of American Sign Language by native and beginning signers. Journal of Deaf Studies and Deaf Education, 14(2), 237-243.

Freeman, W. T., & Roth, M. (1995, June). Orientation histograms for hand gesture recognition. In International workshop on automatic face and gesture recognition (Vol. 12, pp. 296-301).

Garcia, B., & Viesca, S. (2016). Real-time American sign language recognition with convolutional neural networks. *Convolutional Neural Networks for Visual Recognition*.

Hernandez-Rebollar, J. L., Lindeman, R. W., & Kyriakopoulos, N. (2002). A multi-class pattern recognition system for practical fingerspelling translation. In Multimodal Interfaces, 2002. Proceedings. Fourth IEEE International Conference on (pp. 185-190). IEEE.

Hinton, G. E., Osindero, S., & Teh, Y. W. (2006). A fast learning algorithm for deep belief nets. *Neural computation*, *18*(7), 1527-1554.

Image Recognition. *TensorFlow*. (n.d.). Retrieved September, 2018, from <https://www.tensorflow.org/tutorials/images/image_recognition>

Kim, J. S., Jang, W., & Bien, Z. (1996). A dynamic gesture recognition system for the Korean sign language (KSL). IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 26(2), 354-359.

Kindiroglu, A. A., Yalcin, H., Aran, O., Hruz, M., Campr, P., Akarun, L., & Karpov, A. (2012). Automatic recognition fingerspelling gestures in multiple languages for a communication interface for the disabled. Pattern Recognition and Image Analysis, 22(4), 527-536.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *nature*, *521*(7553), 436.

Liang, R. H., & Ouhyoung, M. (1998, April). A real-time continuous gesture recognition system for sign language. In Automatic Face and Gesture Recognition, 1998. Proceedings. Third IEEE International Conference on (pp. 558-567). IEEE.

Liu, X., & Fujimura, K. (2004, May). Hand gesture recognition using depth data. In null (p. 529). IEEE.

Muir, L. J., & Richardson, I. E. (2005). Perception of sign language and its application to visual communications for deaf people. Journal of Deaf studies and Deaf education, 10(4), 390-401.

Murakami, K., & Taguchi, H. (1991, April). Gesture recognition using recurrent neural networks. In Proceedings of the SIGCHI conference on Human factors in computing systems (pp. 237-242). ACM.

Pugeault, N., & Bowden, R. (2011, November). Spelling it out: Real-time ASL fingerspelling recognition. In Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on (pp. 1114-1119). IEEE.

Rautaray, S. S., & Agrawal, A. (2015). Vision based hand gesture recognition for human computer interaction: a survey. Artificial Intelligence Review, 43(1), 1-54.

Rivera-Acosta, M., Ortega-Cisneros, S., Rivera, J., & Sandoval-Ibarra, F. (2017). American Sign Language Alphabet Recognition Using a Neuromorphic Sensor and an Artificial Neural Network. *Sensors*, *17*(10), 2176.

Starner, T., Weaver, J., & Pentland, A. (1998). Real-Time American Sign Language Recognition Using Desk and Wearable Computer Based Video. M.I.T Media Laboratory Perceptual Computing Section Technical Report No. 466

Wang, Y., Yang, C., Wu, X., Xu, S., & Li, H. (2012). Kinect Based Dynamic Hand Gesture Recognition Algorithm Research. 2012 4th International Conference on Intelligent Human-Machine Systems and Cybernetics

Wu Y., Huang T.S. (1999) Vision-Based Gesture Recognition: A Review. In: Braffort A., Gherbi R., Gibet S., Teil D., Richardson J. (eds) Gesture-Based Communication in Human-Computer Interaction. GW 1999. Lecture Notes in Computer Science (Lecture Notes in Artificial Intelligence), vol 1739. Springer, Berlin, Heidelberg

Xu, D. (2012). Real-time dynamic gesture recognition system based on depth perception for robot navigation. 2012 IEEE International Conference on Robotics and Biomimetics (ROBIO)

Yang, M., Ahuja, N., & Tabb, M. (2002). Extraction of 2D motion trajectories and its application to hand gesture recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 24 (8), 1061 - 1074

Zhang, C. & Tian, Y. (2015). Histogram of 3D Facets: A depth descriptor for human action and hand gesture recognition. Computer Vision and Image Understanding, 139 (2015), 29-39

Zheng, L., Liang, B., & Jiang, A. (2017). Recent Advances of Deep Learning for Sign Language Recognition. 2017 International Conference on Digital Image Computing: Techniques and Applications (DICTA). doi:10.1109/dicta.2017.8227483