[[1]](#footnote-1)

American Sign Language Image Classification(April 2019)

Michael Barnard, Nikita Buslov, Rosemond Fabien and Trung Tran, Member, TRM

*Abstract*—American Sign Language (ASL) image classification suffers from noise and overtraining due to limited labeled data and people not following instructions. In this report, we examine how different features can be extracted from an image to assist in classification of ASL images.

In particular, the abstract must be self-contained, without abbreviations, footnotes, or references. It should be a microcosm of the full article. The abstract must be between 150–250 words. Be sure that you adhere to these limits; otherwise, you will need to edit your abstract accordingly. The abstract must be written as one paragraph, and should not contain displayed mathematical equations or tabular material. The abstract should include three or four different keywords or phrases, as this will help readers to find it. It is important to avoid over-repetition of such phrases as this can result in a page being rejected by search engines. Ensure that your abstract reads well and is grammatically correct.

*Index Terms*—American Sign Language (ASL), random forest, hue, saturation, and value (HSV), k-nearest neighbors (KNN) convolutional neural network (CNN), Exchangeable Image File (EXIF)

# INTRODUCTION

D

ESPITE noise, random shoulders, and other items besides the hand itself appearing in the images. The preprocessing was able to crop out the shoulders using a box blur and an HSV threshold. Outlined in this letter is the implementation of the preprocessing, feature extractions, and random forest classifier. Furthermore, an explanation on why a convolutional neural network (CNN) classification was decided in lieu of a random forest will be provided alongside their results. The results of the CNN on the dataset ranged from:

# Implementation

The following section is separated into 4 parts: Preprocessing, Feature Extractions, Random Forests, and CNN.

To insert images in *Word,* position the cursor at the insertion point and either use Insert | Picture | From File or copy the image to the Windows clipboard and then Edit | Paste Special | Picture (with “float over text” unchecked).

## Preprocessing

The dataset comprised of 100x100 pixeled images. Each image is expected to have only one ASL letter signed in the center with a white opaque background. However, upon further inspection of the dataset, people did not follow instructions when captured their images. There were several issues in the images that includes but are not limited to: 75x100 pixeled images, items appearing in the image other than the hand itself such as a shoulder, non-white backgrounds, reflective backgrounds, noisy backgrounds (i.e. brick walls, pillow case), dark shadows in the images, hand not centered, and some of them were rotated by 90 degrees when they were opened.

All the images in the dataset were resized to 100x100 pixels and the orientation of the images were fixed by examining their Exchangeable Image File (EXIF) data.

Another issue was how to account for different skin tones in the images. Red, green, and blue (RGB) images would have not been viable anymore because the training may not account for enough variation of skin tones. Thus, gray scaling the images was the next step to try to mitigate this issue. However, with the darker images, the gray scaling only made the hand more unrecognizable since we applied a threshold mask to the image in order to extract the hand only. Inspecting the properties of the image revealed that on the hue, saturation, and value (HSV) spectrum, the hand always had a significantly higher saturation than the background even on a background where the color closely matched the hand’s skin tone.

At first, a blob detection was used to try and crop out the hand from the background, however, there was too much noise that was picked up by the detector that the blob cropped in pixels that were not the hand, sometimes even the shoulder which rendered the blob detection useless.

A picture containing clothing

Description automatically generatedA close up of a hand

Description automatically generated

A picture containing monitor, screen, television, object

Description automatically generatedA picture containing monitor, television, screen, object

Description automatically generatedFig. 1. Each row of images is a different step in the preprocess starting with the second row. From top to bottom:

The original RGB images.

Apply a box blur of 2.

Get saturation of images.

Threshold saturated images.

Crop images.

Through some research and reading what previous people have tried on the internet, the following steps outlined in Fig. 1 were made for preprocessing the dataset. In order to crop out the shoulder from images, a box blur was first applied to the original RGB image to get rid of as much noise on the image as possible. Get only the saturation of the image. Take a 10x10 pixel window in the middle of the 100x100 pixeled image. Under the assumption that the hand is in the center on the image, the 10x10 window will be comprised of mostly the hand. By finding the mean of the window and using that as the threshold value, a rough outline of the hand can be found. However, this does not work for all the images, as seen in the 4th and 5th images. There is almost no discernable hand shape in the resulting Boolean image. From the 10x10 pixel window, slowly increase the size of the window until the number of pixels on the edges of the window meet a ratio of false and true pixels, 2.5, to signify that the three other edges of the image are the wall and the fourth edge will have the arm in it so it would have to be true values there.

The resulting images are cropped with only the hand in the center and then resized to a 100x100 pixels ready to be trained or tested on. This worked very well in most cases, as shown in Fig. 1. The shoulders are cropped out and the hand makes up most of the image now. This preprocessing step depends on the hand being centered in the image to work correctly, otherwise; the cropping method will give poor results.

## Feature Extractions

An edge histogram was used as a feature. This was a recommendation from the instructor. Using a Gaussian blur to reduce more of the noise in the images. Sobel edge detection was performed in the x and y dimensions to calculate the type (horizontal, diagonal, and vertical) and location of the edges. The resulting image is partitioned into 9 equal sections where the x and y edges are averaged and then placed into an 2x9 array.

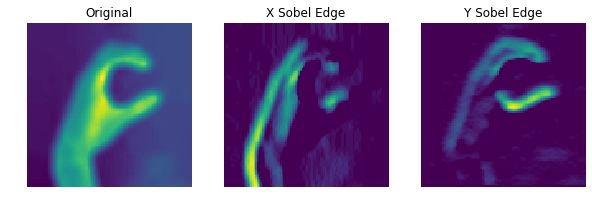


Fig. 2. The three images above illustrate how the edge histogram feature extraction worked. From left to right, that is the original image after preprocessing with an added Gaussian blur to reduce more noise. The following two images are the Sobel edge detection on the x and y dimensions respectively.

## Random Forests

A random forest was decided during the initial development because the team had more experience with it. From the preliminary testing, it was decided that a random forest would be a backup plan if the CNN did not give higher accuracy results.

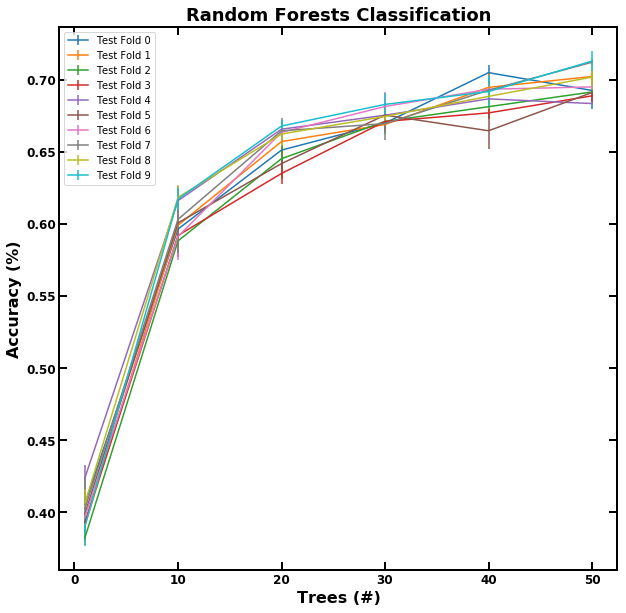


Fig. 3. A random forest classification of 50 trees resulted in approximately 70% accuracy when tested on the class’ dataset. 10 testing folds were performed and increasing the number of trees above 50 did not result in a better score so it was maxed out at 50 trees. “entropy” was used for the criterion; the number of features was set to 50.

## Convolutional Neural Network

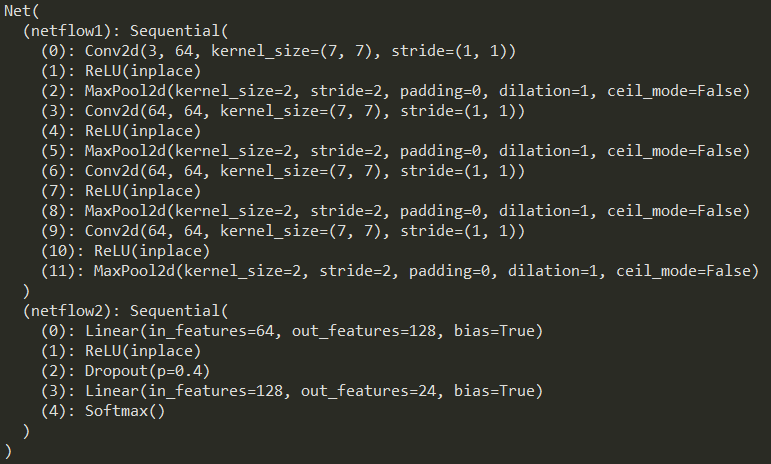


Fig. 4. This is the neural net definition that was used.

TODO: Insert whatever the fuck Nikita did

# Experiments

### discuss how we started with using only our small dataset

### tried most of the algorithms on their default settings

### varied some of the macro parameters such as varying the K value for the KNN or more trees for the random forests

### Pruned the classifiers to use to only just Random Forest and CNN

### Tried to prune our random forest by setting the maximum features parameter

### Then decided to go deeper with the CNN

### Mostly tried to see what some accepted architecture are for this problem

### Found some acceptable architecture to play off of, then played around with the number of epochs, learning rate and batch size to get it right

### MAYBE: if nikita’s stuff works, which is transfer learning

## A. Preliminary Experiments

Before a CNN classifier was decided or even considered. A k-nearest neighbors (KNN) and random forest classifier was used on the small dataset that the team contributed. The labeled data was shuffled and split into a training and testing set. Different combinations were tried where the training comprised of everyone except for one person because that last person’s dataset would be used for testing.

For the KNN, all that was changed was the K-value, leaving the other parameters to their default states and values, as the K-value increased from one, the accuracy decreased significantly. The maximum K-value that was tested was 10 and it gave the lowest results consistently. The random forest on the other hand, showed promised to be the best classifier for the ASL image labeling. As seen in Fig. 3., the random forest reached approximately 70% accuracy when tested on the entire class’ dataset. As mentioned before, the criterion was “entropy”, 50 trees, and a limit on the number of features was 50 to prune the thousands of features that the random forest algorithm deemed important to the classification. The goal of this limit to avoid overfitting and hopefully all the Sobel edge histogram feature to shine. The accuracy seemed to plateau around 50 trees.

## Building Neural Net from Scratch

Beginning with a CNN structure as dictated by a similar problem from a similar project someone has done [1]. This original net does not have weights provided but reached 85.5% accuracy on testing data from the Sign Language MNIST dataset. There are several key differences between this dataset and our dataset, namely number of datapoints, size of the images, and the MNIST dataset is grayscale only. The MNIST dataset has about five times more data, that is four times smaller than our current dataset, additionally, our current dataset has three times the data channels. Iterating through several architecture changes, we found the best results in a slightly deeper net, with larger convolutions at every layer. The goal with these modifications was to better take into account the increased information per image and ignore the increased background noise.

The hyperparameters of learning rate and batch size were also considered with a wide variety tested to find the optimal midpoint between speed and functionality. Learning rates from 0.01 to 0.00001 were tested and 0.0001 was found to be the highest functional value. Batch size was tested between 500 and 64. Smaller batch sizes seem consistently better but going any further would cause the network to train at an unacceptably slow rate. Tests were stopped at 64, which is what is currently used.

The final consideration was for data normalization. Several standard methods were employed, including per pixel mean normalization, min-max normalization, and L2 normalization. All standard methods tested had a strong negative effect on the learning ability of the network and were thus excluded from the final product.

## Fine Tuning CNN

Hello.

A screenshot of a computer

Description automatically generated

Fig. 5. Confusion matrix of resulting outputs

## Transfer Learning CNN

After implementing the CNN organically and testing

various architectures, we found it difficult to pinpoint exact correlations between network architecture and performance. Though some trends between network depth and classifier accuracy were noticed, the micro parameters of the network such as number of convolutions channels at each layer along with corresponding activation functions were problematical to identify. To alleviate these issues and improve overall accuracy of our training algorithm, we decided integrate transfer learning into our machine learning pipeline.

The deep learning architecture of which our design is transferred learned is the ResNets CNN. In general, ResNets address the issue of vanishing gradient which states that as networks become too deep, the model weights in the early layers cannot update correctly due gradient error converging towards zero. However, without a deep network, there may not be enough tunable parameters to tailor a network to a specific problem. With ResNets, this no longer becomes an issue.

In the experiments of our hand designed neural network we noticed that smaller learning rates around 1e-4 were more effective at training the network. Therefore we applied the same rate to our ResNets based system. After five epochs of utilizing the widely accepted one cycle learning policy, the network was completely trained with adequate results.

We tested our CNN against an easier dataset of A and F images in which it yielded an accuracy of 98% and a harder dataset of all the ASL letters in which it yielded an accuracy of 88%.

## Validation & Testing

Number equations con

secutively with equation numbers in parentheses flush with the right margin, as in (1). First use the equation editor to create the equation. Then select the “Equation” markup style. Press the tab key and write the equation number in parentheses. To make your equations more compact, you may use the solidus ( / ), the exp function, or appropriate exponents. Use parentheses to avoid ambiguities in denominators. Punctuate equations when they are part of a sentence, as in

# Conclusion

## A conclusion section is not required. Although a conclusion may review the main points of the paper, do not replicate the abstract as the conclusion. A conclusion might elaborate on the importance of the work or suggest applications and extensions.

Acknowledgment

The preferred spelling of the word “acknowledgment” in American English is without an “e” after the “g.” Use the singular heading even if you have many acknowledgments. Avoid expressions such as “One of us (S.B.A.) would like to thank ... .” Instead, write “F. A. Author thanks ... .” In most cases, sponsor and financial support acknowledgments are placed in the unnumbered footnote on the first page, not here.

References

1. Jain, R. (2019). *Deep learning using sign langugage | Kaggle*. [online] kaggle.com. Available at: https://www.kaggle.com/ranjeetjain3/deep-learning-using-sign-langugage [Accessed 20 Apr. 2019].
2. P. R. Ruiz, “Understanding and visualizing ResNets,” *Towards Data Science*, 08-Oct-2018. [Online]. Available: https://towardsdatascience.com/understanding-and-visualizing-resnets-442284831be8. [Accessed: 20-Apr-2019].

1. Michael Barnard is with the University of Florida, Gainesville, FL 32611 USA (e-mail: michaellbarnard@ufl.edu).

   Nikita Buslov is with the University of Florida, Gainesville, FL 32611 USA (e-mail: author@ boulder.nist.gov).

   Rosemond Fabien is with the University of Florida, Gainesville, FL 32611 USA (e-mail: rosemondfabien@ufl.edu).

   Trung Tran is with the University of Florida, Gainesville, FL 32611 USA (e-mail: tdtran17@ufl.edu). [↑](#footnote-ref-1)