Rock On

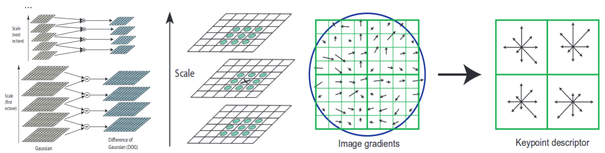
Kappa Krusaderz: Andy Sun, Jacob Patenaude, Paul Westlund, Siddhant Agrawal, Wilson Lee

Simon Fraser University

SIFT - Scale-Invariant Feature Transform

Feature Detection Algorithm by David Lowe

Images are transformed into scale-space by creating image pyramid, at each level, difference-of-Gaussian is found for image.

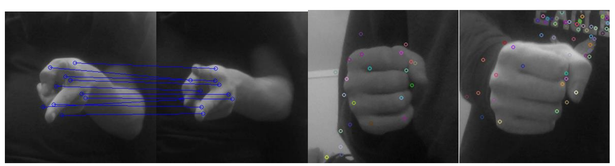


*Images taken from David Lowe's 2004 paper, Distinctive Image Features from Scale-Invariant Keypoints

Using difference-of-Gaussian, keypoints are computed by comparing each pixel to its 26 neighbours to find maxima and minima per level (areas of high curvature or contrast) Keypoint descriptors are calculated by sampling regions around keypoint and calculating image gradient per sample

Regions are then clustered together to form orientation histograms; orientation vector is then normalized along the largest gradient sample to achieve local rotational invariance Keypoints can then be matched based on Euclidean distance of descriptors, allowing for recognition of model objects

Hand gestures such as Rock, Paper, and Scissors have salient curvature-based features for SIFT to pick up; however, in practice there are few keypoints per pose (average ~ 10). To mitigate this issue, images are cropped to minimize background noise causing undesirable keypoints.



- Requires cropped model images, otherwise background keypoints will overwhelm model keypoints
- Brute-force match all test image keypoints against trained keypoints
- Very quick, algorithm capable of real-time matching for our data sets on our testing computer whereas the neural network cannot
- Accuracy of about 55% on validation images

Motivation

To experiment with the effectiveness of convolutional neural networks for the use in classification/object recognition over a traditional method such as SIFT.

Approach

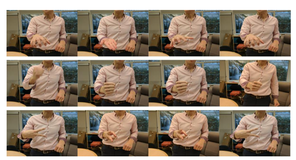
For both SIFT and the Neural Network, we first run back-projection on a skin-tone color histogram to isolate and track the player's hand within the camera, allowing for a more accurate classification. This was a requirement for SIFT, as otherwise its classifications would have been hindered by noise in the background of the images.

While for SIFT we were able to start with an out-of-the-box implementation, our approach for building our convolutional neural network was to start with a baseline network to see what results we could get from the start. From then on we branched out and tweaked with different network structures and hyper-parameters to see how far we could improve it to get better results.

Dataset

Our dataset consisted of 215 training images and 196 validation images. These were separated into classes of Rock, Paper, and Scissors. We then pre-processed the images to isolate for the hand for feeding into both SIFT and the neural networks. For SIFT, the images were checked for keypoints which could be used to identify each of the three gestures.

For the neural networks, we would train for 300 epochs on all training and validation images. We plotted accuracy and tracked benchmarks to see where improvements could be made.

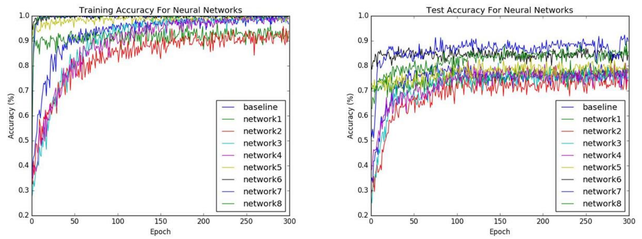


Conclusions

David Lowe's Scale-Invariant Feature Transform feature detection algorithm is a highly performant and quick-to-implement existing approach for the task of gesture/object recognition; however, we found that a sufficiently trained neural network outperformed SIFT on test data by about 30%. If we had collected more data or augmented our data we would likely have seen further improvements in both approaches. Furthermore, if we had a more powerful computer for the real-time capture, we could likely improve the speed of the neural network as well.

Neural Network

We will base our approach by using the pre-trained Inception V3 network as the base for what we will build upon in a technique called transfer learning. This allows us to take advantage of an already powerful fully trained classification model that we can use to retrain for a new set of classes. We can get pretty decent results out of the box this way which will cut down on the amount of work needed to train a completely new network from scratch.



- Because the data was collected manually and in a short timeframe, we encountered overfitting of our training data in all networks
- Classifications tended to be very sharply in favour of a particular class, regardless of whether that class was the correct one or not
- Some networks were redundant, making no improvement over previous iterations
- Training accuracy almost always approached 100%, while validation accuracy oscillated much more and was typically around 80% for most networks.
- The final chosen neural network (network 7) outperformed the SIFT approach in accuracy (90% validation) but was too slow to run in real-time on the capturing computer

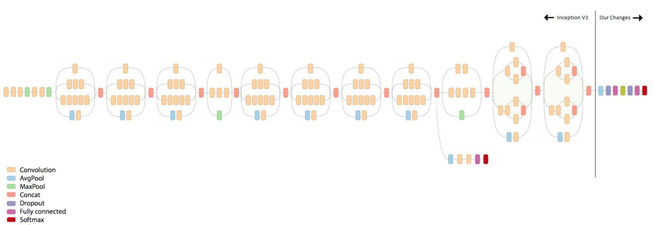


*Image adapted from Inception V3 overview