Machine Learning Engineer Nanodegree

Capstone Proposal

Juan Andrés Ramírez September 15th, 2017

Proposal

Domain Background

Hand gesture recognition is an important research issue because of its extensive applications in virtual reality, sign language recognition, and computer games [1]. Sensors used for hand gesture recognition include wearable sensors such a data gloves and external sensors such as video cameras and depth cameras [2]. Data gloves usually require extensive calibration and restrict natural hand movement [2]. Video-based based approaches addresses this issues but adds other problems, like the hand segmentation from background and oclussion [2]. There are several recent works that uses depth cameras as sensors, but this hardware doesn't have the availability that video cameras have today in people's homes.

Vision-based hand gesture recognition techniques can be divided into two groups: appearance-based approaches and 3D hand model-based approaches. Appearance-based approaches use image features of the hand and compare these parameters with the extracted image features from the input. 3D hand model-based approaches rely on a 3D kinematic hand model and try to estimate the hand parameters by comparison between the input images and possible 2D sinthetic images, generated by the 3D hand model [3].

Problem Statement

The main objective of the proposed work is to create an image classifier capable of detecting the three different hand gestures from the rock-paper-scissors game. This classifier should allow a human user to play this game with a computer using just a screen and a webcam as an interface. To measure the classification performance it is going to be used the accuracy score over a test dataset.

Datasets and Inputs

The proposed classifier will receive a webcam color image as input. The algorithm should work well on different light conditions, postures and backgrounds, so the training and testing datasets should reflect this noisy environments. For this purpose there's the need of using a lot of samples from different sources. The selected databases are: * A subset of the SenseZ3D static hand gesture dataset [4, 5], which contains 30 images of different hand gestures from the same subject in webcam similar conditions. The proposed subset contains just the gestures of Rock-Paper-Scissors (G1, G2 and G5) * A subset from the BochumGesture1998 [6], which contains 3 images per gesture from 19 different subjects. This database just considers hand pictures of size 128x128 pixels (smallest images of all considered databases). * A specially made dataset with the webcam images

of three subjects. This dataset will contain 30 images of each gesture for each subject

Different combinations of this datasets are going to be used for training and testing. This is extensively described on the *Project Design* section.

Solution Statement

The proposed solution is to obtain a classifier by applying *Transfer Learning* to ResNet50 convolutional neural network. The collected rock-paper-scissors training dataset is going to be used to adapt the weights of a fully connected layer that takes its inputs from ResNet50 pre-classification outputs. The resulting algorithm is going to be evaluated as a classification task where the input are the images of people showing hand gestures from rock-paper-scissors game. The accuracy measure is going to be used to evaluate performance over the three different classes of the rock-paper-scissors game.

Benchmark Model

A general hand gesture classifier may be used as a benchmark for this project. This type of classifiers usually work with more than three classes, but they are the closest benchmarking models found in literature.

In the work done over BochumGesture1998 in [6], researchers used a very small training set (images from 3 out of 19 individuals) and a very large testing dataset. They claimed to have achieved 85.8% accuracy on images with complex background when classifiying them into 12 different hand postures. This work seems to be a little bit old and it is not possible to get its source code. However, as most of the recent work is based on range cameras, this research appears to be a good reference and starting point for this project.

Evaluation Metrics

The proposed evaluation metric is the clasiffication accuracy over the test dataset. Depending on the use of the available datasets (see *Project Design*) the main test dataset will be the *Specially Made* testing subset. This subset consists in 30 images per each 3 gestures of one subject, which is previously separated from the training samples.

Project Design

Datasets

Subsets

The following table shows some samples from the considered datasets

As mentioned before, the SenseZ3D and BochumGestures1998 databases aren't specialized on the rock-paper-scissor problem. Some of the gestures are similar but not the same. So there's a possibility that they could impair the classification performance over the *Specially Made* dataset. Also, it will be necessary to enlarge the images of the BochumGestures1998 as ResNet50 requires images of 224x224 pixels and this process will add distortion and noise. Thus, it seems reasonable to train and test with different combinations of sources. This combinations are: 1. Train and test using only the *Specially Made* dataset (training with images from 2 subjects and testing with the third) 2. Train with 2 subjects from *Specially Made* plus the images from SenseZ3D. Testing is made with the remaining subject of the *Specially Made* dataset 3. Train with the dataset mentioned in 2, but adding 3 subjects from BochumGestures1998. With this approach we may obtain a test score on the same validation test used by the benchmark algorithm. Additionally, we may also have a different test score on the testing set used in 1, 2 so we will be able to compare the results with this approaches. 4. Train with all databases but without 1 subject of *Specially Made* dataset and using those images as test

At the end there will be 4 different models. It will be possible to compare just one of this models (3) directly with the benchmarking model (because we don't have access to the code and just have the results showed on their paper).

Preprocessing

- Images from all databases are going to be scaled to fit the input size of ResNet50. Also the preprocessing operations required for this architechture will be applied.
- BochumGestures1998 will have to be transformed from hsi to rgb format

Augmentation

Images from datasets are going to be augmented by means of randomly:

- Rotating images up to 90 degrees
- Horizontally flipping images
- Shifting position by 20%

Training

- A dropout layer will be applied before the fully connected layer to prevent overfitting. The probability of this dropout layer will depend on the training behavior.

References

- [1] Ren, Zhou, et al. "Robust hand gesture recognition with kinect sensor." Proceedings of the 19th ACM international conference on Multimedia. ACM, 2011.
- [2] Suarez, Jesus, and Robin R. Murphy. "Hand gesture recognition with depth images: A review." Ro-man, 2012 IEEE. IEEE, 2012.
- [3] Chen, Qing, Nicolas D. Georganas, and Emil M. Petriu. "Real-time vision-based hand gesture recognition using haar-like features." Instrumentation and Measurement Technology Conference Proceedings, 2007. IMTC 2007. IEEE. IEEE, 2007.
- [4] Minto, L., and P. Zanuttigh. "Exploiting silhouette descriptors and synthetic data for hand gesture recognition." (2015).
- [5] Memo, Alvise, and Pietro Zanuttigh. "Head-mounted gesture controlled interface for human-computer interaction." Multimedia Tools and Applications (2016): 1-27.
- [6] Triesch, Jochen, and Christoph Von Der Malsburg. "A system for person-independent hand posture recognition against complex backgrounds." IEEE Transactions on Pattern Analysis and Machine Intelligence 23.12 (2001): 1449-1453.