The researcher's challenge

Andrei Mihalea

October 2018

1 Introduction

Hand gesture control and recognition is a field which can have a lot of applications and its development can lead to a smoother, easier and more natural interaction between human and machines, which can be used in different scenarios, such as auto industry, sign language recognition and many others. As every popular research topic, which presents interest, it presents many ideas and implementation methods, but nonetheless a lot of challenges and difficulties to surpass, so it is important to define the goals, a clear path to achieve them, and to identify the problems that may appear on the road, such that they can be avoided or easily overcame, without leading to a stagnation in the development process.

2 The plan

2.1 Learning from others

Before deciding on the approach used for implementing the application, it is important to have a background on the already existing methods and to know what are the strong points of various algorithms and ideas, what they do well and what are their drawbacks. When thinking about hand gestures using an RGB camera, the first thing that comes to mind is the how the input will look like and what we can use from it. Most of the methods imply using image and color segmentation, extracting feature vectors from the input image, and then comparing them with a set of labeled images.

2.2 Choosing the method

Even though methods like the one proposed by Azad et al.[1], which uses image segmentation and morphological filtering proved effective in the task of recognizing the gestures used in the sign language, I think gesture recognition is not only about identifying a sign from a static image, but identifying and recognizing a sequence of images, which may represent a single gesture or a chain of gestures which should be interpreted as a single one. For this task, the the most adequate idea I see is using a combination of convolutional long short-term memory neural networks and image segmentation techniques for localizing the position of the hand. The reasoning behind choosing this type of neural networks over classical computer vision methods, or over convolutional neural networks, is the temporal aspect of the inputs, which can't be handled by these algorithms.
The principle of this algorithm is modifying the classic fully connected LSTM, which receives as inputs a unidimensional vector, (which can be the flattened output of a convolutional neural network), into receiving a multidimensional tensor, in our case, the sequence of frames which represent a gesture and instead of doing matrix multiplications at the LSTM cell level, the operations made will be convolutions.
A standard, fully connected LSTM presents 3 gates: input, output and forget and a memory cell. The forget gate has the purpose of deciding what to be kept from the memory cell and transferred to the next cell, while the input gate has two roles: deciding how much and what of the input is combined with the existing memory and the output gate computes the output value of the cell. Each of these gates can be imagined as a single neuron. As stated before, the operations made in the case of, convolutional LSTM nets will be more complicated and more computationally intensive.

The effectiveness of such neural networks has been shown by Xingjian Shi et al.[2] for precipitation forecasting based on a history of images.

2.3 Collect and preprocess data

The first challenge that may arise during the development process is what data to use for the learning process, how and where to collect it from. For starting, the basic idea is to record some simple gestures and try to work with them, to see the behavior of the model. When the behaviour will correspond to the desired one, further records, for more classes of gestures should be added.
A second challenge that may appear in the development of the algorithm is what an input image should look like and in what form it will be passed to the model. There are many options for this decision, from depth images, binary images to the original image captured by the camera. The main challenge faced here is how not to let anything else in the image except for the hand influence the decision that is taken. For example, the brightness may be different from sample to sample, or the background can change, but the gesture may remain the same, or the video quality can vary from camera to camera. The goal is to minimize the impact of such variations in the image, such that the only thing that will have an impact and be relevant to the decision taken will only be the hand.

2.4 Initial tests

When enough data has been acquired and preprocessed, the next step will be running some preliminary tests with this existing data, to see the behaviour of the algorithms implemented and to see what can be improved, removed or added to them, before continuing the development process. These tests should show the direction in which the development should go and may reveal precious information about which is the best way to represent the input and how to better set up the parameters of the model.

2.5 Development process

After the direction of development is clear and the initial tests are finished, it is time to start developing the model for the rest of the cases. This means recording whole new sets of frames with new classes of gestures, then labeling them. As an alternative, or a complementary method, we can also search for already recorded hand gestures and add them to our database.
Other challenges may appear during this process, one of them being the large amount of time that a big neural network needs to be trained. Depending on the amount of data, a training routine may last for hours or even days, so making modification on the model and quickly seeing the result and debugging may be a difficult task and this is why understanding how the model works and how the parameters influence it is very important and why the initial tests should be done beforehand.
Another challenge may be deciding the intention of the user and how to deal with actions that are made accidentally or with a misinterpreted chain of events.
By the end of the six months, a trained model should be developed and able to recognize a set of hand gesture commands from various users in multiple environments and using different input devices.

3 Summary

A research project on a popular topic, with many challenges lies ahead, a project where difficulties may appear at any step. The main mentioned challenges are the dataset acquisition, the extraction of relevant information from the data, dealing with the large training times of the model and the possible difficulty of recognize and classify the gestures in different environments.

References

[1] Reza Azad, Babak Azad, and Iman Tavakoli Kazerooni. Real-time and robust method for hand gesture recognition system based on cross-correlation coefficient. *CoRR*, abs/1408.1759, 2014.

[2] Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. *CoRR*, abs/1506.04214, 2015.