One-shot learning of gestures using a convolutional neural network

Jasper Vaneessen

Supervisor(s): Lionel Pigou

Abstract—**To do**

Keywords— **One-shot learning, convolutional neural networks, gesture recognition, ChaLearn LAP 2014**

I. Introduction

Sign language is the main form of communication in the deaf and hard of hearing community. It is a highly visual-spatial, linguistically complete and natural language. Instead of using acoustically conveyed sound patterns sign language combines hand shapes, orientation and location as well as movement of arms and body. These signs can also be supported by facial expressions.

Some deaf people can read lips and thus understand some of the verbal communication, provided people articulate well and speak slowly. But when they try to convey a message, they have to rely on interpreters or text writing.

Interpreters can be very expensive and their availability is limited while writing down everything you want to say can be very inconvenient and feels limiting. It is very frustrating when your ability to express yourself is dependent of other peoples capabilities or their readiness.

There is also a communication issue among the non-hearing community. Sign language is not at all universal. Different countries have different sign languages and some even have regional ones. So even when someone knows a sign language, this does not mean they can communicate with all deaf people.

II. Methodology

A. Convolutional neural networks

Instead of accounting for all these various edge cases and spending a lot of time on building an optimal feature set we can use another technique. Humans and some animals can recognize objects from a very young age. It is not something we do explicitly, we just see the difference between for example a bicycle and a motorcycle. We have this ability because we learn from previous experiences. Every time humans see a new object they implicitly store features of it in their brain and use these to identify the same object in the future.

There has been a lot of research on the visual cortex of respectively a cat and a monkey. The cortex consists of simple and complex cells. The simple cells perform the feature extraction while the complex cells combine local features from a small spatial neighborhood. This is called spatial pooling and it is crucial to obtain a translation-invariant result.

Convolutional Neural Networks (CNN) try and mimic this principle. They consist of alternating convolutional and pooling layers. The convolutional layer functions as a feature extractor with several feature maps in each layer. Each feature map detects another feature and the amount of feature maps per layer increases the deeper we go into the network. Pooling layers combine these features to a smaller resolution, thus ensuring a translation invariant detection. Using this kind of architecture

Fig. 1. Sample of the ChaLearn, looking at people 2014 dataset

we can feed our network a dataset and let it learn to recognize objects in it. Here the ChaLearn, looking at people 2014 dataset is used. It consists of ten thousand samples of twenty gestures. These gestures are performed by different people in different environments to ensure a general enough result.

The network would learn from these examples as humans do. It looks at an example, tries to predict the gesture and checks its result with the actual gesture. If it is wrong, the network learns from its mistakes and adjusts itself. So all we have to do is let our network learn from a big set of examples and make this learning process optimal. In figure 2 the architecture used in this

Fig. 2. A schema of the used CNN architecture

study is visualized. The network consists of three convolutional and pooling layers and one fully connected or dense layer. This final layer performs the classification of the samples. It takes input from all the feature maps in the third convolutional layer and outputs it to the twenty output nodes of the output layer. Each node represents a gesture which can be recognized.

So at the start of the network, very global and generic features are detected. As we descend into the network all these features get combined and more specific aspects of the image are detected until finally the dense layers makes a decision and outputs the recognized gesture.

III. Conclusion