Hand gesture recognition using convolutional neural network and histogram of oriented gradients features

Alda Kika
Department of Informatics
Faculty of Natural Sciences
University of Tirana
alda.kika@fshn.edu.al

Aldo Koni
Department of Informatics
Faculty of Natural Sciences
University of Tirana
aldo.koni@fshnstudent.info

Abstract

Hand gesture recognition is the core part for building a sign language recognition system for the people with hearing impairment and has a wide application in human computer interaction. The chosen dataset for the construction of the hand gesture recognition system model is fingerspelling alphabet gestures of American sign language. The algorithms that are chosen in this study to create the features of the images that will train the classifier are deep features from a pretrained convolutional neural network AlexNet and histogram of oriented gradients. The feature vectors provided by the extraction methods are used as an input to train support vector machine classifier. Testing results show that the classifiers constructed with two sets of features perform almost with the same accuracy. The combination of histogram of oriented gradient as feature extractor and support vector machine as classifier gives very good results for the classification of images when the dataset of the input is small as in our case.

1. Introduction

Gesture recognition is a very interesting field in computer vision which find practical application in many fields. One of these fields is hand gesture recognition as one of the method used in sign language for non-verbal communication. A hand gesture recognition system provides a natural way of communication for people with hearing impairments and also interactive user friendly way of communication with the computer for the human beings in general.

Convolutional neural network are deep neural networks that recently have reached very high performance in computer vision problems like detection or classification of images. On the other hand handcraft traditional features like histogram of oriented gradients combined with a classifier have resulted also successful in computer vision tasks. Both of these algorithms have been used in sign language hand gesture recognition as in [Ame+17] and [Tav+14].

We have chosen as dataset, Massey Dataset[Bar+11] which is created for American sign language fingerspelling gestures. Pretrained convolutional neural network, Alexnet, and histogram of oriented gradients will be used as feature extractor while support vector machine is chosen as the classifier. In this paper we explore these two methods for feature extraction from a fingerspelling alphabet gesture sign language dataset, compare with each other and discuss the results.

The study is divided into 5 sections. Feature extractors and classification algorithm are discussed in the second section. The dataset is presented in the third section. Experiments and results are discussed in the fourth section. Conclusions are presented in last section.

2 Background

Feature Descriptors

Convolutional neural network are deep learning tools that are very suitable for computer vision taks. They do not only perform classification, but they can also learn to extract features directly from raw images [Siv+12]. They are similar to neural networks because they contain neurons, weights and biases, they have one or more fully connected layers as neural network with many layers have, but differently from them they are easier to be trained because they have less parameters. A very important advantage of using convolutional neural network for computer vision tasks is related to the fact that every layer learns different features of the image. These features can be used to train the classifier.

A convolutional neural network is composed of four different layers [Shoi+16] which are:

Convolutional layer: a set of filters slide on the image. They will be activated when they find the same pattern in it.

Pooling Layer: the aim of this layer is to reduce the dimension of the space, the parameters and the calculations on the net. Several functions can be used but max pooling is more common.
Non-linear Layer: In the architecture of convolutional neural network there are non linear functions like rectifed linear units (RELU), Identity, Tanh, Arctan that have the purpose of introduction of non-linearity in the neural network which will make the training faster and more accurate.
Fully-connected Layer: the neurons in this type of layer connect to every neuron in another layer like in neural networks.
We have used the pretrained AlexNet, deep convolutional neural network, which was used to classify the 1.2 million high-resolution images in the ImageNet LSVRC-2010 contest into the 1000 different classes. The architecture of this network is summarized in Figure 1[Kri+12]. It contains eight learned layers, five convolutional and three fully-connected.

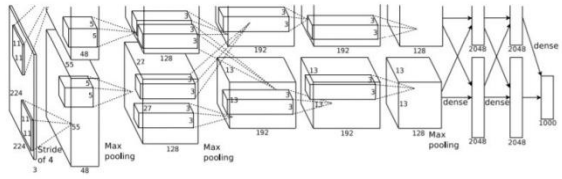

Figure 1: The architecture of AlexNet

Histogram of oriented gradients defined from Dalal dhe Triggs[Dal+05] are the general features in the structure for object dedection and one of the most powerful method for image descriptor. Presentation through HOG has many advantages. Usage of histogram of oriented gradients on the images catches information of local contour like the borders of the structure of gradients. The borders play a very important role in the computer vision tasks and their orientation describe important features for object dedection. Hog uses the borders of the objects to create the feature set that describe the object. In order to calculate Hog descriptors of an image, the image is divided in a number of cells and bins of orientation.

Below some characteristics of each of the methods that we used to extract the features are given.

Convolutional Neural Network:

- Convolutional neural networks are mainly deep learning models which are motivated by the manner that our cornea operate through the alternation of convolutional and pooling layers.
- They are trained feature dedectors making them very adaptable. This is the reason why they reach highest accuracy in image dedection.
- They can learn low level features from training samples as the methods HOG or SIFT do.

Histogram of oriented gradients :

- It is based on first order gradients that are in orientation bins.
- It is dense (it is evaluated in all the image).
- The features extracted from histogram of oriented gradients can't be learned but are hand crafted that means that the information is contained in the image for example in the corners or borders.

Classifier

Support Vector Machines (SVM) presented by Wapnik[VAP98] is one of the most advanced classification method based on machine learning. If we compare it with other classification methods such as decision trees or Bayesian networks it has as advantages higher accuaracy and geometric interpretation. Above all, they do not need a large amount of data for training in order to avoid overfitting [Cam+11]. Support vector machines work well in practice with different types of applications from the dedection of digits, identification of faces, bioinformatics etc.

Classification of the data is a common task in machine learning. The principle of SVM lays in determining the classes to which the data belong. SVM creates a model that delivers new cases to the classes. Training the SVM involves the optimization of a concave function which has a single solution. Other learning paradigms do not provide that the function will be

concave resulting in different solutions depending on initial values for model parameters. The data are saved as kernels which measure the similarity or variability of the objects of data. Kernels can be constructed with different types of objects from continous to discrete data and from sequences to graphical data. In this manner different models of data can be trained with the same model making this approximation very flexible and powerful. Vector support machines are the most known and used method that uses kernels. [Cam+11]

3 Dataset

The chosen dataset is created from Massey University, New Zeland. It contains 2524 images created in such a manner that the hands touches all the borders of the frame. The hands are cropped from original image and placed in a black background. The size of the frame is 500x500 pixels. To construct such a dataset 5 users are used. The hand gestures are based on the american sign language alphabet fingerspelling hand gestures. The main characteristics that distinguish this dataset from other similar datasets are: firstly, the images cover a large variety of hands using different illumination conditions. Secondly, the images are segmented and cropped, but not altered from the original captured images and thirdly, there is no need to use special gloves, or any other apparatus [Bar+11].

In the figure 2 the process of creation of the dataset is shown.

Figure 2: Acquired image with wrist cover. The images are segmented to obtain the final images stored in the dataset

The names of the files follow a simple convention that can easily be used by programmers in their scripts.

For example the convention for the file: handX_G_ILL_seg_crop_R.png is :

- X is the number os the user
- G is the gesture from a to z
- ILL determine the condition of the illumination which can be bot (bottom), top, left, right or diff (diffuse).
- R is the repetition of the gesture.

In the figure 3 the dataset of the data for the american fingerspelling alphabet gestures is presented.

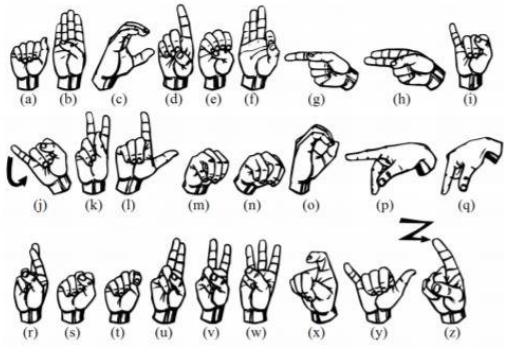

Figure 3: The dataset for the fingerspelling alphabet of american sign language [Asl]

Since two letters "j" and "z" are not static we will remove them from the dataset. The data grouped in 24 classes will serve for the training of the classifier and testing.

4 Experiments and results

Two methods were used to extract the features from the images of the dataset: the pretrained convolutional neural network AlexNet and histogram of oriented gradients. Each feature set is divided in training set and testing set. Two classifiers with each training set are constructed and then tested with the remaining features. The diagram of the experiments is presented in the figure 4.

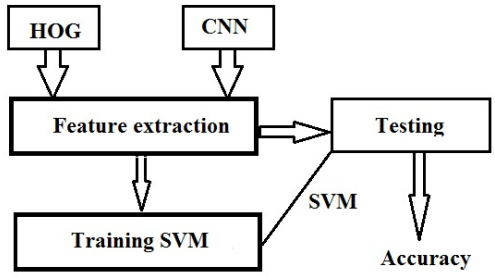

Figure 4: Diagram of the experiments

We will use top-1 and top-5 accuracies. Top-1 accuracy is the conventional accuracy: the answer of the model that has the highest accuracy match the expected answer. Top-5 accuracy means that the expected answer must match one of the model 5 highest probability answers.

The two classifiers were tested using the testing set giving the following results:

Table 1 : The accuracies of the classifiers

Classifier	**Hog**	**Alexnet (CNN)**
Top-1 Accuracy	0.6423	0.6231
Top-5 Accuracy	0.8769	0.8615

The results of the experiments show that the highest accuracy(Top-1 and Top-5) can be reached when the features that are extracted with HOG algorithm are used to train the classifier. Top-5 Accuracy is almost the same with both models.

5 Conclusions

One of the field of machine learning that is giving very good results in complex data analysis is deep learning. Convolutional neural network is a deep neural network that is used in computer vision tasks. We have used a pretrained convolutional neural network and handcraft histogram of oriented gradients to extract the features from a set of hand gesture images of American fingerspelling sign language. The features were used to train a support vector machine classifier. The classifier trained with features extracted with histogram of oriented gradients reaches the highest top-1 and top-5 accuracy.

In the case of convolutional neural network, the number of training sample is very important because it learns from them. We have used the pretrained convolutional neural network, Alexnet, which is trained with millions of images from 1000 different categories which are distinctive among each other while sign languages hand gestures categories have very little difference between them.

Histogram of oriented gradients use predetermined filters while convolutional neural network learn from the training dataset.

Through fine-tuning with a larger sign language dataset the pretrained convolutional neural network will transfer general learned recognition capabilities to specific features of hand gesture classes having more potential for improvement of the results inspiring further research in the future.

References

[Siv+12] M. Sivalingamaiah and B. D. V. Reddy, “Texture segmentation using multichannel Gabor filtering,” *IOSR Journal of Electronics and Communication Engineering*, Vol. 2, pp. 22-26, 2012.

[Shoi+16] Doaa A. Shoieb, Sherin M. Youssef, and Walid M. Aly. Computer-Aided Model for Skin Diagnosis Using Deep Learning. *Journal of Image and Graphics*, Vol. 4, No. 2, pp. 116-121, December 2016. doi: 10.18178/joig.4.2.116-121.

[Li+10] Daoliang Li, Wenzhu Yang, Sile Wang. Classification of foreign fibers in cotton lint using machine vision and multi-class support vector machine. *Comput. Electron. Agric*., 74, 274–279, 2010.

[Vap98] Vladimir N. Vapnik, 1998. *Statistical Learning Theory*. John Wiley & Sons, New York.

[Cam+11] Colin Campbell, Yiming Ying. Learning with Support Vector Machines. *Synthesis Lectures on Artificial Intelligence and Machine Learning*. Morgan & Claypool 2011.

[Dal+05] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. Proc. IEEE Computer Society Conference on Computer Vision and

Pattern Recognition, pp. 886–893, 2005. DOI: 10.1109/cvpr.2005.177. 48, 49

[Bar+11] A.L.C. Barczak, N.H. Reyes, M. Abastillas, A. Piccio and T. SusnjakRes. A New 2D Static Hand Gesture Colour Image Dataset for ASL Gestures. *Lett. Inf. Math. Sci.*, Vol. 15, pp. 12–20, 2011.

[Asl] A. S. L. University. Fingerspelling. ”http://www.lifeprint.com/asl101/fingerspelling/”.

[Kri+12] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*. 25(NIPS’2012), 2012.

[Ame+17] Salem Ameen and Sunil Vadera. A convolutional neural network to classify American Sign Language fingerspelling from depth and colour images. *Expert Systems*, Vol. 34. 2017.

[Tav+14] Neha V. Tavari , A. V. Deorankar. Indian Sign Language Recognition based on Histograms of Oriented Gradient. *International Journal of Computer Science and Information Technologies*, Vol. 5 (3) , 3657-3660, 2014.