BABEŞ-BOLYAI UNIVERSITY CLUJ-NAPOCA
FACULTY OF MATHEMATICS AND COMPUTER SCIENCE
SPECIALIZATION Computer Science in English

DIPLOMA THESIS

AI-Driven Hand Tracking Application for Real-Time Drawing

Supervisor
[Grad, titlu și Tudor-Dan Mihoc]

Author
Sali Arnold

2024

ABSTRACT

Contents

Chapter 1

Introduction

Chapter 2

Related Work

The field of computer vision has evolved from natural vision. Its main purpose is to allow machines to understand visual information based on the natural way humans and other animals gain information from visual input.

This branch of computer science is widely used to gain insight into the real world, through different algorithms and computational techniques, ranging from simpler problems such as object detection in an image or video to more complex ones such as scene understanding and image generation. The solutions from this field open the door for new innovations, which are already present in some capacity in today's society such as a simple social media filter or self-driving cars.

The evolution of computer vision can easily be followed along with the evolution of computational power, given the high requirements of image processing. In the earlier days, such as the 1960s and 1970s, the first algorithms were very limited by the processing power available, but as time passed, more sophisticated ones were created, such as Hough Transform, the Viola-Jones face detection and machine learning.

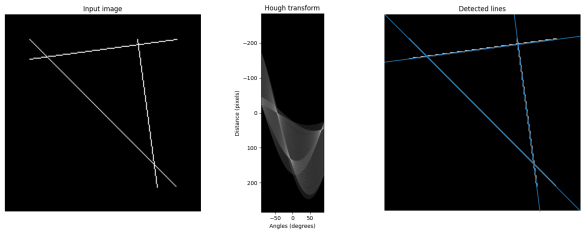

Figure 2.1: Hough transform calculation to find the lines of a triangle [Unk]

2.1 Image segmentation algorithms

2.1.1 Hough Transform

This technique was proposed by Paul Hough in 1962, as a method to identify patterns in images. [DH72]

The Hough Transform is used in computer vision and image processing. The core idea behind it is a voting process made by the curves in the transform space, creating cluster points or local maxima. These points represent the existence of a shape, and since the curves contain the parameters of the shapes, from those parameters the position of the shape in the input image can be detected.

The main strength of this method is that it is capable of identifying shapes even in slightly obscured or noisy data. This made this procedure a massive breakthrough in image processing. Since it's powerful to be able to detect shapes, the method is used in many sectors, from robotic navigation, by identifying edges and lines in the environment, to medical imaging, by detecting circular shapes potentially depicting different biological conditions and many more, where image processing can be utilized.

2.1.2 Viola-Jones Face Detection

Another breakthrough in computer vision was the face detection procedure proposed by Paul Viola and Michael Jones in 2001. [VJ01]

The main idea behind their approach lies in the usage of Haar-like feature and Adaptive Boosting.

The Haar-like features are simple rectangular regions in the input image data. The main goal of using these patterns is first to cut down on the necessary calculations by working with pixels, and second is to get a better understanding about certain regions in the image. By calculating the differences in the sum of pixel intensities between the regions, a contrast can be learned, an incredibly useful information which allows the detection of basic patterns, like edges, lines and textures.

Adaptive boosting is a machine learning algorithm used for boosting weak classifiers into a strong one. The main principle of this method is to iteratively train weak classifier, each focusing on different features. At each iteration pass, the algorithm selects the best performing classifier and merges it with the final strong one, weight proportionally to its performance. It also weighs incorrectly classified instances more for the next iteration for the purpose of shifting the focus on the most challenging parts.

With the help of these two procedures the method of Viola and Jones achieves great accuracy in real-time, more specifically at 15 frames per second on a 700 MHZ

Intel Pentium III, using approximately 50 thousand parameters. Since its low hardware requirements, this way of face detection is still used on very weak devices.

2.2 Deep learning algorithms

With the increase in computational power, image processing algorithms have also evolved in tandem. As seen in the Viola-Jones face detection algorithm, machine learning were already used in the early 2000s, be it at a smaller scale.

Since these algorithms now are able to utilize more resources, they can work with way more parameters, going from thousands to millions, allowing them to learn more complex features than before.

My choice for the following three deep learning architectures is that, they were significant achievements in the field of computer vision, while also resembling my chosen base model the MobilNet.

2.2.1 YOLO

YOLO (You only live once) is an object detection method, based on a Convolutional Neural Network backbone.

The main selling point of this algorithm, is that it achieves real-time performance with a single-pass approach. It divides the image into a grid, each predicting a bounding box, with an associated confidance score and class probability. Utilizing these parameters and various anchor boxes, which are bounding boxes with a predefined shape, size and aspect ratio, the model predicts the final bounding box for the searched cell. [RF18]

Since its real-time performance the method has been utilized in many projects, from autonomous driving, to surveillance systems.

Type	Filters	Size	Output
Convolutional	$64 - 1024$	$3x3/2$	$128x128 - 8x8$
Convolutional	$32 - 512$	$1x1$	
Convolutional	$64 - 1024$	$3x3$	
Residual			$128x128 - 8x8$

Table 2.1: Building blocks for the YoloV3 CNN backbone, Darknet-53 [RF18]

2.2.2 RESNET

ResNet is a convolutional neural network proposed by Kaiming He and co. in 2015, achieving great performances in various computer vision tasks.

The main strength of this model is that it addressed the vanishing gradient problem, via the introduction of skip connections. The residual blocks, the main building blocks of this model, contain the shortcut connections, skipping one or more layers. With this, the network is able to learn residual functions, basically the difference between the desired output and the input data. This feature allows the ResNet architecture to increase the depth of the model, without the gradient vanishing. [HZRS15]

Number of layers	Number of parameters
20	$0.27M$
32	$0.46M$
44	$0.66M$
56	$0.85M$
110	$1.7M$
1202	$19.7M$

Table 2.2: Correlation between the size of the model and the number of parameters used [HZRS15]

2.2.3 MOBILENET

MobileNet is a convolutional neural network designed specifically for mobile and other devices with limited computational power.

At its core the architecture is built using depthwise separable convolutions. The standard convolutions is separated into two separated operations. The first one being a depthwise convolution, in which a single convolution filter is applied for each input channel, capturing features separatly. The second is pointwise operations, which applies a 1x1 filter to the outputs of the the previous calculations, combining the results. This allows the model to significantly reduce the number of parameters needed, compared to normal convolution, thus decreasing the computational power needed. [HZC+17]

Model	ImageNet Accuracy	Parameters
Conv MobileNet	71.7%	$29.3M$
MobileNet	70.6%	$4.2M$

Table 2.3: Difference in parameters between Depthwise separable and full convolutional MobileNet architecture [HZC+17]

Chapter 3

Approaches for better performance with machine learning models

3.1 Size comparisons

3.2 Transfer learning

Chapter 4

Machine learning model specifications

4.1 Modified MobileNet

4.1.1 Transfer learning specifications

4.1.2 Model output

4.2 Data

4.2.1 Data specification

4.2.2 Data preprocessing

Chapter 5

Application structure

5.1 Graphical User Interface

5.2 File Repository

5.3 Image processing

Chapter 6

Performance metrics

6.1 Computing machine specifications

6.1.1 Training specifications

6.1.2 Running specifications

6.2 Training metrics

6.3 Real-time performance

Chapter 7

Future Work

7.1 Machine Learning Model Improvements

7.2 GPU Utilization

Chapter 8

Conclusions

Bibliography

[DH72] Richard O. Duda and Peter E. Hart. Use of the hough transformation to detect lines and curves in pictures. *Commun. ACM*, 15(1):11–15, jan 1972.

[HZC+17] Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications, 2017.

[HZRS15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.

[RF18] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement, 2018.

[Unk] Unknown. Straight line Hough transform. `https://scikit-image.org/docs/stable/auto_examples/edges/plot_line_hough_transform.html`. Online; accessed 23 March 2024.

[VJ01] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, volume 1, pages I–I, 2001.