

# Statistical Hypothesis Testing using CNN Features for Synthesis of Adversarial Counterexamples to Human and Object Detection Vision Systems



**Approved for public release.  
Distribution is unlimited.**

Sunny Raj  
Sumit Jha, PhD  
Laura L. Pullum, DSc  
Arvind Ramanathan, PhD

**May 19, 2017**

## DOCUMENT AVAILABILITY

Reports produced after January 1, 1996, are generally available free via US Department of Energy (DOE) SciTech Connect.

**Website** <http://www.osti.gov/scitech/>

Reports produced before January 1, 1996, may be purchased by members of the public from the following source:

National Technical Information Service  
5285 Port Royal Road  
Springfield, VA 22161  
**Telephone** 703-605-6000 (1-800-553-6847)  
**TDD** 703-487-4639  
**Fax** 703-605-6900  
**E-mail** [info@ntis.gov](mailto:info@ntis.gov)  
**Website** <http://classic.ntis.gov/>

Reports are available to DOE employees, DOE contractors, Energy Technology Data Exchange representatives, and International Nuclear Information System representatives from the following source:

Office of Scientific and Technical Information  
PO Box 62  
Oak Ridge, TN 37831  
**Telephone** 865-576-8401  
**Fax** 865-576-5728  
**E-mail** [reports@osti.gov](mailto:reports@osti.gov)  
**Website** <http://www.osti.gov/contact.html>

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

Computational Science and Mathematics Division

**Statistical Hypothesis Testing using CNN Features for Synthesis of  
Adversarial Counterexamples to Human and Object Detection Vision Systems**

Author(s)

**Sunny Raj, University of Central Florida  
Sumit Kumar Jha, PhD, University of Central Florida  
Laura L. Pullum, DSc, Oak Ridge National Laboratory  
Arvind Ramanathan, PhD, Oak Ridge National Laboratory**

Date Published:  
May 2017

Prepared by  
OAK RIDGE NATIONAL LABORATORY  
Oak Ridge, TN 37831-6283  
managed by  
UT-BATTELLE, LLC  
for the  
US DEPARTMENT OF ENERGY  
under contract DE-AC05-00OR22725

THIS PAGE INTENTIONALLY LEFT BLANK

## CONTENTS

LIST OF FIGURES .....	iv
LIST OF TABLES .....	v
ACRONYMS .....	vi
ABSTRACT .....	1
1. INTRODUCTION .....	1
2. RELATED WORK .....	2
3. STATISTICAL HYPOTHESIS TESTING AND CONVOLUTIONAL NEURAL NETWORKS .....	3
3.1 ERROR MODELS FROM CONVOLUTIONAL NEURAL NETWORKS .....	3
3.2 STATISTICAL HYPOTHESIS TESTING .....	4
4. EVALUATION .....	4
4.1 HUMAN DETECTION WITH OPENCV'S HISTOGRAM OF ORIENTED GRADIENTS .....	5
4.2 OBJECT DETECTION WITH CAFFE CONVOLUTIONAL NEURAL NETWORK .....	7
5. CONCLUSIONS AND FUTURE WORK .....	8
6. REFERENCES .....	8

## LIST OF FIGURES

1.	OpenCV recognizes human beings in images (a)-(c) despite the strong visible distortions in images (b) and (c). .....	2
2.	A typical flow in computer vision detection system involves reducing the image to a set of feature vectors and then using a classifier to render a decision. ....	2
3.	Our approach to testing human detection algorithms using statistical hypothesis testing and the error models derived from a convolutional neural network. ....	3
4.	Visualization of 10 error models used in MAYA. Each pattern was obtained by normalizing the features derived from the Caffe convolutional neural network pre-trained on ImageNet. ....	4
5.	(a)-(e) Original images where OpenCV detects human beings, with (f)-(j) corresponding adversarial counterexample images synthesized by MAYA where OpenCV fails.....	5
6.	MAYA's performance on images from the MIT pedestrian database [12]. ....	7
7.	(a)-(c) Original images where Caffe correctly detects objects beings, with (d)-(f) corresponding adversarial counterexample images synthesized by MAYA where Caffe fails. ....	7

## LIST OF TABLES

1. Adversarial counterexample synthesis against OpenCV's pre-trained HOG based human detection using CNN-based error models and individual pixel perturbations ..... 6

## ACRONYMS

CNN	Convolutional Neural Network
CV	Computer Vision
GPU	Graphics Processing Unit
HOG	Histogram of Gradients
LSVRC	Large Scale Visual Recognition Challenge
ORNL	Oak Ridge National Laboratory
RGB	Red Green Blue
SPRT	Sequential Probability Ratio Test
TNR	Times New Roman
UCF	University of Central Florida



## ABSTRACT

Validating the correctness of human detection vision systems is crucial for safety applications such as pedestrian collision avoidance in autonomous vehicles. The enormous space of possible inputs to such an intelligent system makes it difficult to design test cases for such systems. In this paper, we present our tool MAYA that uses an error model derived from a convolutional neural network (CNN) to explore the space of images similar to a given input image, and then tests the correctness of a given human or object detection system on such perturbed images. We demonstrate the capability of our tool on the pre-trained Histogram-of-Oriented-Gradients (HOG) human detection algorithm implemented in the popular OpenCV toolset and the Caffe object detection system pre-trained on the ImageNet benchmark. Our tool may serve as a testing resource for the designers of intelligent human and object detection systems.

## 1. INTRODUCTION

Computer vision algorithms are increasingly being deployed in cyber-physical systems that directly expose human beings to the real-world consequences of the errors in the design or implementation of such intelligent systems. The expectation of rich dividends from autonomous driving [1] has motivated several car manufacturers (such as Tesla) and public ride-sharing providers (such as Uber) to rapidly explore the market for autonomous vehicles through several innovative offerings. While such innovations are driven mostly by advances in sensor technology and computer vision algorithms supported by the scalability of Moore’s law [2, 3], researchers in safety, formal verification, and cyber-physical systems must discover tools and techniques to test such autonomous self-driving vehicles. Our tool MAYA is an effort towards exploring the correctness of human and object detection systems, and should be of interest to the de-signers and validation teams of intelligent cyber-physical systems such as autonomous cars.

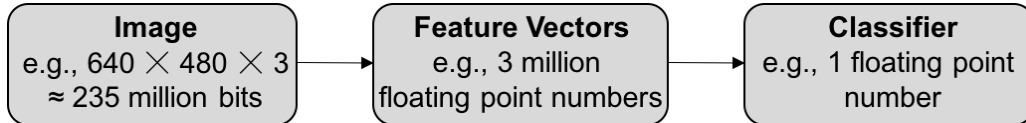
Verifying the correctness of intelligent systems is notoriously difficult due to their probabilistically correct nature, nonlinear computations that deter static analysis efforts, and the high-dimensional nature of their input space. Hence, our tool MAYA uses a combination of statistical hypothesis testing and an error model derived from a convolutional neural network to explore the space of possible inputs to a human or object detection algorithm. Our validation efforts demonstrate that the OpenCV implementation of Histogram-of-Oriented-Gradients (HOG) based human detection [4] is robust to a variety of strong perturbations (See Figure 1). However, MAYA is readily able to successfully generate adversarial counterexamples to OpenCV’s HOG human detection algorithm. We also successfully obtain adversarial counterexamples to the popular Caffe object detection system [5] pre-trained on the ImageNet benchmark.



**Fig. 1. OpenCV recognizes human beings in images (a)-(c) despite the strong visible distortions in images (b) and (c). However, the image (d) synthesized by MAYA using an error model derived from a convolutional neural network is an adversarial counterexample and the OpenCV HOG-based human detection algorithm fails to detect a human being in this image.**

## 2. RELATED WORK

There has been a series of recent investigations [6–9] that demonstrate the susceptibility of deep learning algorithms to adversarial attacks. In some cases, deep learning vision algorithms see shapes and forms of real objects in images that look like pure noise to the human eye. In other cases, deep learning algorithms are deeply affected by small perturbations and produce results not consistent with the human vision system.



**Fig. 2. A typical flow in computer vision detection system involves reducing the image to a set of feature vectors and then using a classifier to render a decision. In convolutional neural networks, the computation of feature vectors and classifiers may be performed by different layers of the same network. Validation of computer vision algorithms is challenged by the size of the input – about 200 million bits for an ordinary image. Even the size of the space of features computed from the image is usually enormous e.g. the number of feature vectors in OpenCV’s Histogram of Oriented Gradients is about 3 million floating point numbers.**

Perturbation of individual bits of the image has been used to generate adversarial counterexamples [10] by exploiting a combination of symbolic and statistical model checking approaches. However, such approaches based on perturbing individual pixels (see Figure 2) are slow and do not exploit information about the structure of the image. Even the insertion of random matrices as errors does not substantially accelerate the search for adversarial counterexamples. In contrast, our tool MAYA uses an error model derived from the Caffe deep learning framework trained on ImageNet and shows an order of magnitude improvement in performance over random perturbations.

### 3. STATISTICAL HYPOTHESIS TESTING & CONVOLUTIONAL NEURAL NETWORKS

An image  $I$  with  $n$  horizontal row of pixels and  $m$  vertical columns of pixels, where each pixel has a depth of  $d$ , naturally corresponds to a tensor of size  $n \times m \times d$ . The space of all possible images can be thought of as members of this set of tensors  $T$ . Given an image  $I \in T$ , a straightforward approach to understand the impact of perturbations on the image  $I$  would be to explore the space of all possible images in an  $\epsilon$ -neighborhood of the image. However, random sampling is not very effective at generating counterexamples to data-rich vision algorithms. Even when an approach based on random pixel-wise perturbations is guided by a fitness metric such as distance from a classification boundary, it is more than 20 times slower than the approach implemented in MAYA (see Table 1 in Section 4.1).

The design of MAYA is illustrated in Figure 3. The user provides an original image, the human or object detection system under test, and Type I error bound [11] on the acceptable error rate from our tool. MAYA stops if it either obtains an adversarial counterexample or enough samples have been observed to meet the expected Type I error bounds.

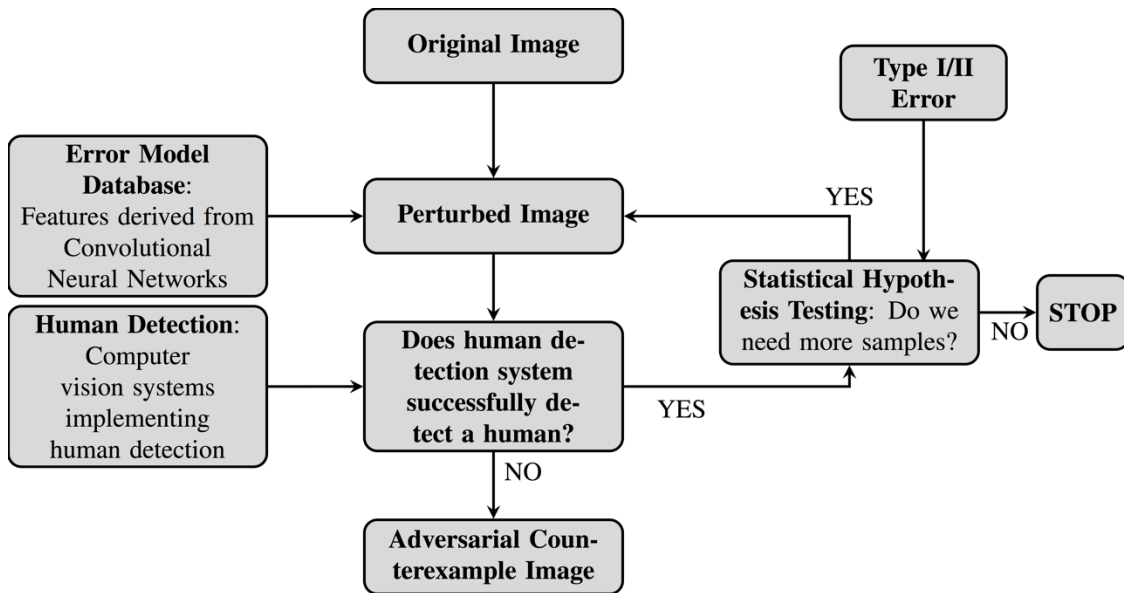
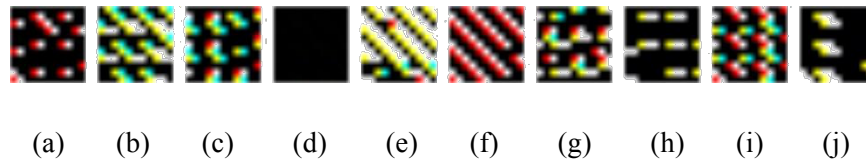


Fig. 3. Our approach to testing human detection algorithms using statistical hypothesis testing and the error models derived from a convolutional neural network.

#### 3.1 ERROR MODELS FROM CONVOLUTIONAL NEURAL NETWORKS

Inspired by the success of convolutional neural networks and earlier results on using individual pixel perturbations to generate counterexamples for machine learning algorithms [10], we build error models (see Figure 4) for perturbing images using the features derived from a convolutional neural network. MAYA explores the space of images using error models derived from the Caffe convolutional neural network trained on the ImageNet database. The lowest layer of Caffe has 96 convolutional filters with

each filter having an input of  $11 \times 11 \times 3$  weights, where  $11 \times 11$  is the pixel width and height of each feature and 3 is the width of the channel corresponding to RGB values. Caffe and other convolutional neural networks internally represent the image using these features. For use in the MAYA tool, each feature or image pattern from the Caffe tool has been modified by re-normalizing the value of each pixel between 0 and a small constant. We emphasize that the image database used to train the convolutional neural network was unrelated to the image database used in our experimental evaluation of OpenCV’s human detection in Section 4.1.



**Fig. 4. Visualization of 10 error models used in MAYA. Each pattern was obtained by normalizing the features derived from the Caffe convolutional neural network pre-trained on ImageNet.**

### 3.2 STATISTICAL HYPOTHESIS TESTING

In our experiments with OpenCV’s HOG human detection algorithm, the MAYA tool always found an adversarial counterexample. However, it is possible that more robust human detection algorithms may not possess any adversarial counterexamples. Hence, MAYA implements Wald’s Sequential Probability Ratio Test (SPRT) [11] to decide the number of samples that the algorithm should observe. Our null hypothesis states that the detection algorithm is no more than 99% correct while the alternate hypothesis states that the detection algorithm is at least 99.9999% correct. Wald’s SPRT does not require the user to make an explicit choice of the prior distribution as it assumes a uniform prior distribution. The algorithm continues to sample until the null hypothesis is rejected with the user-specified bound on the Type I error or an adversarial counterexample is obtained. In our experiments, we chose the Type I error to be 0.00001.

## 4. EVALUATION

We tested the performance of our MAYA tool on OpenCV’s pre-trained implementation of the Histogram of Oriented Gradients (HOG) human detection system and the pre-trained Caffe convolutional neural network for object detection. In our experiments, we used OpenCV version 2.4.12 available from the OpenCV home page <http://opencv.org>. Caffe version 1.0.0-rc3 was obtained from GitHub (<https://github.com/BVLC/caffe>) and its weights were pre-trained using images from the ILSVRC12 data set (<http://www.image-net.org/challenges/LSVRC/2012>) for 310,000 iterations with a batch size of 256. Both computer vision tools are pre-trained popular implementations and worthy of our validation efforts.

#### 4.1 HUMAN DETECTION WITH OPENCV’S HISTOGRAM OF ORIENTED GRADIENTS

The results obtained by the MAYA tool on 5 different images are shown in Figure 5. The top row of the figure shows the original images where OpenCV’s pre-trained HOG implementation can detect humans easily while the bottom row shows perturbed images where the identical HOG human detection implementation suggests that there are no humans in the image. The original images and the adversarial counterexamples look remarkably similar to the human eye and demonstrate the weakness of the OpenCV HOG-based human detection implementation.






The performance of MAYA that uses features from a convolutional neural network (CNN) is compared against an approach based on perturbations of individual pixels in Table 1. It can be readily seen that MAYA needs a few minutes to compute a counterexample on a server with 64GB RAM and 64 AMD Opteron® 6376 2.3 GHz processors. Individual bit perturbation requires about an hour of time on the easiest of these examples. Our CNN features based approach shows a speedup of 24 times or more over the approach based on perturbation of individual pixels. The qualitative nature of these results does not change even when individual bit perturbations are replaced with perturbations using random matrices. In our experiments with images from Table 1, random matrix perturbations needed at least 30 minutes for the synthesis of adversarial counterexamples and sometimes needed more than two hours. Hence, we believe that MAYA’s use of CNN based error models is crucial to its superior experimental performance.



**Fig. 5. (a)-(e) Original Images where OpenCV detects human beings. (f)-(j) Corresponding adversarial counterexample images synthesized by MAYA where OpenCV fails.**

We also studied the performance of MAYA on images from the MIT pedestrian database [12]. Figure 6 shows a plot of the number of images vs. the time taken by MAYA to synthesize an adversarial counterexample for these images. In this plot, the runtime varies from 1 second to about 8.5 minutes.

**Table 1. Adversarial counterexample synthesis against OpenCV's pre-trained HOG based human detection using CNN-based error models and individual pixel perturbations.**

Image	Random Error Model		CNN Error Model		Speedup
	Time Taken (seconds)	Number of Perturbations	Time Taken (seconds)	Number of Perturbations	
	7,842.58	39,806	209.54	1,055	<b>37.43</b>
	>9,864.51	>50,000	180.70	914	<b>54.59</b>
	6,726.06	34,249	273.99	1,385	<b>24.55</b>
	3,531.93	18,180	95.86	479	<b>36.84</b>
	7,140.23	36,295	171.00	856	<b>41.76</b>

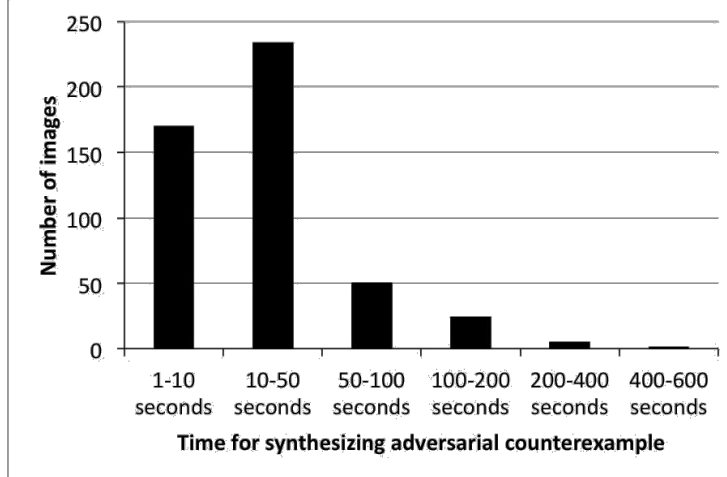


Fig. 6. MAYA's performance on images from the MIT pedestrian database [12].

## 4.2 OBJECT DETECTION WITH CAFFE CONVOLUTIONAL NEURAL NETWORK

Caffe is a deep neural network framework that can be used to classify an image into 1000 categories corresponding to the ImageNet Large Scale Visual Recognition Challenge dataset [13]. The output layer is a softmax function that gives confidence values for each label between a range of 0 to 1. To generate adversarial counterexample images, we computed the difference in the values of the highest confidence prediction and the second-highest confidence prediction, and used this difference as a fitness measure. Figures 7(a), 7(b) and 7(c) show three original images that Caffe correctly classifies as a Rottweiler dog, a tabby cat and a rocking chair respectively. Figures 7(d), 7(e) and 7(f) show the corresponding perturbed adversarial counterexamples generated by MAYA and Caffe fails to correctly recognize the objects in these three images. MAYA performs the adversarial synthesis computation within 4303.0, 496.96 and 360.36 seconds on an Intel i-7-4770k processor with a NVIDIA GTX780 GPU co-processor.

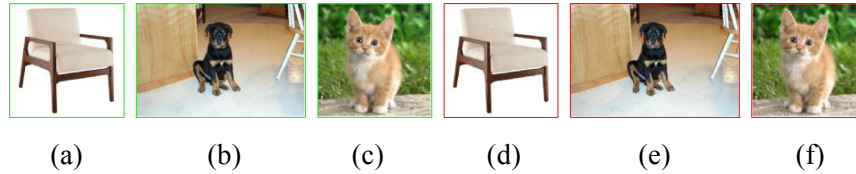


Fig. 7. (a)-(c) Original Images where Caffe correctly detects objects beings. (d)-(f) Corresponding adversarial counterexample images synthesized by MAYA where Caffe fails.

## 5. CONCLUSION AND FUTURE WORK

In this paper, we introduce a new tool MAYA for synthesizing adversarial counterexamples to human and object detection systems using features derived from convolutional neural networks. We show that MAYA is effective in synthesizing adversarial counterexamples against the pre-trained HOG-based human detection algorithm implemented in OpenCV and the pre-trained convolutional neural network Caffe.

MAYA terminates as soon as it determines the first adversarial image to an object detection framework. Hence, the adversarial image produced by MAYA lies very close to the boundary between correctly classified images and incorrectly classified images; hence, such an image can be correctly recognized by well-crafted ensemble sampling methods. However, MAYA can also be used to search for an adversarial image against object recognition systems implementing ensemble sampling. We will continue to investigate the performance of MAYA on recently released and upcoming object detection systems that implement defenses against adversarial attacks.

Adversarial attacks do not necessarily correspond to inputs that would be naturally acquired by machine learning systems in the wild. For example, some attacks may generate images where some color values of some pixels are floating point numbers while a camera always returns integers for all pixels. MAYA voids such obvious physical infeasibility by saving every image as a portable network graphic and then re-acquiring this saved image using the computer vision system. However, printing the image on an ordinary 600 dpi printer and then re-acquiring it using a 600 dpi Canon PIXMA MX340 scanner destroys the adversarial nature of the synthesized image. We will continue to investigate the synthesis of adversarial images that can be acquired using a camera by exploiting Generative Adversarial Networks (GANs) that can synthesize artificial images indistinguishable from images acquired using natural sources. Future versions of MAYA will also focus on parallelizing our computations, performing real-time adversarial synthesis for videos, and on developing counter-defenses.

## 6. REFERENCES

1. Sebastian Thrun. Toward robotic cars. *Communications of the ACM*, 53(4):99–106, 2010.
2. Chris Mack. The multiple lives of Moore’s law. *IEEE Spectrum*, 52(4):31–31, 2015.
3. Igor L Markov. Limits on fundamental limits to computation. *Nature*, 512(7513):147–154, 2014.
4. Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 886–893. IEEE, 2005.
5. Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 675–678. ACM, 2014.
6. Nicolas Papernot, Patrick McDaniel, Somesh Jha, Matt Fredrikson, Z Berkay Celik, and Ananthram Swami. The limitations of deep learning in adversarial settings. In *Security and Privacy (EuroS&P), 2016 IEEE European Symposium on*, pages 372–387. IEEE, 2016.
7. Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.



8. Anh Nguyen, Jason Yosinski, and Jeff Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 427–436. IEEE, 2015.
9. Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial examples in the physical world. *arXiv preprint arXiv:1607.02533*, 2016.
10. Arvind Ramanathan, Laura L Pullum, Faraz Hussain, Dwaipayan Chakrabarty, and Sumit Kumar Jha. Integrating symbolic and statistical methods for testing intelligent systems: Applications to machine learning and computer vision. In *2016 Design, Automation & Test in Europe Conference & Exhibition (DATE)*, pages 786–791. IEEE, 2016.
11. Bhaskar Kumar Ghosh and Pranab Kumar Sen. *Handbook of sequential analysis*. CRC Press, 1991.
12. Constantine Papageorgiou and Tomaso Poggio. A trainable system for object detection. *International Journal of Computer Vision*, 38(1):15–33, 2000.
13. Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. IEEE, 2009.