Haar Cascades – Assignment Report

BM40A0800
MACHINE VISION AND DIGITAL IMAGE ANALYSIS

Authors:
Tikhon Belousko (0458150)
Iakov Lushin (0458163)

07.04.2016

Abstract

This report considers Haar cascades, which are used in the algorithm by Viola and Jones for face detection. The detector, consisting of Haar-like features, utilizing AdaBoost classifiers and incorporating attentional cascades, is described in detail. Application of the algorithm is demonstrated using two examples: detection of faces and detection of road signs. A brief description of other face detection algorithms, that utilize some ideas of the Viola and Jones method, is given. The main conclusion of the work is that the original algorithm by Viola and Jones is significantly outdated and uses significant amount of simplifications. However, more modern and advanced methods are still based on that algorithm and utilize its main ideas and concepts.

Table of Contents

1 Introduction

This work is related to the Haar Cascades which main application area is face detection. The main advantage of the detector was the ability to be rapidly evaluated on average by utilizing simple Haar-like features and a cascaded structure. The paper will describe the method in detail, give demonstrations of its applications. Comparison with other methods, related to face detection, will also be given.

The rest of the paper is organised in the following way. First, the theoretical explanation of the method is given, describing general ideas as well as emphasizing some important details. Next, demonstrations of the method are presented by applying it to the tasks of face detection and road sign detection. This is followed by the comparison of the considered method with some of its extensions and alternatives. The main conclusions are made in the final section of the report.

2 Theoretical Background

2.1 Problem Statement

The main application area of the Haar Cascades algorithm, authored by Viola and Jones [1, 2], is face detection. The same detector, trained with appropriate data, can be applied to other detection problems as well.

Object detection process is usually performed by scanning the original image with a detector. The detector classifies each subimage it processes into two categories, indicating the presence or absence of object in the subwindow. Thus, the detector is performing binary classification in each subwindows. Subwindows are usually chosen in such a way, that they cover all or the most of possible locations and scales of objects in an image. Some preprocessing might also be applied to the original image as to make the detection process more robust to some noise or distortions, such as illumination variance. The process of scanning an image with a subwindow is shown on the Figure 1.

Figure 1. Detection window scanning an image.

2.2 General Detector Structure

Classification task, based on images, might consist a relatively complicated task. Performance of the task in each position and scale of the subwindow can be costly. Therefore, Viola and Jones proposed a cascaded detector structure. It is meant to lessen the average cost of the detector evaluation by gradually narrowing the focus of the detector on the subwindows of more interest (i.e. having more probability of containing object in it).

The detector cascade consists of several classifiers of different complexity. The simpler classifiers, which can be evaluated faster, are meant to discard some proportion of the subwindows, not containing an object, such as face, at low cost. Next, more complex and costly classifiers will be evaluated only on the rest of the subwindows, further narrowing the attention of the detector. This process will continue, until the last classifier in the cascade will be evaluated on the most interesting subwindows, thus providing final detection results.

Each of the classifiers in the cascade utilizes a number of simple Haar-like features. Any such feature can be cheaply computed in constant time, thanks to the concept of Integral Images, proposed by Viola and Jones. The complexity of classifiers is thus determined only by the number of features used in it. Selection of the features and tuning of the parameters in each classifier is performed with the use of Adaptive Boost (AdaBoost) [3] algorithm. All of the elements of such a cascade detector will be described in more detail in the following sections.

2.3 Haar-like Features

The face detection algorithm, proposed by Viola and Jones, utilizes features, that have resemblance with Haar wavelet, which is demonstrated on the Figure 2. The main property of those wavelets is their square-shape, which is derived into the rectangular features used in the algorithm. Those features are calculated as a difference of summed pixels in two, three or four rectangular areas in the image. These rectangles are demonstrated in the Figure 3. On that image, areas, that lie under white rectangles are taken positive in the sum,

and the ones under black rectangles are taken as negative. These kind of simple features are aimed to detect regions of different intensity and edges between them.

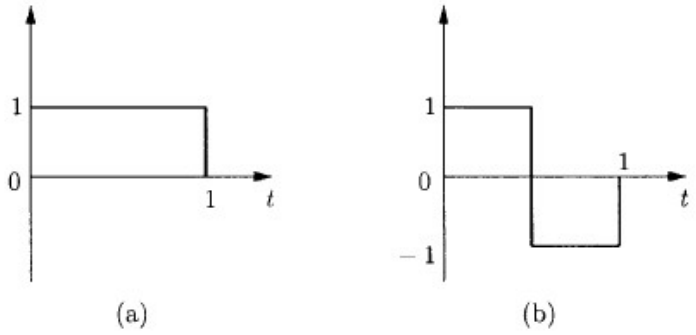

Figure 2. The Haar Basis: a) Scaling function; b) Wavelet [4].

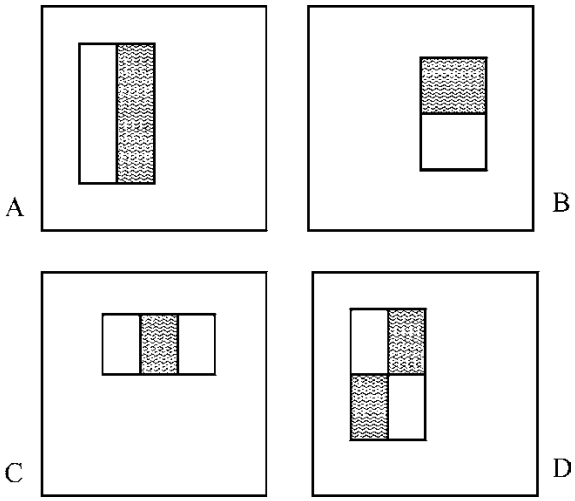

Figure 3. Example rectangle features shown relative to the enclosing detection window. The sum of the pixels which lie within the white rectangles are subtracted from the sum of pixels in the grey rectangles. Two-rectangle features are shown in (A) and (B). Figure (C) shows a three-rectangle feature, and (D) a four-rectangle feature [2].

The benefit of usage of such simple features is the effectiveness, with which they can be calculated. Viola and Jones showed, that with the use of the proposed Integral Images, such features can be computed in constant time with only 6 to 9 memory references on any scale or location.

Integral Image is a matrix of the same size as the original image. The value of such a matrix in each position is equal to the sum of pixels of the original image, that lie higher and to the left of the considered position or with the same vertical or horizontal position. This is demonstrated on the Figure 4. Integral image can be computed from the original image in one pass, using simple recurrent expressions.

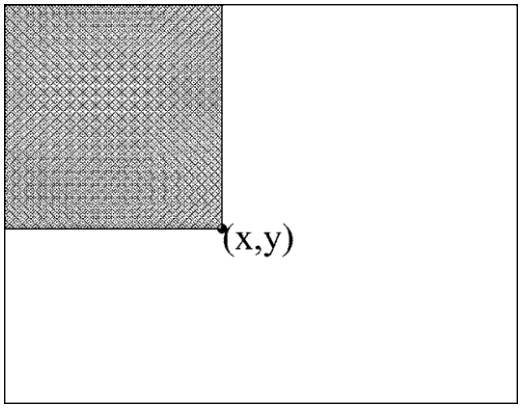

Figure 4. The value of the integral image at point (x, y) is the sum of all the pixels above and to the left [2].

It can be easily shown, that the rectangular features, that Viola and Jones use in their algorithm, can be calculated as a simple summation and subtraction of values of integral image in the locations of corners of the rectangles. Example demonstration of the calculation of the sum of values in a rectangle with the use of integral image is shown on the Figure 5.

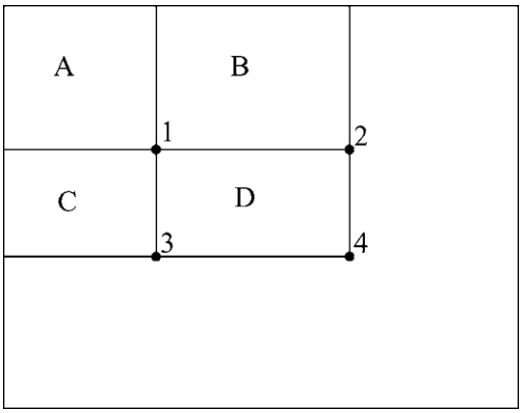

Figure 5. The sum of the pixels within rectangle D can be computed with four array references. The value of the integral image at location 1 is the sum of the pixels in rectangle A. The value at location 2 is A + B, at location 3 is A + C, and at location 4 is A + B + C + D. The sum within D can be computed as 4 + 1 − (2 + 3) [2].

Each rectangular Haar-like feature itself is too weak to be considered a good detector. But the rapid calculation of each such feature allows for their effective combination into more complex and strong classifiers. Viola and Jones achieved it by utilization of Adaptive Boost method, which will be described in the next subsection.

2.4 AdaBoost Classifiers

The set of rectangular features, that can be extracted from a subwindow is very large. Some selection of features, that would be best for the classification process is needed. Viola and Jones proposed the usage of Adaptive Boost (or its modification) method for that purpose.

AdaBoost algorithm constructs a strong classifier from a set of simple classifiers. In the case of the rectangular features the classifier is a binary threshold classifier, based on the feature

output. The construction process starts with one classifier, that demonstrated the best overall result. The next classifier is chosen in such a way, that it has the best classification result on those samples, which the first simple classifier couldn't classify correctly. Their weighted output constitutes a stronger classifier, which can correctly classify a wider range of samples. Subsequent addition of other simple classifiers to that construction in the same way (i.e. its boosting) allows for reaching the desired error rate.

The features selected by AdaBoost algorithm seem to be interpretable. They detect some main patterns that usually appear in people's faces. Such patterns might be the difference between intensities of the eye area and cheeks, or differences of the nose area. These example features are presented on Figure 6.

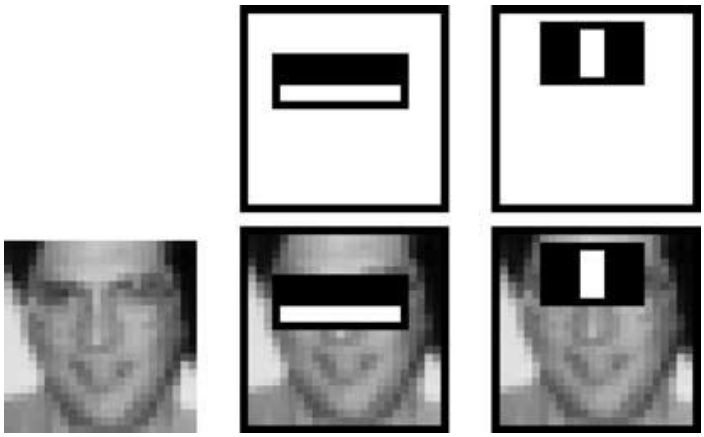

Figure 6. The first and second features selected by AdaBoost. The two features are shown in the top row and then overlayed on a typical training face in the bottom row. The first feature measures the difference in intensity between the region of the eyes and a region across the upper cheeks. The feature capitalizes on the observation that the eye region is often darker than the cheeks. The second feature compares the intensities in the eye regions to the intensity across the bridge of the nose [2].

The problem with strong classifiers constructed with the described procedure is that they might have to contain a large number of features to perform accurate classification. Thus, they may become expensive in terms of evaluation. This problem is addressed by Viola and Jones by incorporating a cascaded design to their detector, which will be elaborated on in the next section.

2.5 Attentional Cascades

Evaluation of strong classifiers, consisting of numerous features, can be costly, when performed on the subwindows in each position and scale. To increase the average speed of the detector evaluation, Viola and Jones proposed to use cascaded structure. It consists of several classifiers, constructed in the same way with AdaBoost algorithm from the same general set of features, but having different complexity. This means different targeting classification accuracy and therefore different number of features.

The classifiers of various strength and computational cost are combined in a cascade. First, the simplest and cheapest classifier is evaluated, sifting a part of subwindows, that are most probable to have an object of interest in it. The result is still inaccurate, and the next

classifier in the cascade is evaluated on that smaller set of subwindows, further narrowing it. Each next classifier in the cascade is evaluated only if the previous one has been passed. If a subwindow was discarded on any stage of the cascade process, it is not processed anymore. The final result is positive only in the case of passing of all the classifiers in the detector's cascade. The cascade detection process is illustrated by the Figure 7.

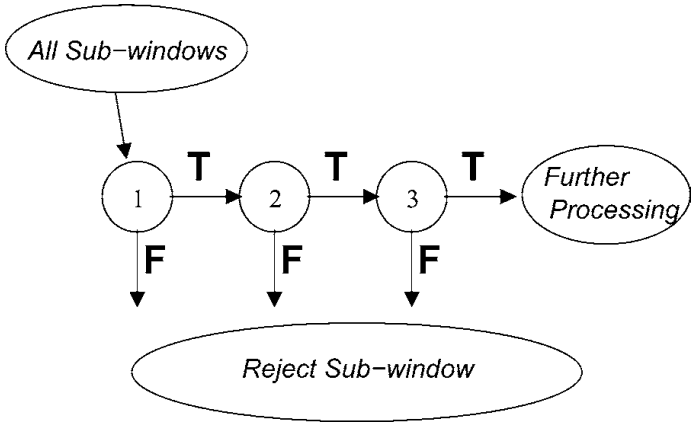

Figure 7. Schematic depiction of a the detection cascade. A series of classifiers are applied to every sub-window. The initial classifier eliminates a large number of negative examples with very little processing. Subsequent layers eliminate additional negatives but require additional computation. After several stages of processing the number of sub-windows have been reduced radically. Further processing can take any form such as additional stages of the cascade (as in this detection system) or an alternative detection system [2].

Such a detection strategy brings a significant boost in the performance. If the first classifiers are trained in a proper way, they will discard a considerable amount of subwindows, thus lessening the number of evaluations of stronger classifiers, which allow recognition of more subtle characteristics of objects. The average number of feature calculations per subwindow is thereby reduced drastically.

The cascade detector has some specifics in the way it should be trained. They are linked to the training of each layer itself and in the relation to other layers. Those specifics will be discussed in the next section.

2.6 Training Procedure

The special characteristic of a cascade detector is that each its layer aims to discard as many subwindows as possible, while keeping all or most of the subwindows, that really contain the objects. That shifts the targeting performance of classifier to getting the number of false negatives close to zero in addition to just lowering the error rate. Thus, the number of false positives might get relatively high. However, these cases will be dealt with by next levels of cascade, as opposed to those already discarded. The shift in the destination functioned is said by Viola and Jones to be easily obtained by shifting the threshold of the AdaBoost binary classifier.

To further increase the effectiveness of sifting of false positives, the training of subsequent layers is performed on different training sets. The first classifier in a cascade utilizes the whole training set (or its random subset) in the training process, thus learning some general

features of objects. The second layer, however, doesn't need to discard subwindows based on general object features. It will receive only those samples, that passed the first classifier. Thus, to train the second layer to act the most effective way on those samples, it is given the set of the negative examples in the training set consisting only of the false positives of the first layer. All the subsequent layers are trained using the false positives of previous partial cascades as their negative samples.

The number of features in each cascade layer and the number of layers themselves are obtained in the training process based on the targeting requirements of the system. Thus, the features are added to the AdaBoost classifier, until it reaches the targeting values of detection and false positives rate. The layers are added to the detector, until it reaches the desired overall effectiveness. Those values are chosen manually, though, since the task of searching for the optimal architecture, minimizing both error rate and processing time, is too complex.

3 Application Demonstrations

To demonstrate the work of the algorithm two experiments were conducted. The first experiment is a real-time face detection, which was meant to prove the ability to rapidly detect faces using images captured from a webcam. The second experiment is a road sign detection which demonstrates the training procedure.

3.1 Real-time Face Detection

Face detection was performed utilizing system object `vision.CascadeObjectDetector` from MATLAB Computer Vision Package. This detector is pre-trained for face detection task by default. For image capturing the Webcam USB addon for MATLAB was utilized. The overall performance of the algorithm is sufficient to recognize faces on images of size 384x216 with frame rate of 11-13 frames per second (FPS). However, for higher resolution it does not suit well due to low FPS. Another problem of the algorithm is its inability to detect faces that are rotated by a certain angle. The demonstration of that is represented on the Figure 8.

Figure 8. Demonstration of the real-time face detection: left image – detected correctly, right image – failed to detect due to face rotation.

3.2 Road Sign Detection

To demonstrate the training process of the `vision.CascadeObjectDetector` the task of road sign detection was chosen, due to the fact that MATLAB standard package has the database of images with positive and negative samples for this task. For the training purposes the false alarm rate was set to 0.1 and the number of cascades was set to 5. With this set of parameters the detector showed better reliability in detection than with other tested values. The result of correct detection is illustrated on the Figure 9.

Figure 9. Road sign detection.

4 Comparison to Other Methods

The original algorithm by Viola and Jones seems not to be the state-of-the-art at the moment. However, it is often referenced, when new face detection methods need to be evaluated. Some of those methods extend the ideas of Viola and Jones, some utilize a part of them, some have little resemblance with it.

4.1 Extensions of Viola-Jones Algorithm

One of the main limitations of the features of the algorithm by Viola and Jones is the lack of rectangle rotations. Some direct modifications of the algorithm aim to extend the set of original features by the rotated ones. Thus, Lienhart et. al. [5] proposed using rectangles, rotated by 45 degrees in addition to those utilized by Viola and Jones. These can also be rapidly computed in constant time with the use of Rotated Summed Area Table (RSAT), an analogue for Integral Image in the rotated case. The proposed extended set of features also includes center-surround features (rectangle inside another rectangle). Leinhart et al. utilize as well a post-optimization technique in order to tune parameters of classifiers on different stages. These improvements significantly increase the effectiveness of the algorithm without much loss in speed.

Further extension of the feature set was proposed by Messom et al. [6]. It adds some more rotations of rectangles to the set, which can be computed in constant time with the use of Integral Images, as preciously. Although, increased amount of rotations may bring additional overhead in the training and performance. This extension is motivated by the need to detect

other types of objects, such as hand gestures, and might not be necessary in the narrower face detection domain.

4.2 Other Face Detection Methods

Some other algorithms utilize a basic cascade approach to the face detection. Bourdev et al. [7] generalize a cascade detector architecture. They introduce the concept of Soft Cascades, in order to ease the training process and enable selection of fewer features. They also overcome some other disadvantages of the original cascades, such as the rejection of a sample by one cascade stage without taking into account how well that sample performed in previous stages.

Li et al. [8] substituted Haar-like features for SURF-based features (Speeded Up Robust Features). This, combined with the proposed AUC learning criterion (Area under ROC (Receiver Operating Characteristic) curve), allowed faster training with selection of fewer cascade stages.

The approach of Jun et al. [9] consisted in the utilization of Local Gradient Patterns (LGP) as features in the detection cascade, which added invariance to local intensity changes, such as shadows or glasses. Their Evidence Accumulation Method (EAM), which is the step of combining responses from several closely located subwindows, significantly reduces false positives rate and enables the usage of less cascade stages.

These methods are comparable in the effectiveness and processing speed and outperform original Viola and Jones algorithm. They may be considered as state-of-the-art in face detection domain.

5 Conclusions

This work described the face detection algorithm by Viola and Jones. It utilizes Haar-like features, incorporated into classifiers of different complexities. These classifiers are constructed and trained by the AdaBoost algorithm and are combined into the attentional cascade in order to simplify the detector evaluation on average. The algorithm has a property of being fast, while retaining considerable performance.

The original method by Viola and Jones becomes outdated nowadays. However, the main ideas and concepts, such as attentional cascade structure of the detector, had a high impact on the later algorithms and the general framework is still utilized. Improvements are made to different parts of the algorithm, allowing to reach higher speed and detection performance.

References

[1] 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, 2001. CVPR 2001, 2001, vol. 1, pp. I–511–I–518 vol.1.

[2] P. Viola and M. J. Jones, “Robust Real-Time Face Detection,” International Journal of Computer Vision, vol. 57, no. 2, pp. 137–154, May 2004.

[3] Y. Freund and R. E. Schapire, “A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting,” Journal of Computer and System Sciences, vol. 55, no. 1, pp. 119–139, Aug. 1997.

[4] P. Kechichian, “On the Partial Haar Dual Adaptive Filter for Sparse Echo Cancellation,” M.E. thesis, McGill University, Canada, 2006.

[5] R. Lienhart and J. Maydt, “An extended set of Haar-like features for rapid object detection,” in 2002 International Conference on Image Processing. 2002. Proceedings, 2002, vol. 1, pp. I–900–I–903 vol.1.

[6] C. Messom and A. Barczak, “Fast and efficient rotated haar-like features using rotated integral images,” presented at the Proceedings of the 2006 Australasian Conference on Robotics and Automation, ACRA 2006, 2006.

[7] L. Bourdev and J. Brandt, “Robust object detection via soft cascade,” in IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005. CVPR 2005, 2005, vol. 2, pp. 236–243 vol. 2.

[8] J. Li, T. Wang, and Y. Zhang, “Face detection using SURF cascade,” in 2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops), 2011, pp. 2183–2190.

[9] B. Jun and D. Kim, “Robust face detection using local gradient patterns and evidence accumulation,” Pattern Recognition, vol. 45, no. 9, pp. 3304–3316, Sep. 2012.