CS34110 Assignment 1: Can you use Viola-Jones face detection for counting people?

Craig Heptinstall (Crh13/ 110005643)
Computer Science Dept.
Aberystwyth University

Novemeber 24, 2014

Abstract

This article, regarding to the module CS34110 (Computer Vision), contains an analysis and critique of a scenario around the topic of face detection using means of a vision technique as outlined in a scientific paper by Paul Viola and Michael Jones (Robust real-time face detection) [1]. This article will critique the paper, as well as look at how effective the methods described could be used in the scenario of 'people counting'.

1. Introduction

As a brief overview of the major points that will appear in this article, they will be as follows:

- A brief description of the problem as outlined in both the paper containing the Viola-Jones technique, and also in relation to this assignment.
- The context of the problem, including reasoning to why this a worthwhile problem to solve.
- A brief resolution description of the issues earlier described.

1.1. Overview of problem the paper tackles

Paul Viola and Michael Jones explain in their paper how their means of face-recognition allows 'Robust Real-time ' [1] face detection that can be used in videos, due to the quick processing of the algorithm and techniques used. The problem this paper relates to however, is a combination of this, and also the accuracy of face detection. Another part of the problem included in this paper is the way the classifiers for basing detection on is achieved. Having a less 'trained' classifier can reduce the accuracy of detection in latter stages. In addition to this, and placing this alongside the assignment scenario, the problem is extended to how this technique could be used to perform detection of people, in order to count how many are for instance in attendance of a concert etc. In general terms, the main issue here is developing a working face detection system that is able to count people at a great accuracy and not to over or under-count amounts of people.

1.2. Context of the problem

Looking more in-depth at the context of this problem, there is a great deal of applications that the solution of this problem could help towards. For instance, as outlined in the assignment scenario, one example could be for safety reasons. By having a well-functioning people counting system, it would mean that venue organises would know when the place is overcrowded, for instance. For purposes like counting attendees of a travel trip or tour, by creating a solution to provide fast and reliable face-detection could mean the ability for organisations to charge more accurate prices for groups. scientific paper however this is broader to solving the issue of the speed and accuracy of general face-detections. By doing this, it means that processing more faces in a short amount of time via a live video is possible.

1.3. A glance at the proposed solution

The solution-system provided in the paper consists of three main parts: **1.** An image representation system that allows features used by the detector to be sorted and checked quickly, **2.** An efficient classifier built from a large amount of images, and also with the help of a technology known as AdaBoost (Freund and Schapire, 1995), **3.** A method that concentrates on the more promising areas of images, in order to find detections quicker.
In essence, what the method(algorithm) aims to do is to search out face-like features, placing these into rectangles, and these rectangles are made of more rectangles, which consist of features such as eyes, mouth, nose etc. These

are usually stored in an XML file, containing a long variety of 'Threshold values' which then decide if a rectangles contents are above a certain number (more is explained on this in the next section). If the threshold value is met, the algorithm then decides it has found a feature and moves on [2]. This is repeated with a wide variety of rectangle sizes, and scales of the image. Through the means of this algorithm, it should be possible to find faces quicker, and also more reliably enabling the use of real-time video with this technique.

2. Critique of proposed method

This section of the report provides an analysis of the face detection described in the paper [1] and highlights the success and failure of such a method in this category. To give a more comprehensive analysis, there is also a range of tests included from my own tests of the algorithm through the use of OpenCV [3].

2.1. Context and explanation of method

Before evaluating the results of the face detector in both my test and the papers instances, I will first give an overview of my understanding for the way the detector works. To begin with, an image classifier was first created using a variant of AdaBoost or Adaptive Boost (Freund and Schapire, 1995), which involved modifying the conventional AdaBoost procedure to allow for both training of the classifier and selecting features.
The original Adaboost as developed by the previously mentioned works by combining a set of weak, simple classifiers [4] eventually resulting in a strong or final classifier. The way that Viola and Jones use this method is to allow the algorithm to search over a set of possible perceptions in a stronger classifier, and use the strongest result (in terms of lowest classification error) for a sequence of learning problems. This is repeated a good number of times to produce several weak classifiers. After this, the examples used to train these classifiers are re-weighted, resulting in the weak classifiers with low perception rates are removed until only one strong classifier remains. The result of this being that an exponential growth in accuracy happens as the method is repeated that removes weaker classifiers.
Continuing from this, an attention cascade is also developed to allow a boost to performance and speed of detection. This is basically performed by only aiming at promising regions of an image, and then focusing on these to reduce overall time of searching.

2.2. Negatives of proposed method

One of the first clear failure modes I found when coming across this algorithm, in both the paper and performing my own tests is the way that it can struggle with doing full profiles of faces. To expand, if an image was at a certain angle, or in more extreme cases facing to the side, then detection might not work. This is pointed out in the paper [1] where at the end it specifies that the detector encounters problems finding faces with more than 15 degree tilt or 45 degrees rotation.
In addition to this, by utilising OpenCV [3] (which primarily uses the same method of detection as described in the paper) I performed some tests to test the rotation issue- See figure 1. Along with the tests reflecting the issue with rotations of faces, I also found that any obstructions to the mouth or eyes significantly reduced accuracy. This ranged from glasses to facial hair- See figure 2.

Figure 1. Example detection in OpenCV, note that some faces are tilted so the detection has not worked. (I was impressed with how the blurry face in the background was detected though.)

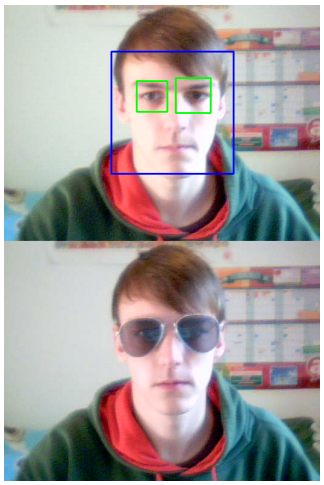

Figure 2. With objects such as sunglasses, detection is not possible because the detector cannot find any difference in pixel values around the eyes.

There is also the issue of false positives, where I found

that some places where faces were not present still were detected by the OpenCV library- See table 1. This occurred when a collection of shapes appeared to make a face, though it would be clear to humans that it was not- See figure 3.
The way the paper describes tackling this issue is by checking areas with overlapping rectangles until a threshold is met, though without a very well trained classifier this may not be accurate to the degree the paper claims (1 in 14084) in terms of false positives. Taking into account the test set they used was large frontal face images, it does not demonstrate the rate in real life scenarios. The paper does then go onto talk about how having a higher threshold would decrease false positives, though also reduce detection rate. This an important finding into calculating thresholds correctly.

Figure 3. With everyone in the photo not facing forward(frontal faces), the detector has failed, though an anomaly has happened where a part of the image has been identified as a face, though this is a false positive.

Test Scenario	False Positive avg. rate
Viola-Jones	20%
My OpenCV tests	40%

Table 1. Based on both tests using 20-Feature classifiers, over 50 images

Another failure mode I found during my own tests and also from sources such as the interview article [2] is the rate of time to get the best test pattern when using AdaBoost to analyse images. In order to train a classifier well, it requires many images (Viola and Jones used 15000 in their paper) causing a very long training time. Analysing images in these quantities (varying on computer processing speed) can take several days to perform well. Again, this could be an issue in real world circumstances.
A final point in terms of negatives for this method comes to how out-dated this seems now in terms of newer and improved algorithms. For instance, there are now face detection algorithms using PCA (Principle component analysis) and ICA (Independent component recognition) which are used in software such as Apples iPhoto, and in Googles PittPatt whose detection rates are vastly improving with systems based from these means achieving average rates of 96% detection [5]

2.3. Positives of proposed method

Amongst the positive points for the Viola-Jones face detection method are most importantly the great results in terms of the true positive rates. In both the papers initial and real world scenario tests, the Viola-Jones method received 94% and 79& respectively. This is a great step up from previous detection methods proceeding this. In my own tests using the OpenCV platform and using a small test set using an online database [6] (20 images)- See figure 4, I received a true positive rate of 90%. In addition to the results from the paper, it also mentions a simple voting scheme further improves results [1] which does this by combining a number of detectors, and allowing them to essentially vote if they detect. If the number of detectors agree for that area outweighs false detections then this usually increases accuracy of the result. I find this to be a very good method of finalising results as well as eliminating any false positives.

Figure 4. OpenCV has identified every face in this photo perfectly, with a true positive rate of 100%.

Although I mentioned in the fall-downs of the paper the speed to train the classifiers could be an issue, the speed in which the classifiers identify faces is very good. For instance, this detection rate is 15 times faster than the

Rowley-Baluja-Kanade detector-1998. The paper also mentions the speed at which is can process images at sizes of around 384 by 288 pixels- 0.67 seconds. This would be just enough for real time detection, though to strengthen this, the tests in the paper was done some time ago where processors were considerably less powerful than today. Having tested with a web-cam myself using OpenCV again, recognition and tracking works much more seamlessly now. Another positive (though this can be reflected as a negative too) is the way the detector works when it comes to recognising faces. The paper mentions how it can fail with clouded faces, but that when mouths are covered the detector will still work. This gives the detector a little bit of flexibility, in that not all the face has to be shown in order for detections to occur. On the downside of this though, if the eyes are obstructed at all it will not be able to find faces. I have found the detector to be too dependent and temperamental.

A very important success of the paper is the use of the integral-image image representation. It means that features can be measured in constant time, giving this means of representation a great advantage over more sophisticated alternatives. To keep this brief, this works by calculating the sum of values in a rectangle grid covering features of a face. It provides a fast way of calculating the sum of the pixels by looking to the top and to the left values of the rectangle. In this case it means that for looking for eyes, it can use a three-rectangle feature (eye, nose, eye) can be calculated in eight array references [7].

A final point I would like to highlight here of a positive note is the types of classifiers used during the papers tests. For classifiers using the cascade architecture mentioned in the paper, these can be faster to learn than single strong classifiers. This is due to rather than having one single classifier trained for a long time, many smaller and weaker classifiers are combined. This does not only reduce training time, but also increases the accuracy of said classifiers.

3. Application of the proposed method to the scenario

Now I have evaluated the Viola-Jones face detection method, I can look at how this could be possibly applied to the scenario described in the introduction stage of this assignment. Considering all the failure modes of the algorithm too can allow me to propose a solution that can avoid these issues and work more effectively.

The main practicalities for using the algorithm described in the paper [1] could be safety (not allowing a location to become too overcrowded), organisation (performing head counts to organise coaches etc.) or for convenience to people queuing (in order to reduce queues by splitting them up). By using the face detection technique described the paper I have analysed, this real-time functionality of counting people is now possible.

Starting with the issues that I highlighted in the previous section, where if a person's eyes were covered in some way (e.g. glasses), this could be turned into a more useful means of detecting a person. Instead of the detection method searching for face directly, it could search for a combination of objects such as certain types of clothing like hats, collars and glasses. By combining this with the usual detection criteria, the ability to count people could be improved. It would also mean that weaker classifiers could be used, each for searching different objects. The voting system described in the paper could then be used to get an average to produce the best results.

This brings me onto the classifiers themselves, to which they could be trained to side profiles as well as frontal face. This would be useful in situations such as where people are entering or leaving an auditorium where not everyone may be facing the front as they walk in. Of course an issue with this practicality is the amount of processing required, as the paper mentions the algorithm only being able to process a face in 0.67 seconds so having to search with more than one position of a face would take longer. This is not such a problem now though considering that processing power has improved vastly since the paper was submitted.

It is important for the detectors to be very well trained too, to ensure maximum accuracy. With more advanced detectors could hurt the possibility of real-time tracking, though as one paper [8] looked at the possibility of taking images every 5 minutes, evaluating them and producing averages could be an option to resolve this. This would mean that rather than looking directly at entrances for people entering/ exiting events or other locations the whole area could be photographed. This may not be entirely possible though in some locations where it is simply too large or inconveniently shaped.

Objects being mistakenly detected can also be an issue here, such as paintings or statues for instance. There would be ways to avoid these problems though, such as the case of background subtraction. By getting an image of the background whilst the venue is empty, then by removing this background whilst detecting could avoid such problems. For the case of moving images whilst detecting, a Gaussian average could be used.

This brings me to the next key point to consider when using an algorithm and techniques described above in order to perform some kind of counting mechanism, which is quantity and placement of cameras. Although it is key for cameras to cover as many people as possible to gain the most accurate counts, it is also important that different cameras do not count the same person twice, resulting in incorrect numbers. Again, depending on the shape of venue, it could simply be done by covering all the people with multiple cameras and getting an average, or by splitting up

the venue into sections and adding numbers.
Along with camera placement, the lighting conditions can affect results, so where images are converted to grey scale like the paper describes, this might not always work for darker areas of a dark hall, meaning that the integral image method may not always work. Here such devices using infrared may be a solution, meaning that the differences in edges and colour would appear more prominent. In paper [9], this looks at the use of NIR (near infrared LEDs) with the use of filters to both cut out background noise and also to ensure prominent dark and light areas for detection.
There is also additional techniques that could be considered whereas to avoid counting a person twice some face recognition is used. There are clear issues with this though, such as accuracy of recognition and also the speed at which the processing would increase vastly.
Considering all these points, and also considering the algorithm in the paper, I think overall the detection rates would be quite high, depending on the type of location on venue. As I highlighted earlier, depending on the location, objects or animals could affect false positives rates. In my opinion, using the Viola-Jones algorithm heavily depends on the cascade used, and the training set also. In the case of the papers cascade file, because of the vast amount of images used, this could be a good starting point to use in most situations.

4. Conclusion

Now I have evaluated the Viol-Jones algorithm, I can provide my conclusions to looking into this subject and also the possible uses of this with scenarios outlined in the assessment. I will also be providing a short self-evaluation and a grade I believe this paper meets.
Starting with the Viola-Jones algorithm, I found out how the main functionality of this worked, splitting the method into its three integral parts: An image representation system, an efficient classifier, and a method concentrates more promising areas of images. A key part of the algorithm I would like to re-highlight here is the integral image use, in which this paper was the first real use of this method in computer vision. Integral image is used in this algorithm to calculate the sum of pixels in certain areas of images (by means of multiplying the upper and left pixel values together), to allow for the method of working out the difference between alongside-areas in terms of pixel value. This then allows for objects such as faces to be found.
The way Viola-Jones trained their classifier was also a major point of the paper, by using a set of weak classifiers and boosting these by the AdaBoost method. This allowed them to create a stronger classifier than they could in the same space of time for a single strong classifier.
After analysing and giving a brief explanation of this method, I went onto providing a range of positives and negatives about the paper and also the method the paper explains.
To highlight some of the negatives, these included:

- Issues with none-frontal faces, e.g. Side profiles. This was due to the face detector used in the paper was only trained on images of frontal faces.
- Obstructions to eyes could reduce accuracy upon performing my own tests- This was due to the classifier used being only trained to faces without glasses.
- False positives was an issue under my own tests, where real world images were used- due to things appearing like faces to the detector.
- The time required to train a cascade file in my opinion was quite long- would require at least a day of processing in order to create a good strong classifier.

Similarly, I found a set of successes in the paper:

- The clear high level of detection- 94% in their set of test images, and 90% in my own tests.
- The speed of detections is very fast- 15 times faster than the predecessor to this technique, meaning real-time detection is possible.
- The combination of weaker classifiers using AdaBoost produced a stronger detector than a single strong one.
- The integral image method used within this algorithm contributed hugely to the increase in speed of the algorithm.

Once I looked at both the up and down-sides of this method of face detection, I then used these to analyse how the methods presented in the Viola-Jones paper could be used for real time people counting scenarios. Along with these methods, I also looked at a range of possible enhancements and additions to the method such as:

1. Dealing with obstructions in faces (e.g. glasses, hats)- but then using this to a possible advantageous means, by training detectors for objects such as the previous to identify people.
2. Not only using the classifiers such as the paper described, but a combined set to allow for side and possible full profiles.
3. Rather than using real-time, use images taken in intervals of five minutes or so, to allow more accurate detection because of a longer time to evaluate each detector.

4. Where objects in the background are mistakenly counted as foreground, a form of background subtraction like Gaussian average could be used to ensure objects are not detected as faces.

5. The correct use of detector cameras is vital in counting people, such as placement, covering the whole scene (e.g. a large auditorium). Schemes like the voting system mentioned in the paper would be useful in getting a good average count using multiple detectors.

6. Infrared use with the cameras could solve a lot of issues when it comes to dark areas in the detection zone, to ensure facial features are highlighted properly.

7. I also mentioned the use of newer techniques such as PCA as used by Apple and Facebook could be considered.

4.1. Self-evaluation

Now for the final section of this report, I can give a brief self-evaluation. To provide a score I think is subject to my assignment solution (from section 5.7 of the students handbook[1]), I will review my understanding and suggestions from the Viola-Jones paper [1].I will then provide my personal opinion into the score I suggest for each section of the report, as broken down in the assignment brief.

Starting with the introduction, I feel I produced a good description of what the assignment was asking for, and gave some examples to why the face detection method might be used in the context of the assignment scenarios. I also gave a good description into the proposed solution from Viola-Jones, including a description of how the algorithm worked in its simplest terms. **Mark: 16/20**.

In the critique of the method in the paper, I felt I expanded on the algorithm more at the beginning of the section, though could have expanded more. The weighting of certain sections held me to keep it brief though. When it came to providing a set of successes and failure-modes, I feel I provided and expanded on an extensive list of the most important of each. Again, I could have expanded more, but the key was to not over-run the section. I also think if I had more time I could have tried out training my own classifier. **Mark: 23/30**.

For the section considering the application of the algorithm to some scenarios described earlier, I think although I described a good mixture of uses with the algorithm and external affects such as cameras, again I could have expanded on some points such as more detail towards edge detection perhaps. **Mark: 16/25**.

As for the final three sections, my conclusion briefly summed up the main points of the paper and summarised on my findings and suggestions **Mark: 7/10**. This section I feel gave a comprehensive breakdown of my efforts in the paper- **Mark: 4/5**, and my references section followed the IEEE referencing convention **Mark: 9/10**.

Generally, and with the help of a few tutorials and also by playing with OpenCV myself to test some of the techniques myself, I feel I have learnt a vast amount this area of vision, and found it equally interesting as important in today's uses of CCTV systems for example. The maths in the paper could appear confusing at certain stages of the assignment, but by sticking with tutorials helped me greatly. By combining the scores I have evaluated, I would score myself **75/100** in the assignment solution, the equivalent of a **1st** in the students handbook.

[1] http://www.aber.ac.uk/ dcswww/Dept/Teaching/Handbook/handbook.pdf

References

[1] Paul Viola and Michael Jones. Robust real-time face detection. 2004. International Journal of Computer Vision, 57(2):137-154.

[2] http://makematics.com/research/viola-jones/ Greg Borenstein. Adam harvey explains viola-jones face detection. 2012. Viola-Jones explanation of alrgorithm in an interview.

[3] http://www.opencv.org/ OpenCV.

[4] Jiri Matas and http://www.robots.ox.ac.uk/~az/lectures/cv/adaboost_matas.pdf Jan S ochman. Adaboost explanation. 2012.

[5] Marian Stewart Bartlett J. Ross Beveridgea http://www.cc.gatech.edu/~isbell/reading/papers/draper_cviu03.pdf Bruce A. Draper, Kyungim Baek. Recognizing faces with pca and ica. 2003.

[6] https://support.bioid.com/About/BioID-Face-Database BioID. Boid face database. 2014. Collection of images used for testing Viola-Jones with openCV.

[7] http://user.engineering.uiowa.edu/~image/BOOK-for-Lectures/Boosting-Slides-Milan-shb-2010.pdf Milan Sonka. Explanation of integral image- lecture slides. 2010.

[8] Shraddha Chitalia Vidya Zope ttp://research.ijcaonline.org/volume84/number17/pxc3892949.pdf Ajit Naik, Divyesh Darde. Rcrowd management using viola jones algorithm and heuristic data mining. 2013.

[9] ShengCai Liao XiangXin Zhu RuFeng Chu Meng Ao Ran He http://www.ics.uci.edu/~xzhu/paper/NIR-FG06.pdf Stan Z. Li, Lun Zhang. A near-infrared image based face recognition system. 2011.

All images used in this report are my own and do not break any copywrite rules.