Can you use Viola-Jones face detection for counting people?

Samuel Jackson
Department of Computer Science
University Of Aberystwyth
Aberystwyth, Ceredigion, SY23 3FL
Email: slj11@aber.ac.uk

I. Introduction

This paper examines the Viola-Jones face detection algorithm [1] to the application of counting people in the real world. The Viola-Jones algorithm presented in [1] proposes an efficient method of performing face detection at frame rate with reasonably accurate results.

Detecting faces is a major challenge in the field of computer vision, and one which has many practical applications. Accurately determining whether a scene or photograph contains a face is provides many challenges. Yang et al. [2] lists a number of different factors affecting human face detection. Faces typically have a wide degree of variation. They have different prominent facial features, skin colour, scars, facial hair, and hair styles. The pose also causes issues. The side of a face is very different from the frontal view. Occlusions prove another major obstacle. In a crowded scene it is easy for a face to be partially or wholly excluded, glasses and sun glasses are also a common source of occlusions. The quality of the capture including the lighting, focus, and contrast can affect the image quality. The characteristics of the camera such as the lenses and sensors also affect final capture.

Another major factor which increases the difficulty of face detection is the complexity of the problem. Photographs and video generally have an incredibly large feature space to search making looking for the face at every pixel and scale an impossibly complex task to practically accomplish at frame rate.

Despite these obstacles efficient detection of faces in a scene continues to be desirable for a multitude of practical applications. One of the most common uses of a face detection system is as part of a larger system, such as a face recognition or image analysis. In order to correctly recognise a face it must first be found. In this way face detection algorithms act as a preprocessor to a larger system. Other applications include uses in image databases, social media, and video conferencing. Automatic detection of faces can be used in image management system to find faces in images as part of a precursor to automatically annotating images with meta data. The most obvious example would be the image tagging feature on sites such as Facebook. Applications to video conferencing include the need to adjust the focus of the camera relative to the speak.

The approach taken by Viola-Jones in [1], is also not restricted to detecting faces. A cascade classifier such as the one presented in [1] can be trained to detect a variety of objects other objects provided that sufficient positive and negative examples can be supplied for training. For example, work based on the Viola-Jones detection algorithm has been used to detect people [3], license plates [4], and facial features (eyes, nose, and mouth) [5].

The major contribution that the Viola-Jones algorithm makes is in the speed at which the algorithm can detect a face. Many approaches prior to Viola-Jones could offer reasonably good detection rates [6] [7] [8], but the Viola-Jones algorithm's strength is in the speed ups offered by their method.

Viola-Jones solves the problem of rapid object detection through the use of three key components. The First component is that the Haar features used to detect faces are computed using an integral image representation. This representation allows for rapid evaluation of Haar features in constant time. The second component is to use the AdaBoost algorithm to choose a small number of features that classify the training set well. The third component is to combine the weak classifiers into a cascade of classifier, with the least complex classifiers at the beginning of the cascade. This ensures that the majority of negative examples are rejected early using classifiers that are quick to evaluate, but that later, more complex nodes in the classifier minimise the occurrence of false positives. Once the cascade of classifiers has been trained it is very efficient to classify new examples by running the image through the cascade.

The rest of this paper is dedicated to reviewing the performance of Viola-Jones object detection in various circumstances and discussing its application to the titular problem. Section II reviews the what problems Viola-Jones can be used to solve and what it cannot solve. Section III discusses how the technique could be applied to the titular problem and what issues the like issues encountered would be. Section IV rounds off the the paper with conclusions drawn from the preceding sections. Finally, section V provides my self evaluation of this paper.

II. Critique of Method

Viola-Jones object detection represents a major breakthrough in the computer vision community by bringing practical speed ups in face detection through intelligent data structures and clever use of the AdaBoost algorithm to make face detection in real time a possibility. However, success of the technique is not without its limitations and the technique has several possible failure modes.

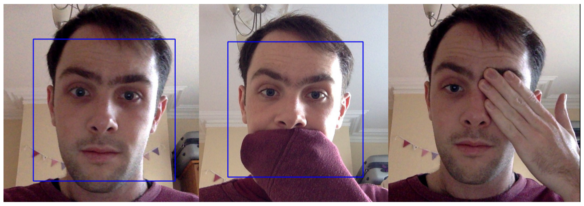

Fig. 1. Example of face detection using the Haar cascade in OpenCV. Occlusions of non essential areas (such as the mouth in the middle image) still result in a positive match, while occlusions of essential features such as the eyes or nose result in a false negative.

At the time of publication of the technique had results that had "detection and false positive rates which [were] equivalent to the best published results" [1] but executed in in a much smaller amount of time. Comparisons between the major approaches published at the time are given at the end of the paper, but the results are not entirely definitive due to most approaches not providing a full receiver operating characteristic (ROC) curves for their classifiers or did not publish there best results. In the case of Sung and Poggio [6] only the MIT part of the common dataset was used (as the CMU dataset did not yet exist). In spite of this, the authors show that their technique provides roughly equivalent levels of performance given the available data. The authors note that the best performance was achieved using a voting regime between three classifiers trained with different negative examples, different negative vs. positive errors, and different criteria for also positives vs. classifier size.

Unfortunately the Viola-Jones has quite a few failure modes in which it will give a high number of either false positives or false negatives. One of the biggest drawbacks of the vanilla implementation presented in the paper is that it can only detect frontal images of faces. Faces viewed for the side or in a different rotation or orientation are unlikely to be correctly detected. The technique can be used to correct detect faces as viewed from the side profile by training a different classifier with sufficient positive and negative examples. Both sides of the face could be accounted for by training a classifier for one side of the face, testing the image then flipping the image and running the classifier again to account for the other side. Similar problems occur when the classifier is presented with faces that are rotated, either in or out of the image plane.

These problems are largely caused by the techniques dependence on such basic features. This leads to the classic time versus complexity trade off where in order for the algorithm to perform fast enough to run practically at real time, the features need to be quick to evaluate. The basic set of features used by Viola-Jones are limited in the amount of domain knowledge they can encode in the learning procedure. The limited range of vertical and horizontal features fail to encode information about diagonal features, such as might be present in a face rotated 45 degrees out of the image plane.

Jones and Viola [9] attempted to alleviate these issues by training multiple classifiers for each of the poses and using additional diagonal Haar like features. The choice of classifier used is decided using a decision tree. This gives a two step approach to classification in which the first step estimates the pose and the second runs selected pose classifier.

Further work by Lienhart et al. [10] provides empirical evidence that better results can be obtained from the Viola-Jones approach though the use of a number of extensions such using the gentle AdaBoost algorithm in favour of the original discrete AdaBoost implementation. They also performed experiments with a broader set of Haar features which were formed from rotated versions of the original features. These appeared to produce more accurate results despite complicating the learning procedure by successfully encoding more domain knowledge into the learning procedure that it would otherwise have difficulty learning. More recent work published by Brubaker et al. [11] also showed that it is possible to improve the speed of face detection as well as reducing the training time without sacrificing accuracy.

Another key failure mode with the Viola-Jones algorithm is its robustness to occlusions. This largely depends on what is occluded in the image. The two key features identified during the training of the Viola-Jones technique show that a vertical rectangular feature over the nose and a horizontal rectangular feature over the eyes are the most successful detection features for a face. This implies that Viola-Jones has some robustness against partial occlusions of non-essential areas such as covering the mouth in the middle image of figure 1 and will still return a positive match. However, occlusions of prominent features used to classify face will result in a false negative. Tweaking the parameters of the cascade can help to increase the detection rate of the occluded image, but this has the disadvantage of subsequently increasing the false positive rate. Lin et al. [12] have experimented with additional robustness against occlusions through by a two step process consisting of reinforcement training with bootstrapping and k-means clustering to reduce the false positive rate and training a selection of "occlusion cascades" which can be checked against to in the case the regular cascade returns a negative match.

The success of the Haar cascade technique is also heavily affected by the quality of the dataset used during training. Careful preprocessing is required to ensure that all of the positive examples used in training are correctly aligned and are of a similar scale. A large number of negative examples also need to be prepared to provided sufficient training against false positives. However, it is also essential that the dataset used in training contains enough variance. As mentioned in section I human faces contain a large amount of variation. A good dataset should therefore capture a high amount of variation while still being well formed. In the dataset used in [1] used 4916 hand labelled images randomly taken from the web. These naturally contained a high degree of variation. Some examples had moustaches and facial hair, others had glasses or sunglasses. Different ethnicities and skin tone is included as re different lighting conditions. This subsequently leads to there being enough variation on the dataset to ensure that the classifier learns the common features shared between faces and therefore generalises well.

One final limitation with the Haar cascade method is the amount of time that it takes to train. This is not as much of a concern as other limitations because the important point about the Haar cascade approach is that fact that once a classifier has been trained, it is fast enough to execute at frame rate. However, the original implementation of Viola-Jones took weeks to train. Exploitation of parallelism [1] and subsequent increases in computing power have brought the training time down, but training still remains a considerable task especially when this is combined with the amount of time required to pre-process a good dataset for use in training.

Despite these limitations the Haar cascade approach face detection remains one of the most popular approaches in the field. This is largely due to the fact that the method runs fast enough to be used at frame rate and therefore has a very wide variety of applications but does not need to sacrifice accuracy in order to achieve it. The framework presented in Ref. [1] is also not limited to the detection of faces. Relevant Haar features can be computed to be representative of a number of different objects, although the low complexity of these features restrict the classifier to working best with rigid body objects. The classifier also provides a small amount of robustness to partial occlusions of an object, but will fail if core features used in detection are occluded. A good dataset, with both positive and negative examples, is required to effectively train a Haar cascade but providing a decent dataset can be acquired the classifier will work well enough for many practical purposes.

III. Application of Method

This section analyses the potential suitability of applying the titular method to counting people. This is a very broad question, the answer to which is largely dependant on the specifics of the application. Counting people is useful in a wide variety of circumstances. Safety applications are the among the most obvious. With automated people counting systems we could avoid disasters such as the one which occurred at the Hillsborough stadium in 1989. Customer service and management is another practical use for such a system. Knowing how many people are queueing at a checkout and alerting management to open another would be lead to more efficient management of people. Such a system would also have financial relevance, such as counting attendance to an event and making sure everyone has paid.

What components are required for a people counting system? One of the most obvious requirements is that it need to be accurate. How accurate is “good enough” dependant on the use. In the case of preventing overcrowding in a stadium the number of people counted does not exact, so long as a good enough estimate is achieved to know when an area is becoming over crowded. In a system that is counting people in order to charge the correct amount you would need an exact and accurate count so that customers are charged the correct price. A customer service scenario would probably demand a less accurate approximation but would still need to be accurate enough to be useful. Another key factor would be the system’s robustness. Groups and crowds of people are a constantly changing and dynamic thing to monitor. Given a typical image of a group of people there is likely to be a large amount of variance in the people in the scene, including their height, body size, clothes etc. Partial or total occlusions are also highly likely especially in very crowded scenes or constantly changing scenes (such as street scene where cars are constantly passing). A successful people counter needs to be able to handle a high degree of variance in a scene as well as have some robustness against partial and total occlusions in order to be accurate enough for practical purposes. Finally, the system will most likely need to be fast. While there are likely to be circumstances where the a people counting system is useful without running in real time, the most interesting practical applications of such a system require that it operates at frame rate. Finding out that a stadium is over crowded is useless thirty minutes after the event.

The practical limitations of a Haar cascade causes some immediate issues when compared with the requirements for a people counter. As mentioned in the preceding section, the basic implementation of a Haar cascade outlined in Ref. [1] uses only very basic rectangular features classify a face from a frontal position. This works well because even though there is a high degree of variance between faces almost all faces share a common structure and it is assumed that the faces to be detected will be viewed from a frontal position. These two assumptions cannot be used when applying the technique to the issues of people counting. Obviously in a real world scene it is unlikely that all of the people in a scene will be facing the directly at the camera. This could be the case for some uses, for example if the application was counting people planned photographs such as wedding photos, but it more likely to not be the case. A modified Haar cascade that has some robustness to rotations such as Ref. [9] could help, but what about situations where there are no clear faces such as people viewed from behind? Or people with hoods up and not looking towards the camera?

This suggests a classifier be trained to detect bodies and not faces in a scene. While faces are not guaranteed to be readably available in a scene, images of human bodies almost certainly will be. However, training a classifier to detect human bodies poses its one assortment of issues. With faces, each face typically has a nose and a set of eyes with the eyes being positioned horizontally above the nose. This gives a common element between all faces that a classifier that uses rigid rectangular features such as a the original implementation of a Haar classifier can be trained to detect. Human bodies on the other hand are much less rigid in there appearance. While human bodies share a degree of similarity (two arms, two legs, one torso etc.) each of these independent parts can the move independently, making the overall structure of a human body far more fluid than the structure of a face. The orientation problem is also exacerbated further when training a body detector over a face detector. A typical street scene will show people from all different angles. They may be viewed from in front, behind, or in profile. They may also be standing, sitting, or laying down. Occlusions will still be a problem for a body classifier. Obviously is a classifier is trained to detector the common features of a body (such as a torso) and these features become obscured (such as another person walking in front) the classifier will fail.

However, despite these issues there is still hope for the application of the technique the problem in question. Viola et al. have themselves carried out experiments adapting the Haar cascade to the practical application of detection pedestrians [3]. In their approach the detecting pedestrians they modify the approach take in Ref. [1] to include an additional set of diagonally shaped features. This is to try and capture some domain knowledge about diagonal shapes within an example. They also realised that motion between two frames of video can encode information whether a section of a scene contains a pedestrian. Using this they created filters which operate on the absolute difference between the two frames shifted in different directions. The Haar like features can still be computed using the integral image representation on the different images. While this approach is shown the the paper to produce decent results approach obviously assumes that you are not counting people in a static image and that people are moving through the scene. If a person stops moving then the classifier will fail to find them as difference between the images will likely be too minimal. Garcia et al. [13] experimented with this approach in combination with background subtraction to try and focus the classifier on certain areas in the scene. Their work found poor performance due to the crowdedness and illumination conditions of their real world dataset. They concluded that a density estimation approach yielded much more positive results.

IV. Conclusion

V. Self-Evaluation

References

[1] P. Viola and M. J. Jones, "Robust real-time face detection," *International journal of computer vision*, vol. 57, no. 2, pp. 137–154, 2004.

[2] M.-H. Yang, D. Kriegman, and N. Ahuja, "Detecting faces in images: A survey," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 24, no. 1, pp. 34–58, 2002.

[3] P. Viola, M. J. Jones, and D. Snow, "Detecting pedestrians using patterns of motion and appearance," in *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*. IEEE, 2003, pp. 734–741.

[4] H. Zhang, W. Jia, X. He, and Q. Wu, "Learning-based license plate detection using global and local features," in *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*, vol. 2. IEEE, 2006, pp. 1102–1105.

[5] P. I. Wilson and J. Fernandez, "Facial feature detection using haar classifiers," *Journal of Computing Sciences in Colleges*, vol. 21, no. 4, pp. 127–133, 2006.

[6]

[7] H. Schneiderman and T. Kanade, "A statistical method for 3d object detection applied to faces and cars," in *Computer Vision and Pattern Recognition, 2000. Proceedings. IEEE Conference on*, vol. 1. IEEE, 2000, pp. 746–751.

[8] H. A. Rowley, S. Baluja, and T. Kanade, "Neural network-based face detection," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 20, no. 1, pp. 23–38, 1998.

[9] M. Jones and P. Viola, "Fast multi-view face detection," *Mitsubishi Electric Research Lab TR-20003-96*, vol. 3, p. 14, 2003.

[10] R. Lienhart and J. Maydt, "An extended set of haar-like features for rapid object detection," in *Image Processing. 2002. Proceedings. 2002 International Conference on*, vol. 1. IEEE, 2002, pp. I–900.

[11] S. C. Brubaker, J. Wu, J. Sun, M. D. Mullin, and J. M. Rehg, "On the design of cascades of boosted ensembles for face detection," *International Journal of Computer Vision*, vol. 77, no. 1-3, pp. 65–86, 2008.

[12] Y.-Y. Lin, T.-L. Liu, and C.-S. Fuh, "Fast object detection with occlusions," in *Computer Vision-ECCV 2004*. Springer, 2004, pp. 402–413.

[13] G. García-Bunster and M. Torres-Torriti, "Effective pedestrian detection and counting at bus stops," in *Robotic Symposium, 2008. LARS'08. IEEE Latin American*. IEEE, 2008, pp. 158–163.