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 applying the Viola-Jones face detection system [1] to counting people in the real world. The Viola-Jones technique presented in ref. [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 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 faces at every pixel and scale an impossibly complex task to 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 face detection is as a component in a larger system, such as a face recognition or image processing/analysis system. 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 speaker [3].

The major contribution that the Viola-Jones technique makes is in the speed at which the algorithm can detect a face. Many approaches prior to the Viola-Jones face detection system could offer reasonably good detection rates [4] [5] [6], but the strength of the method presented in ref. [1] is in the speed-ups offered by their method. The system presented is also not restricted to detecting faces. A cascade classifier such as the one presented in ref. [1] can be trained to detect a variety of other objects provided that sufficient positive and negative examples can be supplied for the training stage. For example, work based on the Viola-Jones detection algorithm has been used to detect people [7], license plates [8], and facial features (eyes, nose, and mouth) [9].

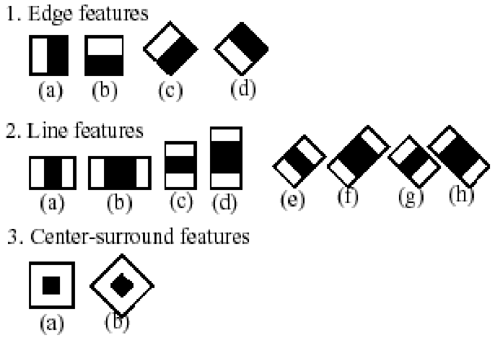

Fig. 1. A collection of Haar-like features used in the OpenCV implementation of a cascade classifier. These include the original features proposed by Viola-Jones [1] as well as additional features suggested by Lienhart et al. [10]. Image taken from the OpenCV site [11].

Viola-Jones solves the problem of rapid object detection through the use of three key components. The first component is the use of Haar-like features to indicate the presence or absence of a face. These features are computed using an integral image data structure. The integral image representation allows for rapid evaluation of Haar-like features in constant time. The integral image is a image where the value of each "pixel" is the summation of all of the pixels above and to the left of it. Using this representation the total area of a subsection in an image may be computed in only four operations. This is perfect for computing blocky Haar-like features which are just the difference between areas of high and low intensity in an image.

The second component is to use the AdaBoost algorithm [12] to create classifiers with high detection rates and low rejection rates [13]. Each of these individual classifiers do not need to be accurate. The only requirement is that they are at least slightly better than random guessing. For this reason such classifiers are referred to in literature as weak learners. In the implementation presented in ref. [1] the AdaBoost algorithm is used both to select the best features and train the classifier. Each weak learner corresponds to a single feature perceptron. Feature selection is achieved through training a bunch of single feature classifiers on the data then selecting the best performing one and combining it with the previously selected classifiers. The data is then re-weighted using the errors in classification to give incorrectly classified examples a higher weighting and then retrained until the desired number of features are obtained.

The third component is to combine the weak classifiers into a cascade of classifiers, with the least complex classifiers at the beginning of the cascade and more complex ones at the end. This ensures that the majority of negative sub-windows are rejected early using classifiers that are quick to evaluate, but that later, more complex nodes in the cascade attempt 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. AdaBoost is also used to train the classifier cascade. It achieves this by adding features to each level of the cascade until the desired rates of detection and false positives are met. Having more features at a later stage in the cascade means that the classifier will achieve higher detection rates and lower false positive weights but at the expense of extra computational time. In this way, classifiers at the start of the cascade will detect a high number false positives but will execute quickly and still remove a large number of irrelevant sub-windows. The following more complex set of features will reject many of the false positives from the previous layer and so on until a sub-window is either rejected or correctly detected as a face.

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 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 the likely 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 face detection represents a major breakthrough in the computer vision community through practical speed-ups, 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.

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 much more efficiently. 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 previous approaches not providing a full receiver operating characteristic (ROC) curves for their classifiers or did not publish their best results. In the case of Sung and Poggio [4] 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 also noted 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 from the side won't be correctly detected by a detector trained on frontal face views. Likewise, training a face on both front and profile images will lead to a bad classifier because it cannot learn features important to both at once. The technique could be used to correctly detect faces viewed from the side by training a different classifier with sufficient positive and negative examples. Both sides of the face can then 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 technique's 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 in real time, the features need to be quick to evaluate and therefore necessarily simple. This is also related to the concept that there is "no free lunch" in search and optimisation [14]. The technique is very good under some circumstances but falls down in others. 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 vertical and horizontal features fail to encode enough information about diagonal features, such as might be present in a face rotated 45 degrees out of the image plane. The detector finds side views particularly difficult to deal with because features presented in the diagonal appear less "blocky" which makes it more difficult for the detector to learn them [13]. The simplicity of Haar-like features also makes the technique unsuitable to detecting objects that are not formed from distinct regular features, such as tree branches [13]. Clearly two images of two different trees will show huge variation in branch position, size, colour etc. which Viola-Jones will be unable to generalise from. The features used will also cause issues when the shape of an object to be identified is itself the most distinguishing feature. The example given in ref. [13] is a coffee mug. Colour, size, and decoration and orientation (e.g. location of the handle) of a mug can vary wildly between images, but the overal shape will remain the same.

Jones and Viola [15] 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 the 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. They performed experiments with a broader set of Haar features which included

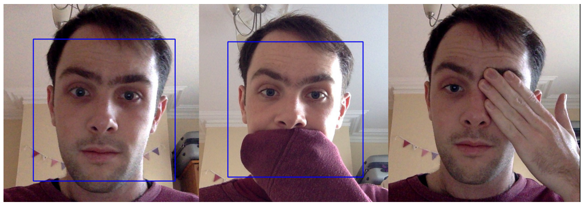

Fig. 2. Example of face detection using the Haar cascade in OpenCV. Occlusions of non essential areas (such as the mouth in the middle image) still result in a positive match, while occlusions of essential features such as the eyes or nose result in a false negative.

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. They also experimented with using small decision trees instead of decision stumps for weak classifiers. They suggested that this allows the weak classifier to learn dependencies between features which cannot be learned in a decision stump classifier. However, this obviously leads to an increase in the execution time as the complexity of the weak learners is increased.

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 important 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 2) and will still return a positive match. However, occlusions of prominent features used to classify a face will result in a false negative (such as occluding an eye in the right hand image of figure 2). 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. [16] 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 when 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. A well formed data should contain examples where the object to be identified shows roughly the same feature set in every image and should only contain one view (such as the frontal view of a face) [13].

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 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 too much accuracy in order to achieve it. In summary 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 large quantities of 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 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. An essential realisation is that counting people in a scene using an approach based on the Viola-Jones technique boils down to a person detection problem. If people in a given scene can be reliably detected independently from one another then they can be counted.

What components are required for a people counting system? One of the most obvious requirements is that it needs to be accurate. How accurate is “good enough” is dependant on the system implementation. In the case of preventing overcrowding in a stadium the number of people counted does not have to be 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 between 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 a 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 overcrowded 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 the 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 counting people in a scene. Obviously in a real world scene it is unlikely that all of the people in a scene will be facing directly at the camera. This could be the case for some uses, for example if the application was counting people in planned photographs such as wedding photos, but this is probably not the case. A modified Haar cascade that has some robustness to rotations such as ref. [15] 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 should be trained to detect bodies and not faces in a scene. While faces are not guaranteed to be reliably available in a scene, images of human bodies almost certainly will be. However, training a classifier to detect human bodies poses its own assortment of issues. With faces, each face typically has a nose in the centre and a set of eyes positioned horizontally above the nose. This gives a common element between all faces that a classifier that uses rigid rectangular features such as 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. 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 detecting pedestrians [7]. In their approach 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 about 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 in the paper to produce decent results approach it assumes that people are moving in the scene. If a person stops moving then the classifier will fail to find them as the difference between the images will likely be too minimal.

Garcia et al. [17] experimented with vanilla Viola-Jones 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. Density estimate approaches have the advantage that they make no assumptions about the features, orientation or motion of people in a scene (which is effectively the inductive bias of Viola-Jones). In ref. [17] the only assumptions were that the background could be successfully extracted and everything in the foreground is a person. The method used in ref. [17] also had some invariance to objects in the foreground that were not human, such as pets or luggage. Another approach is to train a classifier to detect the heads of people. This is exactly the approach taken by Subburaman et al. [18] with their head counting approach to counting people in crowded scenes. This has particular relevance if the target application is in crowded areas where it is likely that the majority of a persons body is occluded by other people in the scene. To help aid the classifier they use a gradient based approach to find regions of interest before applying the classifier. Yet another approach to the problem of human detection has been presented by Mikolajczyk et al. [19] which attempts to produce a human detector using a combination of 7 different invariant body parts such as different orientations of the face, head, upper and lower body and combining them using a probabilistic model for the joint co-occurrence of features. This obviously takes into consideration issues with variance in the positioning of human body parts by modelling each part independently and then using proximity information boost the accuracy of a prediction.

An additional source of information that could be used to extend the accuracy of a Haar cascade based people counter is the use of multiple cameras. Viewing a scene from multiple angle adds additional complexity to the system because the views from each of the cameras need to be correlated in order to avoid counting some people twice. The advantage of multiple cameras is that this allows you to utilise multiple sources of information. Using this as part of a voting system can help with resolving an ambiguous detection from a single detector and prevent problems with occlusions. Multiple cameras mean that people who might be difficult to detect from one view are easier to spot from another angle. Zhang et al. [20] proposed a multi-view face detection system that used an AdaBoost variant FloatBoost [21].

Putting all of these points together it becomes clear that the Viola-Jones method is not enough on its own to produce a system that will be accurate enough for practical purposes in the real world. Weighing the different sources listed in this section suggests a number of improvements that could be undertaken to try and make Viola-Jones applicable to real world people counting. The first extension required for a practical people counting system using Viola-Jones would be to extend the feature set used to include features such as those suggested by Lienhart et al. [10], or if the system is going to be used in a scenario where people are likely to be moving around a lot, by using motion features like those suggested by Viola and Jones in ref. [7]. This would make the classifier more robust to real world circumstances where people are unlikely to be directly looking at the camera and will have different poses and rotations in and out of the image plane. Multiple classifiers could also be used. For example, one could be trained to detect a face in the frontal position, while another could detect the side of a face (and the image flipped to detect the other side).

Whether the classifier used would be trained to detect a face or a the whole body largely depends on the target application of the system. If the system is going to be used in an application with a relatively sparse count of people in the scene then training a body or invariant body parts classifier seems to be the correct method to use, particularly if the subjects are likely to be moving across a range of frames. However, in crowded scenes a whole body detector isn't likely to work as only the very top of a person such as their head is likely to be visible to the camera. In this case it would be better to train a classifier to detect just the head. However, as shown by Garcia et al. [17] this generally performs worse than other methods. Multiple cameras could be used to gain a better perspective on a scene and give some additional information that would catch missed people due to occlusions. The part of a human that the classifier is trained to detect is also reliant on positioning of the camera within the system. For example, a frontal face detector is not going to work very well if the camera is necessarily positioned side onwards to the flow of people through an area. However, the choice of the camera position can be advantageous to the implementation of a people counting system. For example, if the proposed system was supposed to count the number of people in a line at a checkout, it may be better to position the camera side on to the checkout so that the queue of people appear side on to the camera. That way people are not occluding each other as they stand in a line at the checkout.

Depending on the power of the architecture of the equipment to be used in the counting system, the research in this paper suggests that it would be advantageous to utilise multiple classifiers for different likely scenarios in the scene. For example, Lin et al. [16] showed that they could using multiple classifiers could improve robustness by training separate classifiers that deal with occlusions or the work by Jones and Viola [15] which uses different detectors for each pose. Obviously using multiple classifiers adds to the computational complexity of the system and depending on the hardware requirements of a people counting system it may be infeasible if the the equipment has low processing power and the system is required to run at frame rate. Of course this might not be such an issue if the system is not required to run at frame rate or the machine doing the analysis is powerful enough to handle it.

Another hardware consideration would be the use of a thermographic camera to help narrow the search space of by thresholding everything in the original image that does not correspond to a strong enough heat signature in the thermal image, thus focussing the detector. This would provide an advantage over other background subtraction techniques in that it would remove anything that isn't hot (such as furniture). Depending on the environment this could drastically reduce the search space. Obviously things other than people are hot in the real world which would cause this approach to be less useful. Also in a scene which mostly consists of crowds the thermal camera may be overloaded with too much information. Yet another issue would be the additional complexity of computing a homography to map the capture from the thermal camera to the regular camera. Despite the author's best efforts no relevant sources could be found to support or refute this idea.

IV. Conclusion

In conclusion, the first half of this paper has presented a review of the highly susccessfully Viola-Jones approach to face detection and discussed subsequent work building upon the basic implementation of the framework to include new features and enhanced set-ups for different problem domains. The Haar cascade has a wide number of applications a variety of different to object detection problems and is in no way limited to the just face detection. The key insight that the Viola-Jones object detection framework brings to the table are speed-ups allowing technology from the early 21^{st} century to perform object detection at frame rate. The first key contribution from the technique was the use of the integral image data structure to reduce the time complexity of evaluating features to a constant time operation. The second contribution was the use of the AdaBoost algorithm to perform feature selection of Haar like features which can be evaluated using the integral image. Finally they presented a way to combine classifiers into a degenerate decision tree with successively more complex nodes called a classifier cascade. These three contributions produced a framework that was extremely efficient yet very accurate at detecting rigid body objects with "blocky", largely invariant features.

The second half has focused on the application of the Viola-Jones framework and the subsequent work and variants to the titular question of counting people. We have suggested that the Viola-Jones framework presented in its most basic form would not be robust enough to be used practically as part of a useful people counting system unless such a system was used in a highly unrealistic environment. We have also presented a selection of existing work carried out in the field since the Viola-Jones break through with various extensions on the original framework. Combining the major findings of the approaches we have suggested what modifications work probably be desirable to a system wishing to use Viola-Jones for counting people, while keeping a open mind to the broadness of the titular question. We have concluded in that for practical use in people counting a Viola-Jones based system would need to boost its robustness against the high degree of variance in the pose of bodies and face among real life scenes. This would most likely involve adding new features that capture a greater amount of domain knowledge such as diagonal shapes [10], or features present in the motion of pedestrians [7]. Such a system would also need to consider the high degree of occlusions likely to be present in a real world scenario and take advantage of either a multi-detector or multi-camera approach. Either approach would be able to add support to deal with occlusions of essential features that cannot be handled by the vanilla Viola-Jones technique alone. However, due to the instability of the framework in real life systems due to the high amount of variance in a typical scene and the framework's lack of robustness thanks to it's necessarily simple features an alternate approach based on the methods such as density estimation which do not make assumptions about position or pose may yield better results in a practical setting and would be an interesting area for further experimentation.

V. Self-Evaluation

I believe that this piece of work is worth an A- as defined by table 2 in section 5.7 of the Aberystwyth University student handbook. I believe that this would be a balanced choice based on the work presented in this document and on what I would do differently with hindsight. I think I deserve this grade because I have demonstrated that I have read and understood the paper upon which this assignment is based. I have also clearly analysed the major success and failure methods of the technique under examination and outlined why the technique was such a breakthrough in the computer vision field. In the second section I have researched a variety of existing work in the area of detecting people using the Viola-Jones technique and drawn conclusions based on these sources. I feel that I have achieved this fairly successfully and that I have shown I can read a comment on scientific research. I have also shown that I can propose reasonable suggestions on how to implement the system proposed in the scenario. However, there is always room for improvement. I found the amount of background reading and how to write about it effectively one of the biggest challenges and I feel that there are some areas which I could have expanded on further. In hindsight I would have written more in comparison with alternative approaches used to count people using computer vision and focused less on extensions of Viola-Jones. I could of also possibly expanded the application section with some more outlandish suggestions than I listed here. I also feel that I could have played with the detector more in OpenCV and perhaps have trained my own classifier on some pedestrian data and included some actual performance results for the vanilla Viola-Jones implementation. Finally, I feel I could of written a better description of how the technique works. I found it difficult to explain concisely as the system is fairly complex and accurately describing how each component hangs together was challenging.

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] E. Hjelmås and B. K. Low, "Face detection: A survey," *Computer Vision and Image Understanding*, vol. 83, pp. 236–274, 2001.

[4] K.-K. Sung and T. Poggio, "Example-based learning for view-based human face detection," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 20, no. 1, pp. 39–51, 1998.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[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] OpenCV. (2014) Cascade classification opencv api reference. [Online]. Available: http://docs.opencv.org/modules/objdetect/doc/cascade_classification.html

[12] Y. Freund, R. E. Schapire *et al.*, "Experiments with a new boosting algorithm," in *ICML*, vol. 96, 1996, pp. 148–156.

[13] G. Bradski and A. Kaehler, *Learning OpenCV: Computer vision with the OpenCV library*. " O'Reilly Media, Inc.", 2008.

[14] D. H. Wolpert and W. G. Macready, "No free lunch theorems for optimization," *Evolutionary Computation, IEEE Transactions on*, vol. 1, no. 1, pp. 67–82, 1997.

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

[16] 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.

[17] 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.

[18] V. B. Subburaman, A. Descamps, and C. Carincotte, "Counting people in the crowd using a generic head detector," in *Advanced Video and Signal-Based Surveillance (AVSS), 2012 IEEE Ninth International Conference on*. IEEE, 2012, pp. 470–475.

[19] K. Mikolajczyk, C. Schmid, and A. Zisserman, "Human detection based on a probabilistic assembly of robust part detectors," in *Computer Vision-ECCV 2004*. Springer, 2004, pp. 69–82.

[20] Z. Zhang, G. Potamianos, M. Liu, and T. Huang, "Robust multi-view multi-camera face detection inside smart rooms using spatio-temporal dynamic programming," in *Automatic Face and Gesture Recognition, 2006. FGR 2006. 7th International Conference on*. IEEE, 2006, pp. 407–412.

[21] S. Z. Li and Z. Zhang, "Floatboost learning and statistical face detection," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 26, no. 9, pp. 1112–1123, 2004.