ASM Sc. J., 13, 2020
https://doi.org/10.32802/asmscj.2020.sm26(5.13)

Evolution of Face Detection Techniques in Digital Images

Abdullah Bade* and Tulasii Sivaraja

Mathematics, Graphics and Visualization Research Group (MGRAVS), Faculty of Science and Natural Resources, Universiti Malaysia Sabah, 88400 Kota Kinabalu, Sabah, Malaysia

Face detection is a popular application of object detection in computer vision. To detect faces in a digital image or video input requires the computer to utilize a series of algorithms or techniques. The technology of detecting faces has evolved proportional to its usage in various applications such as biometric security, autofocus in cameras, robotics, and social media applications. This aim of this paper is to provide a description on the evolution of face detection techniques such as Viola Jones, Histogram of Oriented Gradients (HOG), Convolutional Neural Network (CNN), and Multitask Cascaded Convolutional Neural Networks (MTCNN) from the year 2000 to recent years. Additionally, a comparison of the face detection techniques is discussed to evaluate the optimal face detection technique.

Keywords: face detection, cascade training, cascade

I. INTRODUCTION

Computer vision is an extended disciplinary field of computer graphics. The term computer vision describes the process, or the action done by the computer that mimics the human visual system. Another interpretation of computer vision is creating a self-sufficient system that can carry out some of tasks of the human visual cortex (Huang, 1996). Most scientists use computer vision to extract information from digital images or videos the same way a human being's visual receptors do. The processes of computer vision can be condensed into three steps, to see or detect, to describe or recognize and to understand. An average person is able to do consecutive steps in split seconds; the individual is able to detect and identify an object and react to it almost immediately. To replicate this using a computer is no easy task. The computer has to be able to see and detect objects from images or videos, which can be done by providing the computer with an image or video via a camera. Then, the computer has to interpret the input and be able to describe and identify the input, in other words recognize an object. Since computers do not perceive information as humans do, the image has to be converted to a form that the computer is able to understand which is in numbers that represent intensity.

Face detection is one of the more popular application of object detection in computer vision. The computer uses a series of mathematical algorithms, pattern recognition and image processing to identify faces from an image or video input. Over the years, the technology of detecting faces has evolved proportional to its usage in various applications. One of the earliest documented usage of face detection was traced back to the work of Woodrow Wilson Bledsoe in the 1960s, where his research using manual measurements to recognize faces paved the way to the various systems developed (Bledsoe, 1964).

The issue of finding faces in images has been one of the most significant undertaking in computer vision. Many social networking companies such as Facebook and Snapchat are employing face detection and face recognition technology to further enhance user experience

*Corresponding author's e-mail: abb@ums.edu.my

(Rajawat, Pandey, & Rajput, 2017). However, there is a sizeable amount of hindrance that is encountered with face detection. The challenges encountered in the field of face detection and recognition is commonly expressed as A-PIE, which represents aging, poses, illumination, and expression (Mahalingam, Ricanek, & Albert, 2014).

Humans are able identify human faces regardless of the age of the person identified. However, computers are not able to do so. The face of an infant and an elderly person significantly differs from the larger population of teenagers and adults as they lack some crucial facial features that is searched for during face detection. Meanwhile, variation in face detection such as smiling or laughing make face detection and recognition much more tedious and difficult to achieve as the structure of the differs with different expressions (Atta & Ghanbari, 2010). The varying facial expressions challenges face detection algorithms which are rigid or trained to detect faces with a neutral expression (Kutty & Lakshmy, 2017). Lighting and angles are constant issues that researchers are trying to address (Meena & Suruliandi, 2011). The brightness of the image influences the visibility of facial features. Most face detection algorithms have a pose angle allowance of [-15°, 15°] yaw. Some of these pose variations creates a situation whereby the faces can no longer be considered frontal faces and some face detection algorithms will fail to detect these faces.

Similarly, partial occlusion due to glasses, hats, scarfs, hair, or other objects also pose a problem to the task of face detection as not all the face features are available for detection. Researchers began noticing that partial occlusion of faces affects the overall detection rate of the algorithm. Most of the face detection techniques work by searching facial features such as mouth, nose and a set of eyes in images in a joint search. Therefore, occlusions such as sunglasses and scarves hinder the detection of faces which consequently affects the performance of the detection algorithms.

The general concept of face detection is most commonly achieved with the following three steps; the first step is to examine the picture or video frame to determine the regions of interest. This is usually done via a sliding window. The second step is to acquire the extracted features or patterns from the region of interest. This is where the main aspects of the face detection algorithm lie such as using Haar-like features, Histogram of Oriented Gradients or deep learning methods such as convolutional neural networks. Finally, the third step is to classify if the detected regions of interest into faces and non-faces for recognition.

For this paper, a series of techniques introduced after the year 2000 are considered. However, there are a few face detections approaches that have been successfully implemented prior to the year 2000, for example Eigenfaces (Turk &Pentland, 1991) and genetic algorithm (Wong & Lam, 1999). For the purpose of this paper, two classical methods, Haar-cascade classifiers and Histogram of Oriented Gradients (HOG) and two variations of deep learning methods are discussed. The recent works in face detection and recognition have gravitated to deep learning approaches that uses convolutional neural networks.

The remainder of this paper is divided into 4 other sections; Section 2 and Section 3 are the classical and current methods of face detection respectively, Section 4 is a comparison of the methods covered in the previous sections and the final section concludes the paper.

II. CLASSICAL METHODS

Viola-Jones Haar cascade classifier and Histogram of Oriented Gradients are some of the face detection approaches introduced in the early 2000s. These are both categorized as feature-based approaches in face detection. Feature-based approaches in face detection requires the extraction of facial features from an image and comparing it with a knowledge base of face features (Modi & Macwan, 2014).

In 2001 the pair of researchers, Paul Viola and Michael Jones, introduced a method of face detection that is still being used till this day. They proposed a framework that produces real time face detection by the means of a novel image representation known as integral image and creating a boosted cascade of weak Haar-like feature classifiers (Viola & Jones, 2001). Specific Haar-like features are used to evaluate the abrupt changes in terms of colour intensity for facial features. Figure 1 shows the

Haar features used in the Viola-Jones face detection algorithm. The rotated features were added in 2002 to expand and Haar features used to evaluate faces and improve the detection rates (Lienhart & Maydt, 2002).

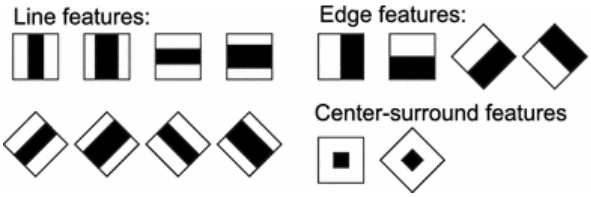

Figure1. Haar-like features (Basic and extended)

There are for stages in the implementation of Viola-Jones Haar cascade classifier face detection, which is illustrated in figure 2.

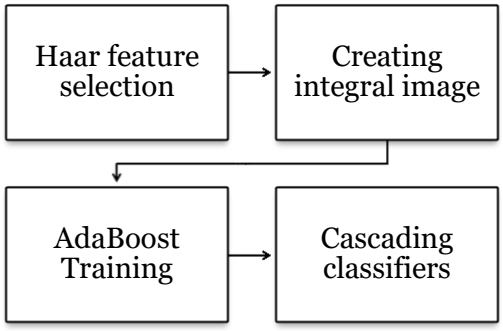

Figure 2. Stages of Viola-Jones method

Haar feature selection covers the process of generating an infinitely large Haar-like feature pool and selecting only the features that obey several conditions to reduce the features selected (See figure 1 for the possible features selected).

Then the integral image is calculated in order to compute the features. This is done by calculating the summed area table $SAT\ (x, y)$ and the rotated summed area table $RSAT(x, y)$ for the 45° rotated Haar-like features. $SAT(x, y)$is defined as the sum of pixel of a rectangle from the top left corner (0,0) to the bottom right corner (x, y). Eq. (1) represents $SAT(x, y)$:

$$SAT(x, y) = \sum_{x' \le x, y' \le y} I(x', y') \quad (1)$$

where I is the intensity value of (x, y). Similarly, $RSAT(x, y)$ is defined as the pixel sum of a 45° rotated rectangle with the bottom corner as (x, y). Eq. (2) represents $RSAT(x, y)$:

$$RSAT(x, y) = \sum_{|x-x|' \le y-y', y' \le y} I(x', y') \quad (2)$$

Once the integral image is computed, the Haar-like features which are weak classifiers are boosted with AdaBoost. The idea is that the usage of a group of boosted classifiers instead of weak classifiers will be able to improve the detection rates while reducing the computation time.

The usage of a cascade structure enables non-object regions to be discarded at the early stages and only focus on the relevant regions of interest. The cascade classifiers have *N* stages that are in a connected pattern of classifiers which are able to differentiate the detected face, in this case, and background. The features evaluated at the later stages are focused on determining if the region of interest is a face or non-face.

The concept of HOG was first introduced in 2005 by Navneet Dalal and Bill Triggs, where they used it to detect humans (Dalal & Triggs, 2005). However, this method has been extended to be implemented for face detection.

The general approach of HOG can be summarized as using a feature descriptor to represent a region in an image which is then used to simplify the representation of the image by extracting important information. The descriptor which is the histogram of gradient directions computed from the difference in surrounding pixels, links together the computed values of the HOG directions for all the cells. Normalizing the values of the cell beforehand will help with computation.

The information of the direction and magnitude of the HOG provides information on the same and edges of an object or in this case the face. This is largely due to the face that there are abrupt changes in terms of color intensities around edges and corner points.

The initial step of the HOG descriptor involves calculating the horizontal and vertical gradients which is achieved using the kernels in Figure 3 to filter the image.

Figure3. Kernel to calculate horizontal and vertical gradients.

Once the kernel value of all the pixels is computed, both values will be used to calculate the magnitude and direction. These values will then be used to construct the HOG, whereby the gradient magnitudes are distributed evenly among the range of gradient directions.

Similar to the Haar classifier, the positive and negative HOG descriptors will be extracted from the positive and negative images supplied in the training. A HOG will be generated at each cell of an appropriate size. The HOG can be normalized by using a 50% overlapping sliding window, whereby a normalized HOG is generated by combining all the cell HOGs in the sliding window. Finally, the extracted HOG feature descriptors will be classified using machine learning methods such as Support Vector Machine (SVM). Figure 4 summarizes the implementation of HOG method.

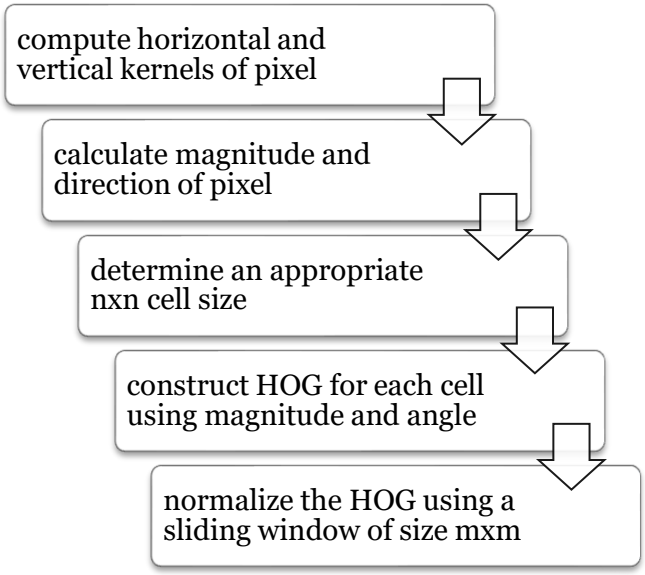

Figure4. Summary of HOG method

III. DEEP LEARNING METHODS

In recent years, most of the face detection approaches are achieved by using deep learning algorithms and neural networks. These approaches have produces successful results in terms of performance and computation time. It is important to note that deep learning approaches are extremely effective in detecting partially occluded faces and faces with various poses.

In order to make sure that the information gained from face detection is useful, a group of researchers proposed a novel method of using deep learning to detect faces and its attributes for images in the wild. Their method involve using two CNNs in cascade, LNet and ANet, whereby LNet detects the faces by means of general object detection for face localization and ANet is used to predict facial attributes (Liu, Luo, Wang, & Tang, 2015).

In 2016, the trend of using a cascaded CNN continues as a group of researchers use a multi-task cascade CNN (MTCNN) framework to boost the performance of face detection by exploiting the inherent correlation between detection and alignment. Their approach makes use of a cascaded architecture that has three stages of mindfully designed deep CNNs that predict faces and landmark locations in a coarse-to-fine manner (Zhang, Zhang, Li, &Qiao, 2016). Other than that, another group of researchers proposed an end-to-end multi task discriminative learning framework that integrates a CNN with a 3D mean face model, where it uses the estimation of facial key-points and the 3D transformation parameters such as rotation and translation in order to compute the bounding box proposals as well as using the configuration pooling of facial key-points to prune and refine proposals in order to compute the detection results (Li, Sun, Wu, & Wang, 2016). This approach addresses the issue of CNNs in a heuristic design of predefined bounding boxes for the region proposals, and also the region of interest of the pooling layer.

Similarly, Chen, Hua, Wen and Sun proposed method described as Supervised Transformer Network which is a cascaded CNN whereby the first stage of the cascade uses a multi task Region Proposal Network (RPN) to predict candidate faces and its facial landmarks, followed by a RCNN that validates the candidate face regions in the second stage (Chen et al., 2016).

In 2017, Ranjan Patel and Chellappa introduced HyperFace which simultaneously detects faces, facial landmarks, head pose estimation and gender recognition using CNNs. They execute the process in three modules, whereby in the first module the images are used to generate and scale independent region proposals, the second module uses CNN to classify the region proposals as face or non-face, which will also provide information on face landmarks location, estimated head pose and gender

information if it is classified as face and in the final module, post-processing steps are taken to improve the performance of each task (Ranjan et al., 2017). Figure 5 shows the output of HyperFace, whereby the gender recognition is conveyed through the color of the bounding box (pink for females and blue for males), head pose estimation is given by the numbers on the top of respective bounding boxes and the facial landmarks are highlighted using green dots.

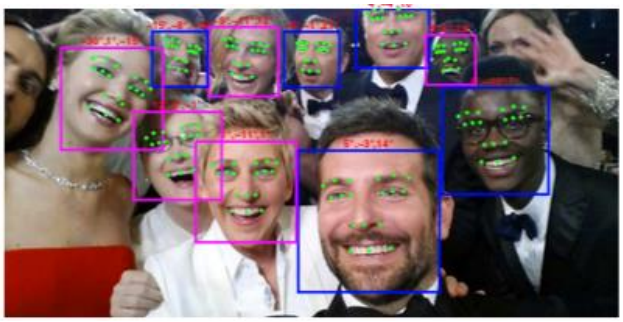

Figure 5. HyperFace output results

In 2018, a group of researchers suggested a deep learning approach of face detection from facial parts responses. They proposed a deep convolutional network that finds faces from a new perspective by using the scores of facial parts responses from their spatial structure and arrangement, with consideration the challenges that partially occluded faces would pose (Yang, Lau, Loy & Tang, 2018).

IV. COMPARING CLASSICAL METHODS WITH DEEP LEARNING METHODS

The comparison between classical and deep learning methods is summarized in Table 1.

The classical methods and deep learning approaches have their own advantages in terms of implementation, accuracy and computation. However, deep learning approaches tend to produce better results in terms of accuracy as it adapts to new information and is continuously learning and adding the new information to its base of knowledge. On that note, it is also important to note that classical methods are the base of deep learning approaches. Deep learning approaches integrate classical methods with the emerging technology to become more efficient and effective in the field of face detection.

Table 1. Comparison of classical and deep learning methods

	Classical Methods	Deep learning
Training Data	Training images has to be uniform to ensure the model is able to establish a standard pattern	Training images can cover a wide range of variation to allow the CNN to learn and adapt the patterns
Detection	Frontal faces with limited pose and lighting variations	Faces with varying levels of occlusion, lighting, pose, emotions
Multitask capabilities	Requires additional machine learning algorithms to achieve classification	Able to extract and classify information from detected faces
Accuracy (Correctly detects faces)	Accuracy depends on the training images used. A large collection of only the face view/type needs to be used	Highly accurate with large amount of any case of face training images used. Various face conditions do not affect accuracy

V. CONCLUSION

The techniques used in achieving the task of face detection as evolved from the classical approaches to deep learning approaches mostly due to the extensive research and advancements achieved in the area of big data problems, which has improved the implementation of training algorithms. Most of the major challenges surrounding face detection and recognition are being solved using one of the variations of deep learning approaches. The next step is to optimize the performance of these algorithms so that the trade-off between computation time and detection rates are marginalized.

VI. REFERENCES

Atta, R., & Ghanbari, M. (2010, October). Face recognition based on DCT pyramid feature extraction. In Image and Signal Processing (CISP), 2010 3rd International Congress on (Vol. 2, pp. 934-938). IEEE.

Bledsoe, W. W. (1964). Facial Recognition Project Report. Technical report PRI 10, Panoramic Research, Inc.

Chen, D., Hua, G., Wen, F., & Sun, J. (2016, October). Supervised transformer network for efficient face detection. In European Conference on Computer Vision (pp. 122-138). Springer, Cham.

Dalal, N., & Triggs, B. (2005, June). Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on* (Vol. 1, pp. 886-893). IEEE.

Huang, T. (1996). Computer vision: Evolution and promise.

Kutty, L. T., & Lakshmy, S. (2017, July). Fast and efficient compact feature descriptor for face recognition. In Networks & Advances in Computational Technologies (NetACT), 2017 International Conference on (pp. 305-309). IEEE.

Li, Y., Sun, B., Wu, T., & Wang, Y. (2016, October). Face detection with end-to-end integration of a convnet and a 3d model. In European Conference on Computer Vision (pp. 420-436). Springer, Cham.

Lienhart, R., & Maydt, J. (2002). An extended set of haar-like features for rapid object detection. In Image Processing. 2002. Proceedings. 2002 International Conference on (Vol. 1, pp. I-I). IEEE.

Liu, Z., Luo, P., Wang, X., & Tang, X. (2015). Deep learning face attributes in the wild. In Proceedings of the IEEE International Conference on Computer Vision (pp. 37303738).

Mahalingam, G., Ricanek, K., & Albert, A. M. (2014). Investigating the periocular-based face recognition across gender transformation. IEEE Transactions on Information Forensics and Security, 9(12), 2180-2192.

Meena, K., & Suruliandi, A. (2011, June). Local binary patterns and its variants for face recognition. In Recent Trends in Information Technology (ICRTIT), 2011 International Conference on (pp. 782-786). IEEE.

Modi, M., & Macwan, F. (2014). Face detection approaches: A survey. International journal of innovative research in science, engineering and technology, 3(4).

Rajawat, A., Pandey, M. K., & Rajput, S. S. (2017, February). Low resolution face recognition techniques: A survey. In Computational Intelligence & Communication Technology (CICT), 2017 3rd International Conference on (pp. 1-4). IEEE.

Ranjan, R., Patel, V. M., & Chellappa, R. (2017). Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence.

Turk, M. A., & Pentland, A. P. (1991). Face recognition using eigenfaces. In Computer Vision and Pattern Recognition, 1991. Proceedings CVPR'91., IEEE Computer Society Conference on (pp. 586-591). IEEE.

Viola, P., & Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. In Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on (Vol. 1, pp. I-I). IEEE.

Wong, K. W., & Lam, K. M. (1999). A reliable approach for human face detection using genetic algorithm. In Circuits and Systems, 1999. ISCAS'99. Proceedings of the 1999 IEEE International Symposium on (Vol. 4, pp. 499-502). IEEE.

Yang, S., Luo, P., Loy, C. C., & Tang, X. (2018). Faceness-Net: Face detection through deep facial part responses. IEEE transactions on pattern analysis and machine intelligence, 40(8), 1845-1859

Zhang, K., Zhang, Z., Li, Z., & Qiao, Y. (2016). Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Processing Letters, 23(10), 1499- 1503.