ISSN 2278-3091
Volume 9, No.4, July – August 2020
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse364942020.pdf
https://doi.org/10.30534/ijatcse/2020/364942020

Room Surveillance using Convolutional Neural Networks - Based Computer Vision System

Roy Francis Navea
De La Salle University, Philippines, roy.navea@dlsu.edu.ph

ABSTRACT

Intelligent systems are capable of performing several tasks with high reliability and efficiency. Hence, these systems were used to perform tasks which are usually done by humans. In the event of facility breach or in times when primary security systems were compromised, a call for secondary line of security is needed. In this study, it is intended to design a convolutional neural network - based computer vision system that can possibly determine whether a person entering a vicinity is authorized or not using face, height, and built recognition with gender sensitivity. The designed system was able to obtain balanced precision and recall as well as achieving more than 0.9 F1 scores. This is a complementary technology that can work with automated locks or security systems.

Key words : CCTV , Convolutional Neural Networks, Face Detection, Gender Skewness, Security Systems

1. INTRODUCTION

Technological developments have contributed to the development of vision-based systems that interact with humans especially in terms of safety and security. More and more surveillance systems are being deployed in both private and public areas which are manually monitored in local command centers. Incorporating intelligence in these passive surveillance systems is a trend which targets wide range of applications [1]. One of these applications is in surveilling areas in which only limited or authorized personnel are allowed. Door lock security systems are currently available however, in the event that these primary security measures are compromised, a secondary security measure is necessary.

Closed Circuit Television (CCTV) systems are widely used for security purposes. For this purpose, the main objective of CCTV surveillance implementation is deterrence. An early version of this kind of system acts like a virtual guard used for detecting potential criminal activity in certain public areas. It provides an alarm for whenever an observed activity matches pre-defined suspicious behaviors criteria programmed into the system [2]. The presence of the surveillance cameras is thought to have a deterrent effect on potential offenders for as long as they are aware that they are being watched. Hence, crime prevention is a feature of the perception of the offender which might produce self-discipline in which individuals could control their acts [3]. CCTV cameras capture instances within a given period of time. Analyzing the video feeds calls for algorithms which can possibly detect and recognize the objects shown in the feed.

Different algorithms are already available and are implemented in computer vision systems designed for a specific application [1][4][5]. For instance, surveillance videos are analyzed by incorporating learning algorithms like neural networks to detect the presence of fire in a certain vicinity [6]. In addition, image enhancement techniques are also used to further provide good images for recognition [7][8][9]. Object detection is one of the techniques that connects image processing and digital vision systems. It considers physical objects and their characteristics with an in-depth analysis of their distinct features as extracted from images and video frames [10]. Electronic devices use face detection algorithms to detect faces for purpose of entertainment and security. This is a biometric data used for personal digital identification, authentication, authorization, pin locks, and account verification. Establishments have been using face recognition in their security cameras, access controls and local data centers which are also useful when it comes to law enforcement and security [5].

One of the emerging machine learning algorithms is the convolutional neural network (CNN). It is a multi-layer perceptron characterized by a deep feed-forward artificial neural network. CNN has been one of the most innovative learning models in the field of computer vision and was proven to perform a lot better than the traditional ones especially in the field of image classification, object detection, segmentation and face recognition. Multimodal information

can be used to learn face representations with deep learning frameworks [11]. CNNs can extract complementary facial features then concatenate them to form a high-dimensional feature vector that can be used of detection, recognition and classification. Cascading CNNs increases discriminative capabilities while maintaining high performance. Cascaded CNNs can operate at multiple resolutions that can evaluate several significant candidates in the terminal high resolution and remove background sections in the fast low resolution stages [12]. In [13], cascaded CNNs were used for face localization and attribute prediction. This was introduced since early face detection algorithms failed to locate faces when unconstrained face images with complex variations are present. In effect, with wrong localization, features obtained are not really from faces but instead, from other parts of the image. CNN imposes less computational cost and can possibly run without GPU or just even in mobile phones [14]. In [6], the CNN framework used consumes lesser time in computation as it automatically learns features from the given raw data. However, in some applications, higher computing machines are necessary [15].

With the availability of advanced computer vision system algorithms in terms of face detection, recognition and classification, these can be implemented to further improve area security which are commonly performed by surveillance or monitoring systems. Adding intelligence to passive surveillance systems will improve its capabilities to secure a specific area of consideration. This study generally aims to develop a CNN-based computer vision room surveillance system. Specifically, the system will use an IP-Camera for video surveillance which will then be equipped with intelligence in terms of detection, recognition and classification using convolutional neural networks. An algorithm will be developed to enable the system to detect faces, recognize the detected faces and classify them whether authorized or not. The system should be able to obtain the height and built of persons who were classified to be unauthorized. A data base system should be made available to record information about the persons captured by the camera.

Having an intelligence-equipped CCTV system, building a secondary computer vision –based surveillance system is possible. In this study, an authorized personnel or a non-authorized personnel in a strictly prohibited area is attainable. This study focus on developing a CCTV system using IP-cameras that automatically determine if a personnel is authorized or not inside a room. The system allows passage of authorized personnel but for non-authorized, it captures an image of the person, determine the person's height and built and store information in a database.

This study finds its significance in the field of vision-based security systems. Intelligence finds its meaning in applications and implementations. For computer vision, machine learning and artificial intelligence (AI) in general, systems in which they can be fully utilize for the purpose of, for example, safety and security is important in both narrow and broad perspectives [16][24]. CCTV systems are almost everywhere but intelligent ones are rarely visible locally. Intelligent CCTV systems work in small scale enterprises but if expanded into a larger coverage in terms of people, area and infrastructures, this system not only applies to high security rooms of enterprises but also in almost every area where safety and security is needed.

2. THEORETICAL CONSIDERATIONS

This section basically talks about the CCTV systems and the convolutional neural networks. Theories behind these two concepts are combined to build an intelligent computer vision system.

2.1 Closed Circuit Television (CCTV) Systems and Internet Protocol (IP) Cameras

A CCTV system is generally composed of four major components: the camera, lens, monitor and a video recorder. The most important one is the camera because it is the one that collects the images. In image processing, the quality of the images obtained matters. For versatility of views, some CCTV cameras are equipped with motor that help it to move the zoom parts which depends on the lenses used. Once an image is captured, it is taken to the monitor and then recorded on a digital video recorder [17]. Typical cameras are made-up of lenses that focus light to create images. For cameras using charge coupled devices (CCD), images are recorded by altering the electrical charge whether high or low. Brighter images are represented by high electrical charges while darker images are represented by low electrical charges. In this effect, black and white images are produced. Colored images are produced by detecting not only the total light levels, but also the levels of each wavelengths of light. Resolution outcomes for identification and recognition shall be the main consideration in camera selection [18].

CCTV is mostly used in proprietary mode as in business establishments or personal mode as in homes or personal spaces. Most CCTV systems are used for monitoring and surveillance to guard, keep-safe, and secure.

Internet protocol (IP) cameras have brought a new dimension in modern security systems in different establishments. IP cameras record digital signals and use transmission and

security features of the TCP/IP protocol. Since digital recording is present, it results in a much higher capacity for higher resolution videos than traditional CCTV cameras [19]. IP cameras can be wired or wireless according to their make and designs.

2.2 Convolutional Neural Networks

Convolutional neural networks are multi-stage architectures that can be trained to learn invariant features. It is composed of filter banks, some non-linear and feature pooling layers [16]. In general, there are three layers that make up the convolutional neural network, namely convolutional layer, pooling layer and the fully - connected layer. A typical architecture of CNN is shown in Figure 1. The convolutional layer is the core part of the network because it contains the local connections and weights of the shared characteristics obtained say, from an image. The convolutional layer is where features of the input/s are learned. There are feature maps composed of neurons which extract the local spatial characteristics in the former layer [14]. For learning a new feature, the convolution of the input feature maps with a learned kernel is performed and then the results are passed into a non-linear activation function like the sigmoid, tanh and Relu [20] . With this, many different features are obtained.

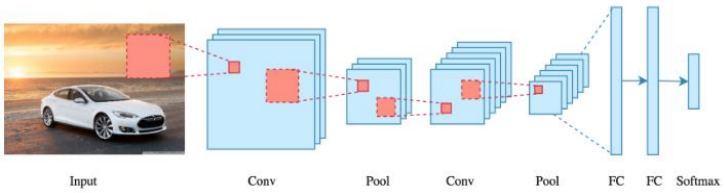

Figure 1: CNN layers [21]

Decision making in CNN is governed by these hyperparameters: kernel size, filter count, filter steps or stride, and the padding. Stride refers to the step-size of the convolution filter with a default value of unity. This means movement of one pixel at a time. Increasing stride makes the filter to jump with larger pixel intervals resulting to a lesser overlap between cells. The size of the feature map is always smaller than the input and to prevent it from shrinking, a layer of zero-value pixels is added to surround the input with zeros (padding). This improves the performance of the CNN and makes sure that the kernel and stride size will fit in the input [22].

The pooling layer comes next after the convolution layer. Its task is to continuously reduce the dimensionality to decrease the number of parameters that affects the computation time in the network. In effect, training time is reduced, and overfitting is controlled. Max pooling is the commonly used technique as it takes the maximum value in each window as shown in Figure 2. Feature map size is decreased while significant information are preserved [22]. The pooling layer functions as a secondary feature extraction layer that reduces dimensionality of features and increases the robustness of the feature extraction capabilities of the CNN. High level characteristics can be obtained by stacking several layers of convolutional and pooling layers [14].

In the classification layer, all neurons in the previous layers are connected to every single neuron of the present layer. Neurons in a fully connected layer have full connections to all activations in the previous layer. The last fully-connected layer is followed by an output layer. CNN's are trained using backpropagations or gradient descent. For tasks like classification, softmax regression is commonly used because it generates a well-performed probability distribution of the outputs [14]. Other methods can be used like support vector machines (SVM).

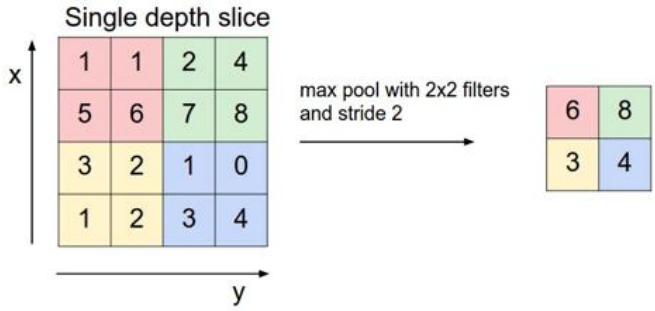

Figure 2: Sample of max pooling [22]

3. METHODOLOGY

Figure 3 shows the block diagram of the system. A video feed will be provided by the camera and once a person is detected, it will proceed to recognition and then classify whether the person is authorized or not. If the person is authorized, the system will just be passive, log time and anonymized name, and continue to accept video feeds. On the other hand, the system will record video and image information which includes the face, height and built of the person. Afterwards, the video feed will continue.

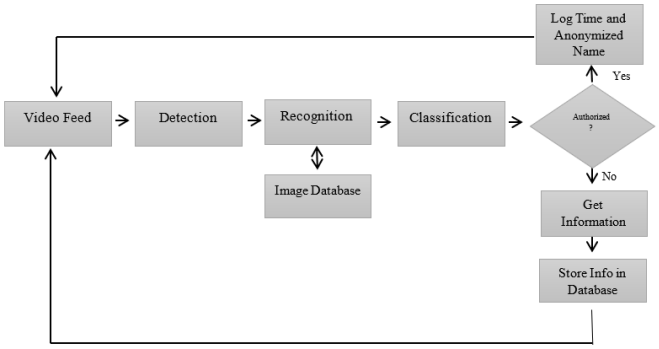

Figure 3.: System Block Diagram

3.1 Participants

A total of thirty participants were invited in which three of them are considered authorized and the rest are unauthorized personnel. The participants are eighteen years of age and above and there is no restriction on the number of male or female participants. The participants were asked to randomly enter a room where the system is in place. Different scenarios with many potential participants were considered as well as the aspect of gender skewing.

3.2 Data Gathering Procedures

The IP camera was placed facing the door to strategically be able to capture the face of the person coming in clearly. It stands approximately 5ft and is accessed using a laptop computer via a router. The processing and monitoring of video feeds are performed using a CNN-based algorithm that detects if a person is present or not. Whenever a person is detected, it analyzes whether the person is authorized/unauthorized. If authorized, the system will record the time and the anonymized name of the person. If unauthorized, the system will record the time, height, body built, will take a snapshot of the person as well as the video clip and will send an alarm or notification.

A graphical user interface (GUI) was created to show the live feed of the video. The GUI features (a) the live feed, (b) the log of the authorized person, (c) the log of the unauthorized person with the height, built and snapshot. The performance of the system was assessed based on its accuracy, F1 score in particular. Confusion matrices were used to further evaluate the performance of the system. Ethical considerations in terms of the image data gathered was considered. A written consent was obtained from the participants for the purpose of data privacy and security.

4. RESULTS AND DISCUSSION

The GUI of the system is shown in Figure 4. It is a straight-forward display that shows the video feed and the details of the person detected. The classification is either Authorized (appearing with green colored texts) or Unauthorized (appearing with red colored texts). The date and time are synchronized with the system clock of the computer. These information are logged whenever an authorized or unauthorized person is detected. A name code was used for authorized personnel and is placed together with the date and time of their entry for data logging. If an unauthorized personnel is detected, the date, time, height, and built will be logged into the data log. In addition, a snapshot of the person will be taken and will displayed on the screen. An alarm will be sent to the authorities leaving the "BREACH" button blinking. A mechanical or electronic lock could be implemented to secure the area though it not covered by this study.

Face detection is highly influenced by the environment which includes illumination, background, distance, and accessories to name a few. Nonetheless, this was dealt with the CNN-based algorithm trained with video feeds that cover different entry scenarios. Similarly, the height and built estimate algorithm was also trained to match the actual physical measurement. This is also correlated with the images taken that shows the image of the unauthorized person.

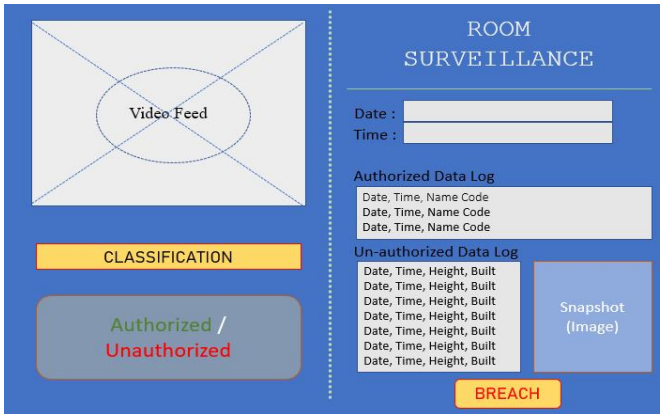

Figure 4: Room Surveillance GUI

Table 1 shows the precision, recall, and the F1 scores which are derived from the confusion matrices. The estimate and image snapshot correlation coefficients are also presented.

Table 1: System Task's Precision, Recall and F1-score

Task	Precision	Recall	F1-Score	R
Face Recognition	0.903	0.921	0.912	-
Height Estimation	0.988	0.974	0.981	0.91
Built Estimation	0.982	0.971	0.976	0.87

Precision indicates the correctly positive observation to the total predicted positive observations. The system shows good precision as correctly recognized faces and measures match the actual personnel considered and the height and built of the participants. It is also noticeable that the Recall of the system significantly good that result to a good balance as indicated by the F1 score. The system was able to correlate well the height and built estimates with a correlation coefficient of 0.91 and 0.87, respectively.

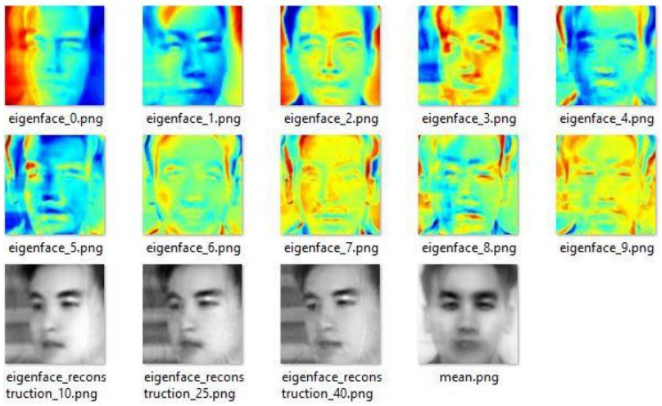

Figure 5: Male Eigenfaces and Image Reconstruction

Gender skewing was dealt by taking in consideration the features of a male from a female participant with their eigenfaces [25]. For this purpose, the CNN was trained with gender balanced VGGFace2 and MS1MV2 datasets [23]. Though there's a good balance of male and female participants, gender skewing data were obtained by repeatedly capturing the faces of the participants at random entry of either individually, with a partner or in groups. Figures 5 and 6 show sample male and female faces depicting their Eigen and reconstructed images.

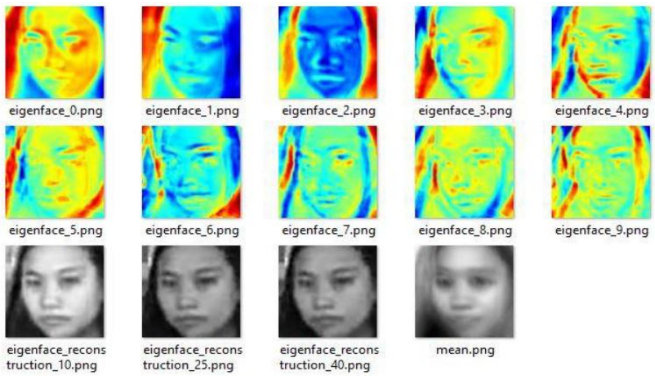

Figure 6: Female Eigenfaces and Image Reconstruction

Table 2 shows the gender skewness level of the system trained after the datasets, VGGFace2 and MS1MV2. Skewness levels between 0.5 to 1 or -0.5 to -1 are considered to be moderately skewed while if between -0.5 to 0.5, data are close to being symmetric. Results show that the gender skewness levels are close to being asymmetric showing a good balance of recognition between the male and female gender.

Table 2: Gender Skewness Levels

Gender	VGGFace2	MS1MV2
Male	0.35	0.22
Female	0.32	0.28

5. CONCLUSION

Security systems have become an integral part of different institutions and establishments. These systems were built in order to meet the demands of assets and personal protection. A CNN-based computer vision system was proposed which is equipped with human, face, and body detection for classifying whether a person entering a secured room is authorized or not. The system was capable of calculating the height and body built of a person and was able to compare it to the height and body built of the authorized person.

ACKNOWLEDGEMENT

This project is funded by the University Research Coordination Office (URCO) / College Research Fund / De La Salle University – Science Foundation.

REFERENCES

1. A, Jain, S. Basantwani, O. Kazi, and Y. Bang. **Smart surveillance monitoring system**, *International Conf. on Data Management, Analytics and Innovation*, pp. 269-273, 2017.
2. G. Theil. **Automatic CCTV surveillance – Towards the virtual guard**, *IEEE Aerosp. Electron. Syst.Mag.*, vol. 15, no. 7, pp. 3-9, 2000.
3. B. Welsh and F. David. **Effects of Closed-Circuit Television on Crime**, *Ann. Am. Acad. Pol. Soc. Sci.*, vol. 587, pp. 110–135, 2003. https://doi.org/10.1177/0002716202250802
4. L. Meinel, M. Findeisen, M. Hes, A. Apitzsch, and G. Hirtz. **Automated real-time surveillance for ambient assisted living using an omnidirectional camera**, *IEEE International Conference on Consumer Electronics*, pp. 369–399, 2014.
5. N. A. Abdullah, M. J. Saidi, N. H. A. Rahman, C. C. Wen, and I. R. A. Hamid. **Face recognition for criminal identification: An implementation of principal component analysis for face recognition**, *AIP Conference Proceedings*, vol. 1891, 2017.
6. K. Muhammad, J. Ahmad, I. Mehmood, S. Rho, and S. W. Baik. **Convolutional Neural Networks Based Fire Detection in Surveillance Videos**, *IEEE Access*, vol. 6, pp. 18174-18183, 2018.
7. R. K. Sadykhov, D. V Lamovsky, V. A. Kharlanov, and A. S. Kirienko. **Combined approach for face frontal view estimation for video surveillance purposes**, *Proceedings of the 5th IEEE International Workshop on Intelligent Data Acquisition and Advanced Computing*

Systems: Technology and Applications, pp. 430-433, 2009.

8. S. Purbaya, E. Ariyanto, D. W. Sudiharto, and C. W. Wijiutomo. **Improved Image Quality on Surveillance Embedded IP Camera by Reducing Noises**, in *3rd International Conference on Science in Information Technology (ICSITech) Improved*, 2017, pp. 156–160.
9. L. Goldmann, A. Samour, M. Karaman, and T. Sikora. **Extracting high level semantics by means of speech, audio, and image primitives in surveillance applications**, in *ICIP*, 2006, pp. 2397–2400.
10. A. Sharifara, M. S. Mohd Rahim, and Y. Anisi. **A general review of human face detection including a study of neural networks and Haar feature-based cascade classifier in face detection**, in *Proceedings - 2014 International Symposium on Biometrics and Security Technologies*, 2015, pp. 73–78.
11. C. Ding and D. Tao. **Robust Face Recognition via Multimodal Deep Face Representation**, *IEEE Trans. Multimed.*, vol. 17, no. 11, pp. 2049–2058, 2015. https://doi.org/10.1109/TMM.2015.2477042
12. H. Li, Z. Lin, X. Shen, J. Brandt, and G. Hua. **A convolutional neural network cascade for face detection**, *in IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 5325–5334.
13. Z. Liu, P. Luo, X. Wang, and X. Tang. **Deep Learning Face Attributes in the Wild**, *in IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 3730–3738.
14. T. Guo, J. Dong, H. Li, and Y. Gao. **Simple convolutional neural network on image classification**, *in 2017 IEEE 2nd International Conference on Big Data Analysis*, 2017, pp. 721–724.
15. A. Giyenko, A. Palvanov, and Y. Cho. **Application of convolutional neural networks for visibility estimation of CCTV images**, *in International Conference on Information Networking*, 2018, pp. 875–879.
16. Y. LeCun, K. Kavukcuoglu, and C. Farabet. **Convolutional networks and applications in vision**, *in ISCAS 2010 - 2010 IEEE Int. Symp. Circuits Syst. Nano-Bio Circuit Fabr. Syst.*, pp. 253–256, 2010. https://doi.org/10.1109/ISCAS.2010.5537907
17. Cctvinstallers, "How does a CCTV Camera Works," 2012. [Online]. Available: https://www.cctv.co.uk/how-does-a-cctv-camera-work/.
18. A. Ziadi, X. Maldague, L. Saucier, C. Duchesne, and R. Gosselin. **Visible and near-infrared light transmission: A hybrid imaging method for non-destructive meat quality evaluation**, *Infrared Phys. Technol.*, vol. 55, no. 5, pp. 412–420, 2012.
19. "How IP Cameras Work: What is an IP Camera," 2018. [Online]. Available: https://www.protectamerica.com/home-security-blog/tech-tips/realblogging-how-ip-cameras-work_11072.
20. G. E. Hinton and V. Nair. **Rrte Amount**, *in Proceedings of the 27th International Conference on Machine Learning*, 2010, no. 3, pp. 807–814
21. A. Dertat, "Applied Deep Learning - Part 4: Convolutional Neural Networks," 2017. [Online]. Available: https://towardsdatascience.com/applied-deep-learning-part-4-convolutional-neural-networks-584bc134c1e2.
22. D. Cornelisse, "An Intuitive Guide to Convolutional Neural Networks," 2018. [Online]. Available: https://medium.freecodecamp.org/an-intuitive-guide-to-convolutional-neural-networks-260c2de0a050.
23. V. Albiero, K.S. Krishnapriya, K. Vangara, K. Zhang, M.C. King, and K.W. Bowyer. **Analysis of Gender Inequality in face rcognition accuracy**, in *arxiv, vol. 1*, 2020. https://doi.org/10.1109/WACVW50321.2020.9096947
24. M. Murugesan, M. Santhosh, T. Sasi Kumar, M. Sasiwarman, and I. Valanarasu. **Securing ATM Transactions using Face Recognition**, *International Journal of Advanced Trends in Computer Science and Engineering*, vol. 9, no. 2, March - April 2020. https://doi.org/10.30534/ijatcse/2020/59922020
25. H. Almohamedh, and S. Almotairi. **Facial emotion recognition using eigenface and feature optimization**, *International Journal of Advanced Trends in Computer Science and Engineering*, vol. 8, no. 4, July - August 2019. https://doi.org/10.30534/ijatcse/2019/28842019