Preserving Women Public Restroom Privacy using Convolutional Neural Networks-Based Automatic Gender Detection

D Kristiyani[1], **A W Wijayanto**[1,2]

[1] Politeknik Statistika STIS, Jl. Otto Iskandardinata No.64C, Jakarta, Indonesia
[2] BPS-Statistics Indonesia, Jl. Dr. Sutomo 6-8, Jakarta, Indonesia

*Corresponding author's e-mail: 221810237@stis.ac.id, ariewahyu@stis.ac.id

Abstract. Personal safety and privacy have been the significant concerns among women to use and access public restrooms/toilets, especially in developing countries such as Indonesia. Privacy-enhancing designs are unquestionably expected to ensure no men entering the rooms neither intentionally nor accidentally without prior notice. In this paper, we propose a facial recognition approach to ensure women's safety and privacy in public restroom areas using Convolutional Neural Networks (CNN) model as a gender classifier. Our main contributions are as follows: (1) a webcam feed automatic gender detection model using CNN which may further be connected to a security alarm (2) a publicly available gender-annotated image dataset that embraces Indonesian facial recognition samples. Supplementary Indonesian facial examples are taken from a government-affiliated college, Politeknik Statistika STIS students' photo datasets. The experimental results show a promising accuracy of our proposed model up to 95.84%. This study could be beneficial and useful for wider implementation in supporting the safety system of public universities, offices, and government buildings.

1. Introduction

In gender-diverse public spaces, male-female segregated restroom (toilet) design has an important requirement to preserve high standards of privacy and safety [18]. For instance, in a public women's restroom, it is expected to restrict men from entering the space during operational times unless under certain maintenance conditions [1]. Numerous recent studies have stated that public restrooms in developing countries commonly fail to provide basic safety instruments and are vulnerable to violent attacks, crimes, and harassment [17,23]. Further vulnerability is increasing in remotely accessed public spaces where limited security officers are available.

The United Nations Commission for the Status of Women mandates comprehensive approaches of Information & Technology (IT) to overcome unpleasant violence towards women and girls at schools, workplaces, parks, public transportation, etc [17]. Facial recognition technology offers an abundant potential advantage for achieving such goals. Retrieval of information from facial recognition can be in the form of a person's gender classification. This can be used to maintain women's privacy while in the women's restroom. The safety and comfort of toilet users are very important because it is the most gender-sensitive place. Several cases cause discomfort, for example, the mistake of a man entering the women's toilet or not being warned when the men's cleaning service enters the women's toilet. The most disadvantaged are women in terms of their privacy. Therefore, this study aims to design a mechanism that can protect women's privacy while in the women's toilet with gender classification through webcam-based facial recognition.

Face recognition is a type of biometric that identifies human characteristics based on faces or images. However, the information received for facial recognition is very susceptible to brightness or illumination, noise, and defects from the specifications of the camera lens used. To eliminate the lighting effect that hinders the face recognition system, a facial representation called Local Quantized Patterns (LQP) is proposed in [13]. Illumination invariance in the representation leads to improved performance for advanced methods in the Labeled Faces in the Wild (LFW) database [21]. In [7] a robust illumination normalization technique is proposed using Gamma correction, Gaussian difference filtering, masking, and contrast equalization. The preprocessing step improves recognition performance on some benchmark databases. In [8], a near-infrared image was used for face recognition regardless of the visible illumination changing in the environment.

Computer vision is a science and technology capable of acquiring to understanding an image or video streaming to retrieve information and make a decision [3]. The decision is in the form of gender classification based on gender detection from a person's face. So, computer vision is a combination of cameras, computers, and pattern recognition.

Digital image processing (image processing) is a method to improve the quality and extract information from the image. In [11], images of the same subject with different expressions are viewed as an ensemble of mutually correlated signals, and scattering accounts for the variation in expression. Humans can classify a person's gender by seeing it directly. However, computers need a lot of data to learn the characteristics or specifications of gender based on the person's face. Variations in face position are needed to study the angle of face tilt and the distance of the human face to the camera [15]. Another critical concern is that numerous existing studies discussed the racial bias of the facial recognition model where facial samples from Western countries are more represented. Darker color people such as those from Africa and Asia, including Indonesia, have less representation on the publicly available annotated dataset. Hence, in this study, we use not only datasets of male and female facial samples from several photo angles taken from the Kaggle website, but also introduce supplementary gender-annotated Indonesian samples taken from a student photos dataset of a government-affiliated college, Politeknik Statistika STIS, especially Batch 60 students. A total of 2,763 images have been grouped into 2 folders, the female folder with 1,289 images and the male folder with 1,474 images

2. Research Method

2.1. Data and Preprocessing

As shown in Figure 1, the research begins by collecting datasets in the form of images or faces. which can be seen in Figure 2 and Figure 3, the pre-processing is carried out. The process of data transformation is to change the data into a form that is suitable for the next process. This study uses one of the Deep Learning algorithms, namely the Convolutional Neural Network (CNN) method. This method accepts input in the form of an image from the pre-processing dataset. Furthermore, it is processed to determine aspects in recognizing and distinguishing one image from another.

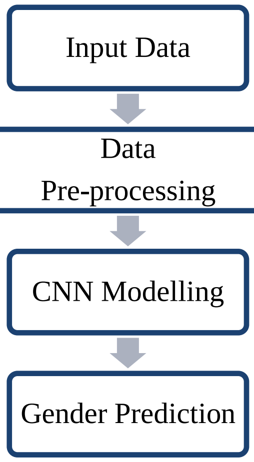


Figure 1. Research stage.

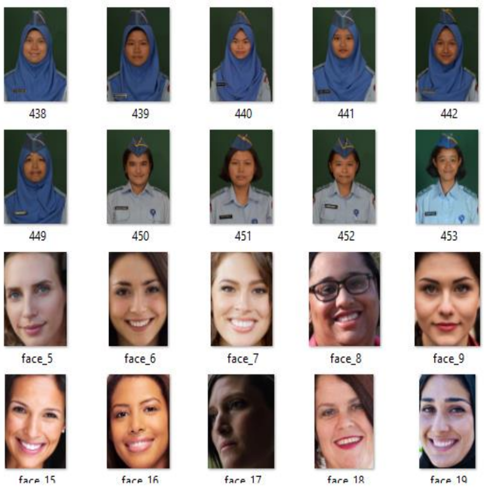


Figure 2. Female image dataset.

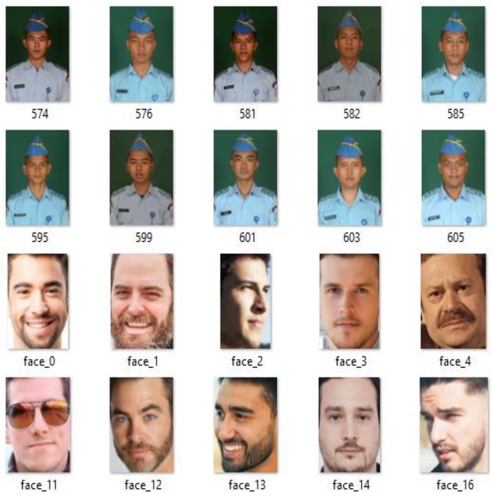


Figure 3. Male image dataset.

2.2. Prediction Model using Convolutional Neural Networks (CNN)
Convolution prediction is an operation process to generate a feature map or convolved feature by using the kernel repeatedly [22]. The kernel moves from top left to bottom right. There is a convolution operation on two-argument functions that have real values.

$$S(t) = (x.t)(t) = \sum_{\alpha=-\infty}^{\infty} x(\alpha).w(t-\alpha) \tag{1}$$

Formula (1) explains that S(t) is a single output in the form of a feature map, x is an input of 3x3 size, and w is a weight (kernel). Determination of the output volume can also be determined from each layer with hyperparameters. Hyperparameter is used to calculate the number of activation neurons in one output [10].

$$The\ spatial\ size\ of\ the\ output\ volume = \frac{W-F+2P}{S+1} \tag{2}$$

Formula (2) explains that W is the size of the image volume, F is the size of the filter, P is the padding value used, and S is the size of the shift (stride). Pooling is the process of reducing the image size. The process consists of a filter with a certain size and stride that will alternately shift over the entire feature map area. A feature is a component in a convolutional layer that contains a weight matrix (kernel). If there are 6 filters, then each filter has 3 kernels with a size of 5x5. Stride is the number of filter shifts in the convolution process. If the stride value is 1, then the filter will shift by 1 pixel horizontally and vertically. Padding is the addition of pixel size with a certain value around the input data so that the result of the receptive field (sensory space) is not too small. Adaptive Moment Estimation (ADAM) is a type of optimizer for CNN. The optimizer functions to update the weights to minimize the loss function.

This study has a target which is a label or class for gender classification decisions whether it is 1 for women or 0 for men. The sigmoid activation function is performed on the CNN output vector before the loss calculation. Using binary cross-entropy loss because of the type of binary classification on its output. Processing the algorithm using the OpenCV and Tensorflow packages in Python. OpenCV is used as the main program between the webcam and the computer to process dynamic images in real-time. Meanwhile, Tensorflow is a computational framework for building CNN models.

$$f(s_i) = \frac{1}{1+e^{-s_i}} \tag{3}$$

$$CE = -\sum_{i=1}^{C'=2} t_i \log(s_i) = -t_1 \log(f(s_1)) - (1-t_1)\log(1-f(s_1)) \tag{4}$$

At the prediction stage, the accuracy and loss of the training data and validation data are calculated. Formulas (3) and (4) explain that $f(s_i)$ is a sigmoid activation function for the CNN output layer, s_i is the CNN output vector. While CE is Cross Entropy with C`=2 (binary), namely classes C_1 and C_2, t_1 is the interval of [0,1], s_1 is ground truth (fundamental truth) and the value of class C_1, t_2 is (1- t_1), and s_2 is (1- s_1) or ground truth and grade C_2. Figure 4 illustrate the CNN layers.

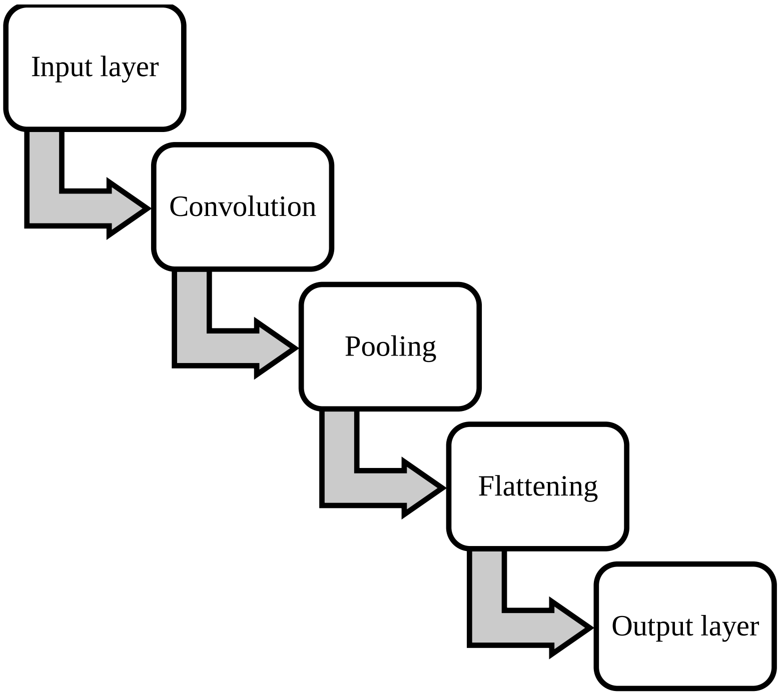


Figure 4. Structure of CNN layers.

2.3. Related Work

The currently existing literatures have been investigating the utilization of Convolutional Neural Network based gender detection model. Table 1 provides the summary of relevant studies related to our work. Our study is focusing on the use of the student photos dataset of a government-affiliated college, Politeknik Statistika STIS, especially Batch 60 student in this study has a unique image because it contains the faces of various ethnic groups in Indonesia. This is because STIS accepts students from various parts of the country.

Table 1. Summary of Related Work.

No	Study	Objective	Research Gap
1	Using Convolutional Neural Networks to Discover Cogntively Validated Features for Gender Classification [19]	In this study, the classification results are compared with different regularization techniques and other standard classifiers. The CNN models yield higher accuracy (88.46%) than both SVMs (Support Vector Machines) and Random Forest classifiers.	Discusses how to improve CNN accuracy by flipping and L2 Regularization. However, the accuracy obtained is still below 90%.
2	Age, Gender, and Fine-Grained Ethnicity Prediction using Convolutional Neural Networks for the East Asian Face Dataset [14]	This study focuses on predicting the age, gender, and fine-grained ethnicity of an individual by providing baseline results. It provides baseline results using a convolutional neural network (CNN) on the Wild East Asian Face Dataset (WEAFD).	Discussing inter-class variation of facial attributes across East Asian women and men. However, the accuracy of gender prediction is still below 90%, which is 88.02%.
3	CNN Based Aerial Image Processing Model for Women Security and Smart Surveillance [4]	This study intended to deliver serious protection encompasses violence against women and sexual or gender-based violence. This project helps in utilizing drone technology in an effective way to solve different issues of society by proposing a CNN-based image processing model.	The model formed has a high precision of 98.63%. However, it does not explain the accuracy obtained because it is quite accurate only on high-quality input with proper lighting conditions.
4	Implementation of Machine Learning for Gender Detection using CNN on Raspberry Pi Platform [5]	This study proposed system describes gender detection based on Computer Vision and Machine Learning Approach using CNN. The whole system is introduced by Raspberry Pi programmed using Python.	The research discusses real-time applications for gender detection to close the gap of Google Cloud Vision technology. It also proposes the system for the field of security in public but does not explain in detail its use.

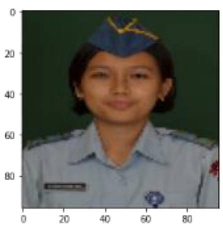
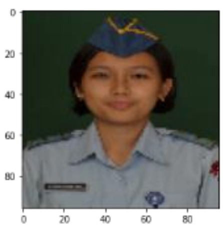
The classification program is becoming increasingly rich in information on the uniqueness of the male and female gender faces from all over the archipelago. At the Overseas, the application of facial recognition in toilets was successfully carried out by China in 2017. According to the BBC.com website [2], China makes tissue dispensers with facial-recognition toilet paper dispensers to limit how much paper a person can take. Therefore, the idea of women's toilet safety can be a new invention that can be developed again and become a pioneer of women's toilet safety technology by an embedded system.

3. Result and Evaluation

3.1. Input Layer

The way the input in the form of an image to a computer works is by converting it into an array. The standard image is known as an RGB (Red Green Blue) image and has a 24-bit color system consisting of 8 bits each [3]. The color representation of the integer type varies from 0-255, while the float type ranges from 0-1. The color value is stored in a 3-dimensional matrix which is denoted in intensity I(X, Y, channel). The dimensions of the array depend on the resolution of the image. In [12], a 3D model for each subject in the dataset was constructed using a single 2D image. Since faces are usually detected by automatic face detectors [20], the cropped faces were not aligned. However, since most feature annotators require alignment before feature extraction, the discrepancy lowers the performance value of the facial recognition system. The pixel representation can be in the form of binary (black and white), greyscale (grey), color, and many grooves (multi-channel) images. This study processes datasets with color images.

Table 2. Data pre-processing.

	Image	Array	Size
Input		array([[[24, 41, 27], [25, 42, 28], [25, 42, 28], : ... : [141, 137, 132], [140, 137, 129], [141, 137, 132]]], dtype=uint8)	2963 pixel high, 2222 pixel width, and 3 dimension
Resize		[[[26 43 29] [26 43 29] [24 42 27] : ... : [138 135 127] [140 136 131] [140 136 131]]]	96 pixel high, 96-pixel width, and 3 dimension
Rescale		[[[0.10196078 0.16862745 0.11372549] [0.10196078 0.16862745 0.11372549] [0.09411765 0.16470588 0.10588235] : ... : [0.54117647 0.52941176 0.49803922] [0.54901961 0.53333333 0.51372549] [0.54901961 0.53333333 0.51372549]]]	96 pixel high, 96 pixel width, and 3 dimension

It can be concluded that face alignment can reduce the performance of the facial recognition system suddenly [6]. Before entering the input layer, pre-processing the dataset which can be seen in Table 2 is carried out, namely rescale or dividing all pixel values by the largest value (255) and producing an interval [0,1] or float type. This normalization is used so that the training process becomes more effective because synaptic weights on neurons are usually generated at small intervals and make it easier for the network to carry out the learning process.

3.2. Convolution→ Pooling → Flattening

This CNN method uses a Multi-Layer Perceptron or Fully Connected Layer model because it combines all the features to form the final model. Using the Feedforward training algorithm which only moves in one direction, namely forward and there are no cycles or cycles in the network (network). The convolutional layer is CNN's first layer which is used to store pixel values from image to array and extract image features. ReLU (Rectified Linear Unit) is an activation function that converts the normalized negative pixel value to zero and other values are kept constant. The pooling layer is the process of reducing the image size as much as possible by reducing the total number of nodes in the next layer [9]. In the flattening process, the pooling data in the form of a 2-dimensional array is converted into 1-dimensional single vector data. Dense is a function to add a fully connected layer. This study uses 1024 units of the number of nodes that must be in the hidden layer. The output layer is the final classification result of whether an image has a value of 1 or 0. The sigmoid function (also known as the logistic function) is a basic activation function that assigns values from 0 to 1. Therefore, this function is appropriate to use in the output layer that produces binary numbers. The summary of the CNN layer can be seen in Table 3 and the CNN architecture can be seen in Figure 5.

Table 3. Summary of CNN layers.

Layer (type)	Output Shape	Param #
conv2d_5 (Conv2D)	(None, 96, 96, 32)	896
max_pooling2d_3	(MaxPooling2 (None, 32, 32, 32)	0
conv2d_6 (Conv2D)	(None, 32, 32, 64)	18,496
conv2d_7 (Conv2D)	(None, 32, 32, 64)	36,928
max_pooling2d_4	(MaxPooling2 (None, 16, 16, 64)	0
conv2d_8 (Conv2D)	(None, 16, 16, 128)	73,856
conv2d_9 (Conv2D)	(None, 16, 16, 128)	147,584
max_pooling2d_5	(MaxPooling2 (None, 8, 8, 128)	0
flatten_1 (Flatten)	(None, 8192)	0
dense_2 (Dense)	(None, 1024)	8,389,632
dense_3 (Dense)	(None, 2)	2,050

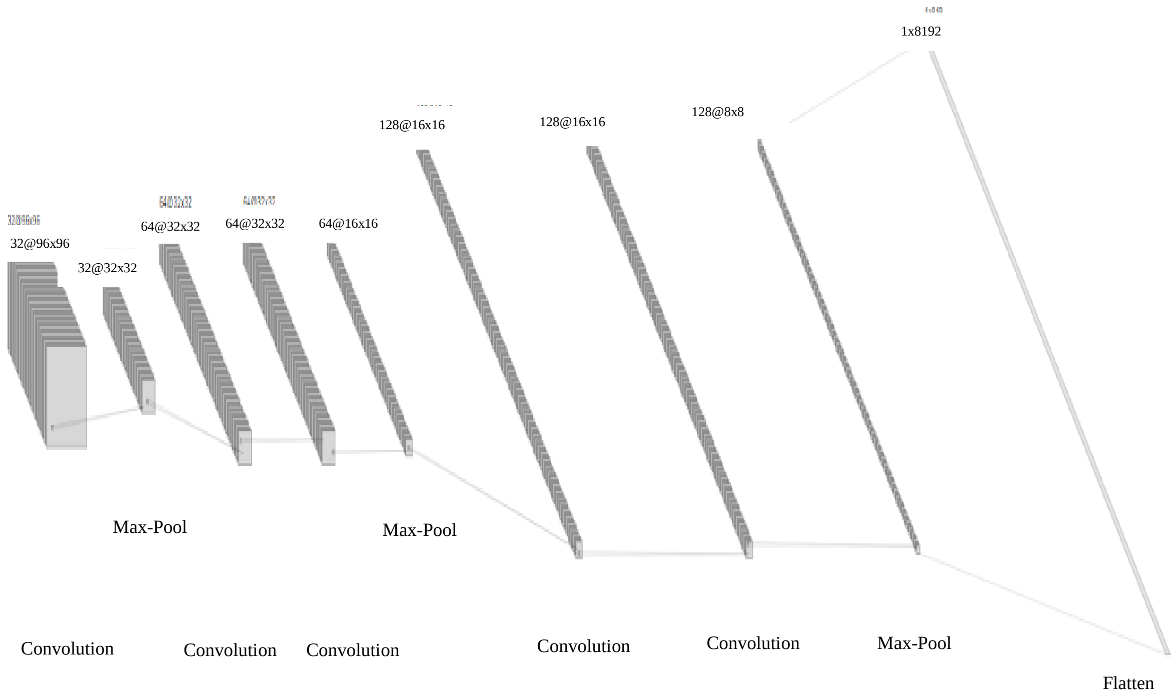


Figure 5. Proposed CNN architecture.

3.3. Output Layer

Determining the learning rate greatly affects the training process. If it is too large it will decrease the MSE until it rises and falls uncontrollably, while if it is too small it will decrease the MSE slowly [16]. Therefore, this study set a learning rate of 0.001. The number of iterations or epochs of 100 is used as a stop condition during the training process. To avoid overfitting is to divide the dataset into two, namely 20% (553 images) for testing data and 80% (2210 images) for training data with a size of 96 x 96 pixels and color. The training data is used to build the model, while the testing data is used to calculate the accuracy of the model that has been built.

Figure 6. Experiment with a laptop webcam in real-time.

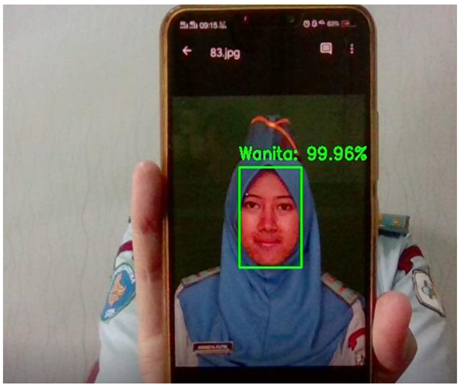


Figure 7. Experiment with an image of a female in hijab with a laptop webcam in real-time.

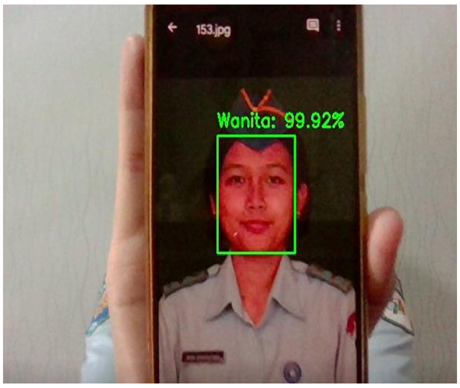


Figure 8. Experiment with an image of a female with a laptop webcam in real-time.

Figure 9. Experiment with an image of a male with a laptop webcam in real-time.

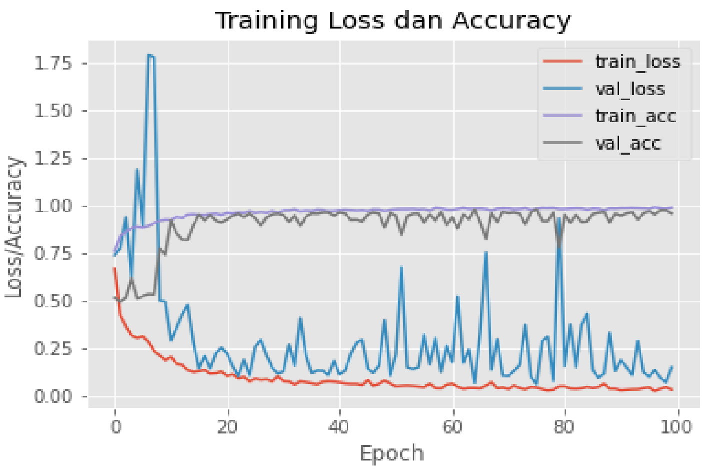


Figure 10. Loss and accuracy graph.

In the experiment with the webcam, the results of gender classification were obtained. Accuracy is calculated based on the maximum confidence interval of the prediction model. Hence, it can classify the gender of women (*wanita*) or men (*pria*) correctly. The model as illustrated in Figure 10 is already quite good because the validation loss is 0.1510 and the training loss is 0.0274 which is almost close to the value of 0.

Table 4. Loss calculation and accuracy.

Epoch	Loss	Accuracy	Validation Loss	Validation Accuracy
1	0.8528	0.6988	0.7408	0.5172
2	0.4331	0.8390	0.7764	0.4937
3	0.3747	0.8649	0.9379	0.5172
…	…	…	…	…
99	0.0425	0.9858	0.0721	0.9747
100	0.0274	0.9921	0.1510	0.9584

Table 4 demonstrates the calculation of loss function and performance accuracy. The training accuracy in the 100th epoch has reached 0.9921 or 99.21%, while the validation accuracy is 0.9584 or 95.84%. Because the accuracy is close to 100% of the model created by the training data, the gender classification is very good.

4. Practical Implementation Design of Privacy Preserving Model

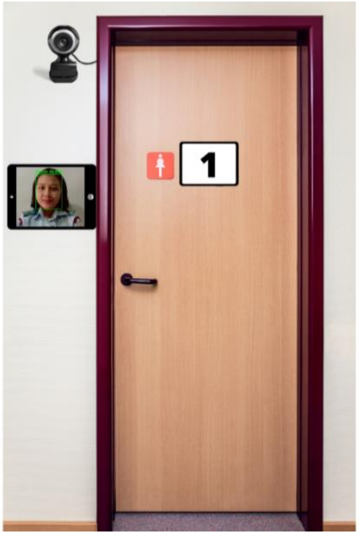


Figure 11. Female restroom exterior's mockup.

Figure 12. Female restroom interior's mockup.

To further enhance the usability of Convolutional Neural Networks model as gender classifier in practical implementation, we construct a mockup design of how the webcam feed model can be installed at public restroom. In Figure 11, there is a webcam and monitor that shows the number of female faces detected entering the restroom. When the woman comes out of the restroom, the monitor will show a reduced number. Meanwhile, in Figure 12, there is a speaker that will sound "male enters the restroom" when a male's face is detected entering the restroom. The mockup design can further provide a security warning for women who are in the restroom. Thus, the existence of this system can help male cleaning services to see whether the female restroom they are going to clean is empty or not and for mutual convenience.

In the application of face counting, there will be a reduction number when a woman's face comes out of the restroom. The idea of reducing this number is if a woman's face is detected a second time then the number decreases. But in practice, when women come out of the restroom, we cannot point the camera into the restroom for privacy reasons. Therefore, it is necessary to add a motion detection algorithm to the camera in front of the restroom door. Hence, the number on the monitor will increase when a woman's face is detected entering the restroom and will decrease when the motion sensor detects someone coming out of the restroom. The monitor can be a smart tablet, mini LCD, or another hardware. Furthermore, the design of this system can replace the webcam with an embedded system (a low-power microprocessor-based computer intelligent system that is capable of performing specific tasks). One of the embedded systems that are often used is Closed Circuit Television (CCTV). However, the Rasberry Pi can be the best choice because it is economical and can handle our proposed Convolutional Neural Network model on the OpenCV-based image processing. Figure 13 shows a schematic illustration of the utilization of the embedded system to implement our privacy preserving automatic gender detection model.

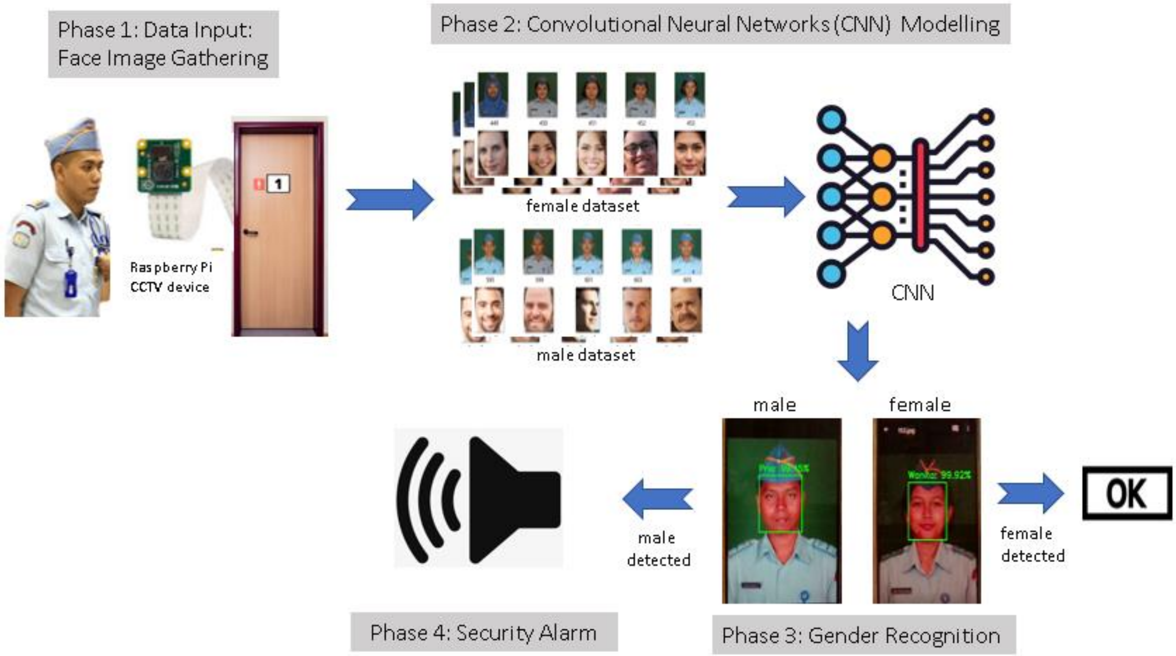


Figure 13. Schematic illustration of model implementation.

5. Conclusion

Based on the extensive evaluation results, the performance of our proposed CNN model is quite good with validation loss and training loss approaching the value 0. Gender classification can be performed with validation accuracy up to 95.84%. While the accuracy of the results of gender detection with a webcam can be calculated with the maximum confidence interval of the prediction model.

Our proposed model and system idea mockup can be further developed into an integrated embedded system to be implemented in wider public places. The system is able to detect facial features and classify gender correctly, to ensuring women safety and comfortability. Additionally, the mockup system is designed to emit a "male entering the restroom" sound when a woman is in it through the speaker. This can be done by counting the number of women's faces that accumulate when entering or leaving the restroom and can be displayed on the monitor screen. When a man enters the restroom and the system detects a woman in it, a warning sound will sound. With this system, women can find out whether the restroom is empty or there are people. Men can also respect women's privacy more so that they are not careless and even careful not to enter the women's restroom. A company that only employs male cleaning services can benefit from the convenience of its employees. The male cleaning service can find out if there are still people or it is deserted by looking at the number of users on the monitor screen in front of the entrance to the women's restroom.

References

[1] ASEAN 2016 ASEAN Public Restroom Standard (Jakarta: ASEAN Secretariat)

[2] British Broadcasting Corporation 2017 Beijing Park Dispenses Loo Roll using Facial Recognition. Retrieved from BBC: https://www.bbc.com/news/world-asia-china-39324431

[3] Budiharto W 2018 *Pemrograman Python untuk Ilmu Komputer dan Teknik* (Yogyakarta: Andi)

[4] Dandamudi A G B, Vasumithra G, Praveen G and Giriraja C V 2020 CNN Based Aerial Image processing model for Women Security and Smart Surveillance *3rd Int. Conf. on Smart Systems and Inventive Technology (ICSSIT)* pp 1009-1017

[5] Gauswami M H and Trivedi K R 2018 Implementation of Machine Learning for Gender Detection using CNN on Raspberry Pi Platform *2nd Int. Conf. on Inventive Systems and Control (ICISC)* pp 608-613

[6] Hussain, Sibt, Napoleon, Thibault, Jurie and Frederic 2012 Face Recognition using Local

Quantized Patterns *BMVC 2012 - Electronic Proc. of the British Machine Vision Conference* pp 99.1-99.11

[7] Lei Z, Pietikainen M and Li S 2014 Learning Discriminant Face Descriptor *IEEE Transactions on Pattern Analysis and Machine Intelligence* pp 289–302

[8] Li S, Chu R, Liao S and Zhang L 2007 Illumination Invariant Face Recognition Using Near-Infrared Images *IEEE Transactions on Pattern Analysis and Machine Intelligence* pp 627–639

[9] Mada F U 2021 Convolutional Neural Network. Retrieved from Universitas Gadjah Mada Menara Ilmu Machine Learning: https://machinelearning.mipa.ugm.ac.id/tag/convolutional_neural_network/

[10] Nada M 2021 Penerapan Deep Learning Menggunakan Convolutional Neural Network (CNN). Retrieved from Medium: https://medium.com/@mukhlishatunnada02/penerapan-deep-learning-menggunakan-convolutional-neural-network-cnn-d02dc6532f5b

[11] Nagesh P and Li B 2009 A Compressive Sensing Approach for Expression Invariant Face Recognition *IEEE Conf. on Computer Vision and Pattern Recognition* pp 1518–1525

[12] Prabhu U, Heo J and Savvides M 2011 Unconstrained Pose-Invariant Face Recognition using 3D Generic Elastic Models *IEEE Transactions on Pattern Analysis and Machine Intelligence* pp 1952–1961

[13] Shan S, Chang Y, Gao W, Cao B and Yang P 2004 Curse Of Misalignment in Face Recognition: Problem and a Novel Misalignment Learning Solution *6th IEEE Int. Conf. on Automatic Face and Gesture Recognition* pp 314–320

[14] Srinivas N, Atwal H, Rose D C, Mahalingam G, Ricanek K and Bolme D S 2017 Age, Gender, and Fine-Grained Ethnicity Prediction Using Convolutional Neural Networks for the East Asian Face Dataset *12th IEEE Int. Conf. on Automatic Face & Gesture Recognition (FG 2017)* pp. 953-960

[15] Suharso Aries 2017 Pengenalan Wajah Menggunakan Metode Viola-Jones dan Eigenface dengan Variasi Posisi Wajah Berbasis Webcam *Techno Xplore : Jurnal Ilmu Komputer dan Teknologi Informasi*

[16] Suyanto 2014 *Artificial Intelligence* (Bandung: Informatika)

[17] UN Women 2018 Fourth UN Women Safe Cities and Safe Public Spaces Global Leaders'Forum: *Proc. report Shaw Conf. Centre, Edmonton, Alberta, Canada*

[18] UNICEF, WaterAid and WSUP 2018 Female-Friendly Public and Community Restrooms: A Guide for Planners and Decision Makers (WaterAid: London, UK). Available at http://washmatters.wateraid.org/female-friendly-restrooms

[19] Verma A and Vig L 2014 Using Convolutional Neural Networks to Discover Cogntively Validated Features for Gender Classification 2014 *Int. Conf. on Soft Computing and Machine Intelligence* pp. 33-37

[20] Viola P and Jones M J 2004 Robust Real-Time Face Detection *Int. J. on Computer Vision* pp 137–154

[21] Wang H, Kang B and Kim D 2013 PFW: A Face Database in the Wild for Studying Face Identification and Verification in Uncontrolled Environment *2nd IAPR Asian Conf. on Pattern Recognition* pp 356-360

[22] Wijayanto AW, Choong JJ, Madhawa K and Murata T 2019 Towards Robust Compressed Convolutional Neural Networks *6th IEEE Int. Conf. on Big Data and Smart Computing*

[23] Yan R, et al. 2021 Operating Status Of Public Restrooms in the Hutong Neighborhoods of Beijing: An Empirical Study *J. of Environmental Management* p 287