Face recognition system using Viola Jones, histograms of oriented gradients and multi-class support vector machine

Face recognition system using Viola Jones, histograms of oriented gradients and multi-class support vector machine

M Arafah[1,2], A Achmad [3], Indrabayu [4] and I S Areni[5]

[1]Doctoral Student of Electrical Engineering Department, Hasanuddin University, Makassar, Indonesia
[2]Informatics Study Program STMIK AKBA Makassar, Indonesia
[3]Department of Electrical Engineering, Hasanuddin University, Makassar, Indonesia
[4]Department of Informatics, Hasanuddin University, Makassar, Indonesia
[5]Department of Electrical Engineering, Hasanuddin University, Makassar, Indonesia

arafah@akba.ac.id, andani@unhas.ac.id, indrabayu@unhas.ac.id, intan@unhas.ac.id

Abstract. This study aims to determine the best distance to install a CCTV camera in identifying a person's face in the passenger inspection area at the airport, which can be developed for suspect detection systems. The training data used is in the form of an image with five different angles per person, while the testing data is in the form of video. The initial stage conducted is face detection based on local features (a pair of eyes, nose, and mouth) in the input data using the Viola-Jones method. The face detection results process with pre-processing, feature extraction, and classification. In the pre-processing stage, the Brightness Enhancement (BE), Grayscale and Contrast Limited Adaptive Histogram Equalization (CLAHE) methods are used to improve the quality of the detected face. Furthermore, the feature extraction and classification stages use the Histogram of Oriented Gradient (HOG) and the Multi-class Support Vector Machine (MSVM) methods, respectively. The result shows that the best accuracy obtained is 86.76% for a CCTV camera distance of 300 cm and a height of 250 cm.

1. Introduction

Terrorist activities involving Indonesian citizens have raised people's fears and have an impact on political life, economics, socio-culture, security and public order, national security, and international relations. In anticipation of acts of terrorism, the government issued a law on the eradication of criminal acts of terrorism which is also included in the National Research Master Plan of the Ministry of Research, Technology and Higher Education of the Republic of Indonesia in 2017 - 2045 [1].

The identification of terrorist perpetrators carried out today is based on the faces of the perpetrators. For this reason, the development of a terrorist detection system can begin with face recognition, with or without occlusion. In recent years, video-based face recognition has become a widespread concern and is one of the most important topics in research in the field of image processing in identifying a person's face [2]. Video-based face images, expressions, and scene recognition are fundamental problems in human-machine interaction, especially when videos have a quite short duration [3]. The appearance of faces captured with video cameras varies significantly due to changes in poses, illuminations, scales, blur, expressions, and occlusions [4]. In addition, human

faces in surveillance videos often experience blurry images, dramatic variations in poses, and occlusion [5].

Jia-Ching et al., in 2017 conducted a face detection based on global features and local features (a pair of eyes, nose, and mouth) using the Local Ternary Pattern (LTP) method, Principal Component Analysis (PCA), and Support Vector Machine (SVM). Face detection based on the global features obtained an accuracy of 93.26%, while the local features obtained an accuracy of 97.39%. In this study, the training data and testing data used are in the form of image [6].

There are usually several stages of pre-processing to improve the image quality, such as those carried out by Indrabayu et al. in 2017. In this study, the brightness enhancement and grayscaling method are used to determine a person experiencing a state of awake and a state of drowsy. The input data used consisted of validation data of 300 colored eye images from 4 people using the image matrix size of 40x95 pixels. The test results of system performance obtained an average accuracy of 93.5% [7].

Furthermore, in 2018, Zheng Xiang et al. used the Histogram of Oriented Gradient (HOG), Gabor wavelet, and Local Binary Pattern (LBP) methods for face detection with FERET Database in the form of images. The levels of accuracy obtained for the HOG, LBP, and GFC methods were 61.1%, 50%, and 49.1%, respectively [8].

Based on the results of previous studies, this study uses the brightness enhancement and grayscaling method to improve image quality before the processes of feature extraction and classification. Whereas for feature extraction, the HOG method is used, and the Multiclass Support Vector Machine (MSVM) method is used for classification. For the anti-terrorism system to be implemented at airports with high accuracy, then the first step in this research is to find the best conditions for the position of CCTV and the walk metal through detectors.

2. Proposed Method

The system design in this study is shown in Figure 1. The system software uses Matlab R2017a.

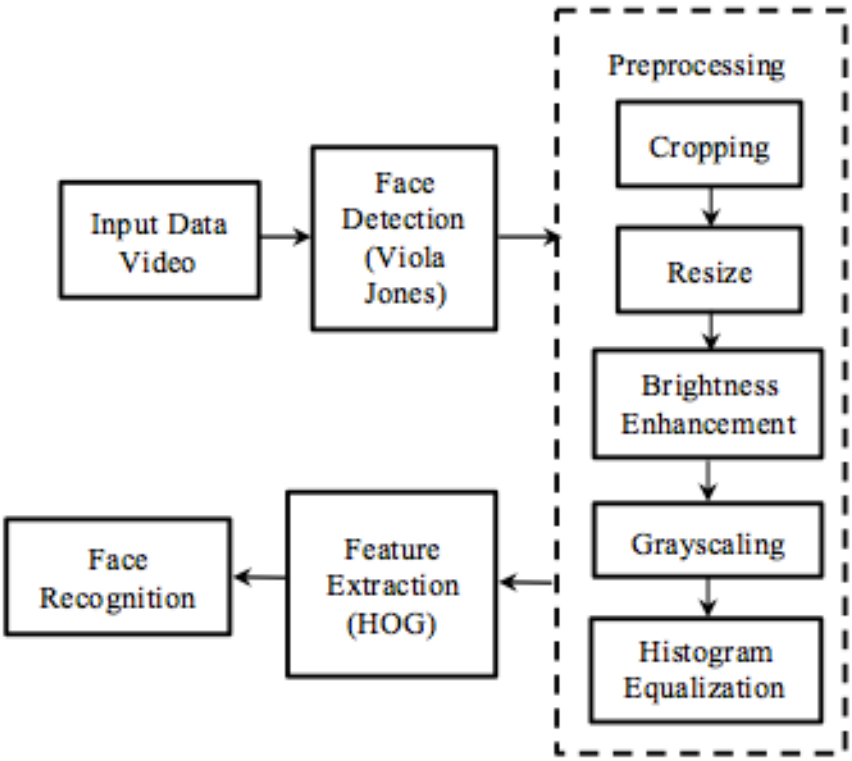


Figure 1. The proposed design system

The data used is divided into two parts, i.e., training data and testing data. Training data consists of five face images with five different angles with the size of 1920 x 1080 pixels, as shown in Figure 2. While the testing data used is in the form of video data from CCTV cameras. In taking video data, the camera position is placed with a height (h) of 250 cm and with a distance (d) from the walk-through metal detector varies from 200, 300, and 400 cm to get the best distance (d_{opt}) for the CCTV installation. The number of data testing frame for each distance is 750 frames for five people as objects. The illustration of data testing is shown in Figure 3. The proposed system is designed to detect suspects based on facial features in the passenger inspection area at the airport.

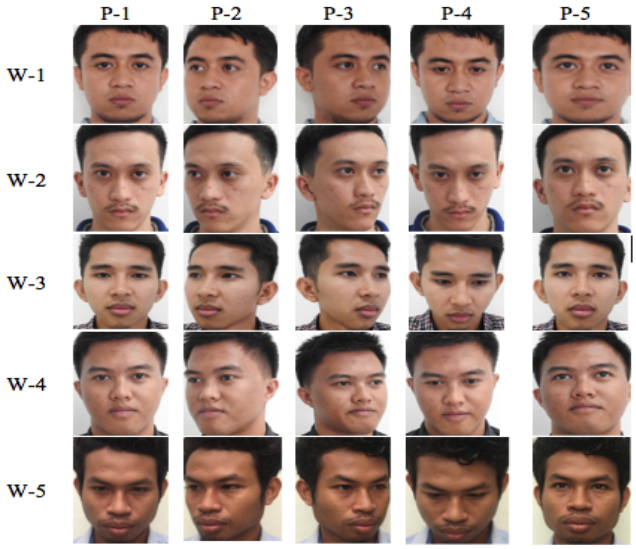


Figure 2. Training Data

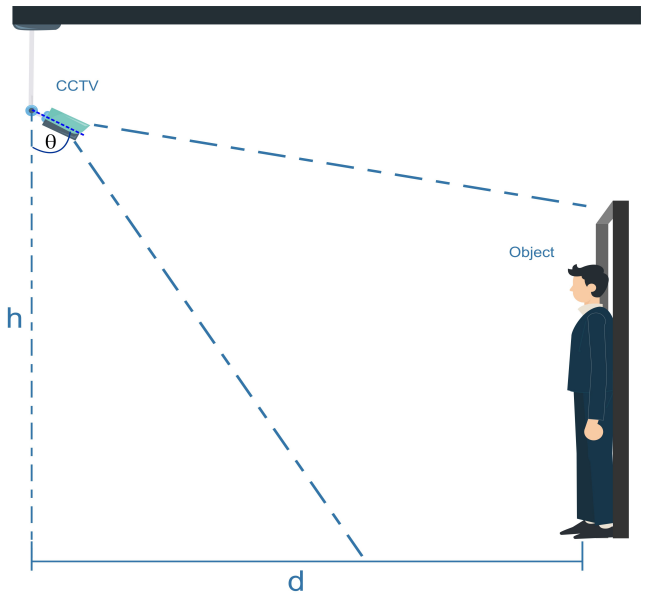


Figure 3. Illustration of testing data collection

Training data and testing data will be processed in stages, as shown in Figure 1. At the stage of input the training data, there are 5 faces where each face (W) has 5 angles (P), the first angle ($P1$) faces forward perpendicular, the second angle ($P2$) faces to the right with a slope of 15 degrees, the third angle ($P3$) faces left with a slope of 15 degrees, the fourth angle ($P4$) faces downward with a slope of 15 degrees, and the fifth angle ($P5$) faces upward with a slope of 15 degrees, as shown in figure 2.

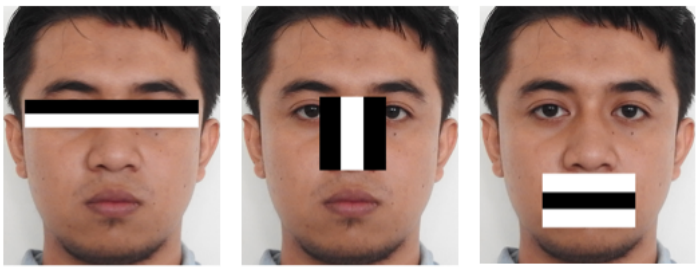
At the stage of inputting the testing data, there are three classifications of video data for each object based on the distance, 200 cm, 300 cm, and 400 cm. Furthermore, the video frame acquisition process in the form of an image will be carried out, after that the image will be processed using the Viola-Jones method to detect local features, like the eyes, nose, and mouth. Viola Jones combines four main keys in detecting features, such as Haar-like Features, Integral Image, AdaBoost Learning, and Cascade Classifier [9]. Those four stages are described as follows.

1. Haar-like features are features that are used to capture dark and bright areas based on high and low interval values. Examples of Haar-like features in the face area are shown in Figure 4.
2. Integral Image is an integral value calculation that carried out by referring to the Haar-like features in this stage that are formed to make it easier to get the difference between dark and bright areas. The total value of feature $F(H)$ is obtained by using the following equation.

$$F(H) = \sum F_{WH} - \sum F_{B} \quad (1)$$

Where $\sum F_{WH}$ is the feature value in the bright area and $\sum F_{B}$ is the value of the feature in the dark area. The value of $F(H)$ does not exceed the specified threshold.

3. Adaboost Machine Learning is the learning stage using the AdaBoost method by distinguishing strong classifications and weak classifications. At this stage, the value of the Haar-like feature will be calculated. If the value obtained is greater than the predetermined threshold value, the Haar feature is classified in AdaBoost learning, the process will continue until all haar features have been used.
4. Cascade Classifier is a stage to classify each classification of the Adaboost process in determining the face or not the face. This process utilizes several levels to ensure the face or not the face.



(a) (b) (c)

Figure 4. Haar-like features on the face area (a). Haar on eyes detection (b). Haar on nose detection (c). Haar on mouth detection.

In the face detection process, the threshold value is highly influential on the detection results. Determination of the incorrect threshold value will cause a greater detection error. After testing the threshold parameter values from 0 to 20, the optimum threshold value obtained is 10. Several examples of different threshold values are shown in Figure 5.

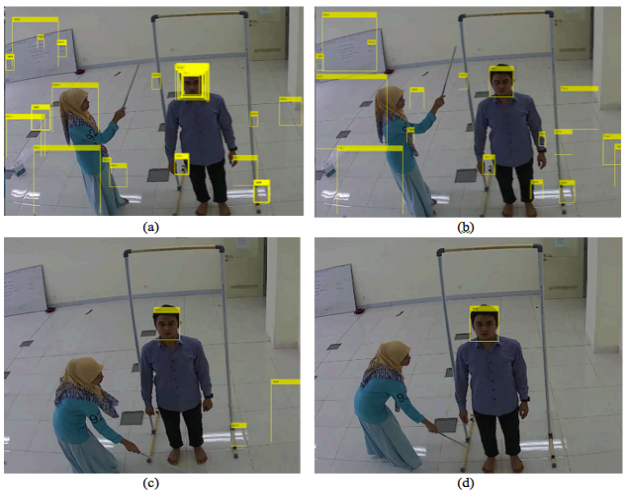


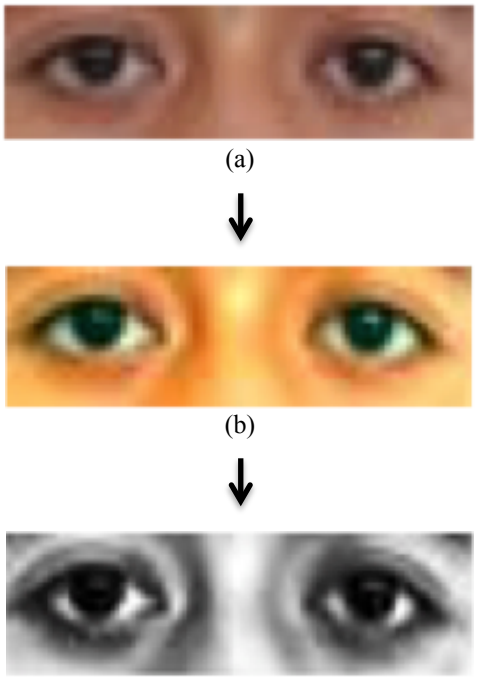
Figure 5. Examples of detection results based on different threshold values: (a) threshold = 0, (b) threshold = 1, (c). threshold = 5, (d) threshold = 10.

Figure 6. Example of cropping on pre-processing

To improve the quality of image detection results, then a preprocessing which consists of 5 stages are carried out, as shown in Figure 1, namely cropping, resizing, brightness enhancement, grayscaling and histogram equalization. These stages are described as follows.

1. Cropping is a process for creating new images that are sourced from existing images. The cropping process is done by taking the values from each side of the face bounding box that has been formed, as shown in Figure 6.
2. Resizing is completed by referring to features that have been detected using Viola Jones and will undergo a geometric transformation by changing the image to equalize the input dimensions on the system. The initial size of the face frame will be resized by referring to the original bounding box size from the feature detection results. The specifications of the input data are 130 x 110 pixels for the size of the facial features, 20 x 65 pixels for a pair of eyes features, 35 x 40 pixels for the nose features, while 30 x 50 pixels for the mouth features. The resizing process utilizes the nearest neighbor method.
3. Brightness Enhancement is a process used to improve the quality of lighting in an image by adding a contrast value to each RGB pixel. One of the brightness enhancement examples is shown in Figure 7(b). In the picture, there is a contrast change in the image.

4. Grayscaling is a process carried out to convert RGB (Red, Green, Blue) images into grayscale images. An example of grayscaling results is shown in Figure 7(c).
5. Histogram Equalization with the CLAHE method (Contrast Limited Adaptive Histogram Equalization) is performed for leveling histograms of facial images that have passed the grayscaling process. The CLAHE method works by limiting the predetermined contrast level (cliplimit) of the image to avoid excessive contrast.



(a) (b) (c)

Figure 7. (a) input image, (b) brightness enhancement output, (c) grayscaling output

CLAHE operates in a small area of the image called tile. The contrast contained in each tile will be fixed so that the histogram generated from that area matches the specified histogram shape. Adjacent tiles will be connected using bilinear interpolation. This method is done so that the results of the tile incorporation look smooth. The CLAHE parameter used is cliplimit parameter with a value of 0.005 and NumTile with size [2 2].

After the preprocessing stage, the feature extraction process is performed using the HOG (Histogram of Oriented Gradients) method with the stages shown in Figure 8. The function used for implementing the HOG method is extractHOGFeatures with the cellsize parameter of [2 2]. Figure 9 shows an example of an eye feature pattern with a size of 20x65 pixels.

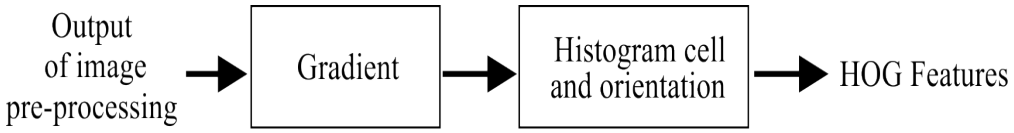


Figure 8. Stages on the HOG method

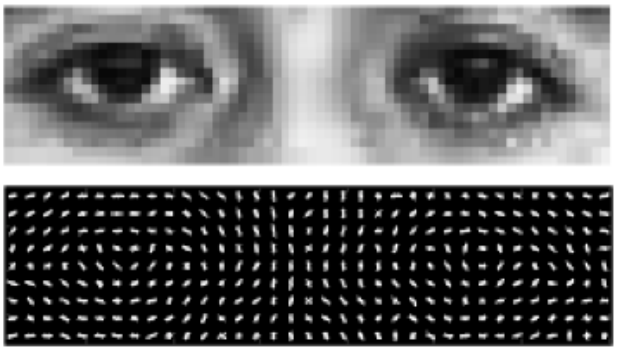


Figure 9. Example of a pair of eye patterns with a size of 20 x 65 pixels.

The last stage implemented in this study is the classification of features extracted using the MSVM (Multiclass Support Vector Machine) method which is divided into a classification in the training process and a classification in the testing process. The classification in the training process will classify 5 faces, where each face has 5 angles, then the class is made like "class = [1 1 1 1 1 2 2 2 2 2 3 3 3 3 3 4 4 4 4 4 5 5 5 5 5] ". Features are classified in the form of vectors [m x r], where m is the number of faces to be trained and r is the value of features that the face will have, such as the value of facial features, a pair of eyes, nose, and mouth.

After that, a train feature model will be created to match the train features with the available classes. The train feature model will generate a struct file with the size of [1x1]. The training model that has been made using predict function on the same features will be tested until they reach 100% accuracy.

In the testing process, faces in the form of video frames are classified according to the class provided. The new tested feature is matched with the trained model that has been created and the MSVM method will classify the value of the incoming features according to the value of the features in the trained model. The feature will be classified in the form of a vector [*n* x *p*], where n is the number of faces in the video frame that has full feature pieces (full face), while p is the value of the features that the face has, such as the value of facial, eye, nose and mouth.

3. Result and Discussion

The results of face recognition by changing the distance value (*d*) between the camera and walk through metal detector **shows** in Table 1 to Table 3 by using a confusion matrix. These results are obtained based on the classification of local facial features extracted using the HOG method. The classification results produce a final value called the score, where the score determines the outcome of the prediction by referring to the highest value.

Table 1. Confusion Matrix d = 200 cm

		Prediction					Accuracy (%)
		T1	T2	T3	T4	T5	
Actual	T1	**119**	27	2	2	0	79.33
	T2	10	**115**	24	0	0	77.18
	T3	0	10	**81**	0	6	83.51
	T4	18	6	35	**30**	24	26.55
	T5	0	0	20	0	**57**	74.03
Overall Accuracy (%)							**68.60**

Table 2. Confusion Matrix d = 300 cm

		Prediction					Accuracy (%)
		T1	T2	T3	T4	T5	
Actual	T1	**153**	8	4	0	3	90
	T2	15	**134**	1	0	0	89.33
	T3	0	0	**150**	0	0	100
	T4	12	0	16	**102**	20	68
	T5	0	2	16	0	**115**	86.47
Overall Accuracy (%)							**86.77**

Table 3. Confusion Matrix d = 400 cm

		Prediction					Accuracy (%)
		T1	T2	T3	T4	T5	
Actual	T1	**147**	0	3	0	0	98
	T2	6	**125**	19	0	0	83.33
	T3	1	15	**132**	0	0	89.19
	T4	19	11	20	**51**	9	46.36
	T5	0	10	41	0	**83**	61.94
Overall Accuracy (%)							**77.75**

In Table 1 to Table 3, it can be seen that the highest accuracy for the three scenarios is obtained on *T1* and the lowest accuracy on *T4*. The lowest level of accuracy on *T4* is caused by many *T4* targets detected with closed eye conditions. As explained earlier, the eye is one of the local features used. The overall system accuracy results for face identification of HOG features and the SVM Multiclass method at a distance of 200cm, 300cm, and 400cm are 68.60%, 86.77%, and 77.75% respectively.

4. Conclusion

Face recognition using the Multiclass SVM method has been carried out in this study with training data is in the form of images from 5 different angles, while the testing data is in the form of video data from CCTV with 5 suspects and 150 video frames per change in distance between camera position and walk through metal detector. The result shows that the distance of 300 cm gives the highest accuracy value of 86.76%. Face recognition errors caused by the condition of the face image at a distance of 2 meters do not display the overall facial features so that the value of the hog feature obtained is not recognized by the MSVM method. At a distance of 4 meters, the condition of the detected face image has a low resolution, and this has an impact on the condition that the face image becomes blurred so that it is not recognized by the MSVM method.

To increase the accuracy value of face recognition, then in the next study it is necessary to design a system that can extract the features in the condition that the image does not fully appear (with occlusion) and in the condition of blurred face images.

References

[1] K. Riset and D. A. N. P. Tinggi, "National Riset 2017-2045 (28 February 2017 Edition)," vol. 2045, 2017.
[2] A. Raghuwanshi, "An Automated Classroom Attendance System Using Video Based Face Recognition," pp. 719–724, 2017.
[3] F. Hajati, M. Tavakolian, S. Gheisari, Y. Gao, and A. S. Mian, "Sparse Representation : Application to Video-Based Face Recognition," pp. 1–13, 2017.
[4] M. Parchami, S. Bashbaghi, and E. Granger, "Video-based face recognition using ensemble of haar-like deep convolutional neural networks," *2017 Int. Jt. Conf. Neural Networks*, pp. 4625–4632, 2017.
[5] C. Ding and D. Tao, "Trunk-Branch Ensemble Convolutional Neural Networks for Video-Based Face Recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 1002–1014, 2018.
[6] J.-C. J. Jian, C.-H. Wu, C.-C. Lai, S.-T. Pan, S.-J. Lee, and C.-S. Ouyang, "Cascading global and local features for face recognition using support vector machines and local ternary patterns," *2017 IEEE Int. Conf. Inf. Autom. ICIA 2017*, no. July, pp. 404–409, 2017.
[7] Indrabayu, A. R. Tacok, and I. S. Areni, "Modification on Brightness Enhancement for Simple Thresholding in Eyelid Area Measurement," pp. 101–104, 2017.
[8] Z. Xiang, H. Tan, and W. Ye, "The Excellent Properties of a Dense Grid-Based HOG Feature on Face Recognition Compared to Gabor and LBP," *IEEE Access*, vol. 6, no. c, pp. 29306–29318, 2018.
[9] P. Viola and M. Jones, "Rapid Object Detection using a Boosted Cascade of Simple Features," 2001.