Implementation of Face Detection and Tracking on A Low Cost Embedded System Using Fusion Technique

Aryuanto Soetedjo*
Dept. of Electrical Engineering
National Institute of Technology
Malang, Indonesia
aryuanto@gmail.com

I Komang Somawirata
Dept. of Electrical Engineering
National Institute of Technology
Malang, Indonesia
kmgsomawirata@yahoo.com

***Abstract*—This paper presents the fusion techniques for detecting and tracking the face. The proposed method combines the Viola-Jones method, the CamShift tracking, and the Kalman Filter tracking. The objective is to increase the face detection rate, while reduce the computation cost. The proposed method is implemented on a low cost embedded system based-on the Raspberry Pi module. The experimental results show that the average detection rate of 98.3% is achieved, and it is superior compared to the existing techniques. The proposed system achieves the frame rate of 7.09 fps in the real-time face detection.**

***Index Terms*—Face detection, Viola-Jones, CamShift, Kalman Filter, Raspberry Pi.**

I. Introduction

Face detection and tracking is an important and popular research topic in the image processing area. An example of the real application that employs the technique is a system for detecting the driver fatigue using the camera systems [1]-[5]. In the system, the driver fatigue is examined from the facial features, such as the eye closure, eye blinking, and mouth openness. The face detection technique is a crucial task for localizing the face area for further process, especially for finding the eyes precisely [1]. Usually, the face tracking is performed after the face detection to improve the performance. By tracking the face, the search area on the next image frame is limited.

Due to the real-time requirement, the driver fatigue detection systems have been implemented using the embedded systems [2]-[5]. The low cost embedded systems using Raspberry Pi module was adopted for real implementation [4], [5].

Many face detection techniques have been proposed by the researchers, namely based on the skin color models [6]-[9] and Haar-like classifier [2]-[5],[8],[10],[11]. In the previous techniques, a face is detected by thresholding the image on a particular color space, such as the normalized RGB [6], the generalized LHS [7], and YCbCr [8],[9]. The latter methods employ the Adaboost learning to detect the face from the Haar-like features called as the Viola-Jones method [12].

In general, the face detection methods based on the Viola-Jones methods offer the better detection rate compared to the skin color model techniques [6[. The computation cost of the Viola-Jones method is relative low. Therefore, this method becomes the most popular method for face detection. However, the method fails to detect the face when the face is occluded by the other objects. To overcome the limitations, the modified version or fusion techniques are proposed by the researchers.

The computation time of face detection from video images could be improved by introducing the tracking technique. By tracking the face in every frame, the search area to find the face in the next frame is localized in the limited area only. Thus it will reduce the computation time.

The CamShift tracking was employed to track the face on the video image once the face was detected [4]. In the system, tracking was used to find the center of face image for judging three conditions, i.e. driver alert condition, drowsiness condition, and out of box condition. The face detection and tracking was implemented on a single board computer equipped with a camera.

The CamShift tracking is a simple and efficient tracking method [13]. It was suitable to track the face for the driver fatigue detection system due to the several reasons [10]: (1) only hands and face are the biggest objects on the captured image; (2) the background of the image is almost stationary. To increase the performance in the varying lighting environment, they proposed to limit the search window of the CamShift method.

Other methods to improve the performance of Camshift tracking are by combining with the Kalman Filter tracking [9], [14]. The Kalman Filter tracks an object by considering the velocity and position of the moving object. It is used to predict the next position of object. The predicted location is then analyzed by the CamShift method for finding the face.

In this paper, we propose a new fusion technique based on the Viola-Jones, the CamShift, and the Kalman Filter techniques for face detection. The main contribution of our proposed method is two folds, i.e.: (1) instead of combine the Viola-Jones and the CamShift techniques in cascade arrangement, we combine them as complimentary or parallel arrangement; (2) the decision output consists of two bounding boxes, namely the detection bounding box and the prediction bounding box. The second contribution could be achieved by the assumption that the proposed face detection is the earlier stage to extract the further features of the face, such as eyes or

978-1-5090-2218-2/16/$31.00 ©2016 IEEE

mouth. Our method provides two bounding boxes representing the face that could be further validated by the next stage to find the eyes or mouth.

The proposed technique is simple and fast. Therefore it could be implemented on a low cost Raspberry Pi module equipped with 5 Mega Pixels Raspberry Pi camera to achieve the real-time implementation.

The rest of the paper is organized as follows. Section 2 presents our proposed approach. Section 3 discusses the experimental results. Conclusion is covered in Section 4.

II. Proposed Approach

A. System Overview

The proposed face detection and tracking is illustrated in Fig. 1. The method is divided into two stages: detection stage and tracking stage. In the detection stage, the Viola-Jones face detection technique is combined in parallel with the CamShift tracking. The objective of combining them in parallel fashion is described in the following.

The Viola-Jones method could detect the face effectively. However, it could not detect the occluded face. Further, the detection is affected by the type of classifier which is used during detection. For instance, if the frontal face classifier is used, it could not detect the profile face and vice versa. One solution is by adopting both the frontal face and profile face classifiers. But it consumes the computation time.

The problem of occlusions, in a certain degree, could be solved by employed the CamShift tracking. When a small object occludes the face, the Viola-Jones may fail to detect the face. But the CamShift method could detect the face properly. In some cases, when the skin colored objects distract the face, for instance when a man holds the ears using his hands, then the CamShift method will detect the face wrongly. Fortunately the Viola-Jones could detect the face properly.

Related to the problem of the frontal and profile faces, since both types of the face appear in the same color, they could be detected properly using the CamShift method.

From the above explanation, it is clear that both the Viola-Jones and CamShift techniques are complement each other. To exploit both advantages, it is suggested to combine them in parallel fashion as illustrated in Fig. 1.

The output of detection stage is the detected face region obtained by the Viola Jones or the CamShift techniques. Then the detected region is used by the Kalman Filter tracking to predict the location of the face. The predicted region is assigned as the new search window for searching the face in the next frame.

As stated previously, the proposed approach generates two bounding boxes (gray boxes in Fig. 1) to provide the better face detection as described in the following. When the face is detected by the Viola-Jones method, the detected face region is assigned as the bounding box of detected face. Otherwise, the bounding box is defined from the CamShift tracking. This approach works properly when the conditions as described previously are satisfied.

In some cases, the bounding box defines the wrong face region as explained in the following. Let us assume that in the k^{th} frame, the face is detected by the Viola Jones method. However the Viola-Jones fails to detect the face in the $(k+1)^{th}$ frame. Therefore in the $(k+1)^{th}$ frame, the detected face region is defined by the CamShift method. It could be a wrong position due to the occlusion or lighting changes. Thus the bounding box does not localize the face properly. Fortunately, since the Kalman Filter tracking considers the position and velocity of the face in the previous frame, the predicted region in the $(k+1)^{th}$ frame is still closely to the detected region in kth frame. It leads to propose the approach that generates both bounding boxes, i.e the bounding box defined by the Viola-Jones or the Camshift methods, and the bounding box which is predicted by the Kalman Filter tracking.

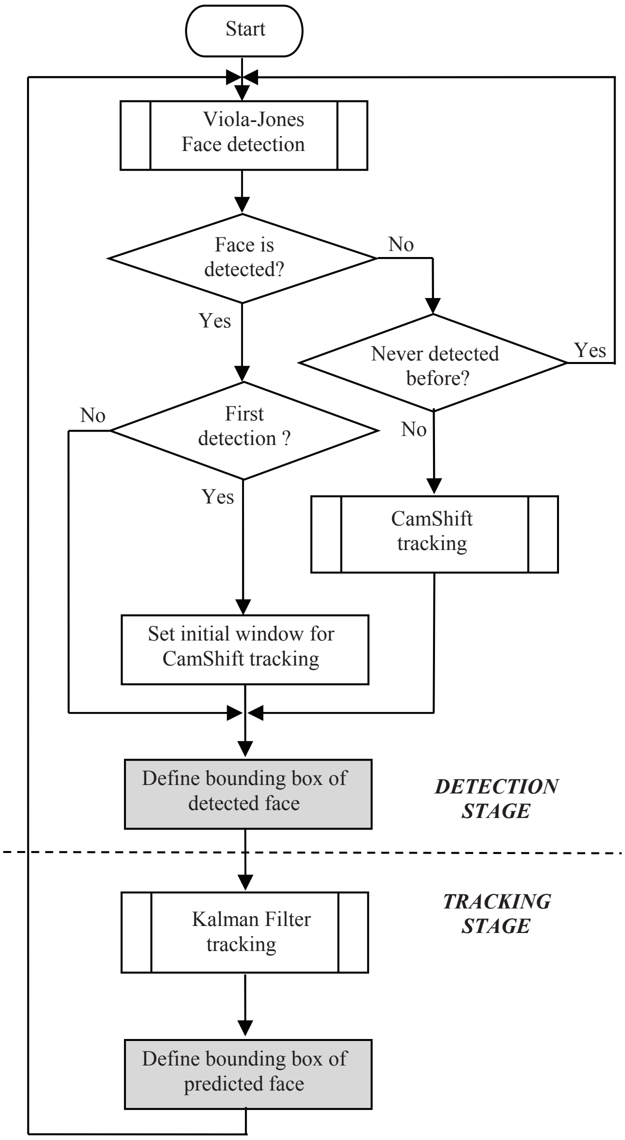

Fig. 1. Proposed face detection and tracking.

B. Face Detection

In the proposed approach, the Viola-Jones method is considered as the primary detection technique, in the sense that in every frame the face is searched by the Viola-Jones first. When it is failed then the CamShift method will take over. The detected face region obtained by the Viola-Jones in the first time is used by the CamShift method as the target tracking in the successive frames.

Since the Viola-Jones method works on a grayscale image, it is required to convert the captured RGB image to the grayscale image. The CamShift tracking starts to run if the Viola-Jones does not detect the face and the face is already detected for the first time. The RGB image is converted to HSV image before it is processed by the CamShift tracking.

C. Face Tracking

Once the bounding box of face is defined, the face is then tracked by the Kalman Filter tracking. The Kalman Filter tracking is used to predict the face in the next frame. This prediction defines the new search window used by the Viola-Jones method to find the face. By limiting the search window, the computation time is reduced.

The parameters of the Kalman Filter tracking consist of the state vector and the measurement vector. In this work, the state vector consists of the *x*-coordinate of the center position of the face, the *y*-coordinate of the center position of the face region, the velocity in the *x*-direction, the velocity in the y-direction, the width of face region, and the height of face region. While the measurement vector consists of the *x*-coordinate of the center position of the face, the *y*-coordinate of the center position of the face region, the width of face region, and the height of face region.

The Kalman Filter tracking is divided into two processes: the time update (prediction) and the measurement update (correction). The measurement update uses the observation data from the bounding box of detected face obtained by the detection stage.

III. Experimental Results

The proposed system is implemented on a low cost embedded system based on the Raspberry Pi 3 Model B with 1.2 GHz 64-bit quad-core ARMv8 CPU and 1 GB RAM. To capture the video image, the Raspberry Pi camera module is employed. The camera module uses the image sensor Omnivision 5647 and supports image resolution of 2592 x 1944 pixels. In the experiment, the image resolution is set to 320 x 240 pixels to speed up the execution time.

The Raspberry Pi runs under the Raspbian operating system. The proposed algorithm is implemented using C++ language and OpenCV library. To evaluate our proposed algorithm, four methods are compared. The first method is the Viola-Jones only method (**VJ**). The second method is the Viola-Jones method and the CamShift tracking (**VJ-CS**), in which the Viola Jones method is used to detect the face for the first time only. Once the face is detected, the rest detection is performed by the CamShift tracking. The third method is the Viola-Jones method and the Kalman Filter tracking (**VJ-KF**), in which the Kalman Filter is used to predict the face region, while the detected face region obtained by the Viola-Jones method is used to update the Kalman Filter. The fourth method is our proposed method (**PM).**

In the experiments, four methods are tested using the same hardware, i.e. the Raspberry Pi and camera module. Since only one camera module is employed, it is difficult to prepare the tested object (human) that is able to repeat the same movement for every method under testing. Instead, the recorded video is employed, in which it is played back to evaluate every method. Therefore the camera module is placed in front of the computer's monitor that playing the tested video. This arrangement ensures that every method captures the same video images. The tested video images are taken from NRC-IIT Facial Video Database [15]. Four video sets are used for evaluating the methods. The image samples of the four video sets are shown in Fig. 2.

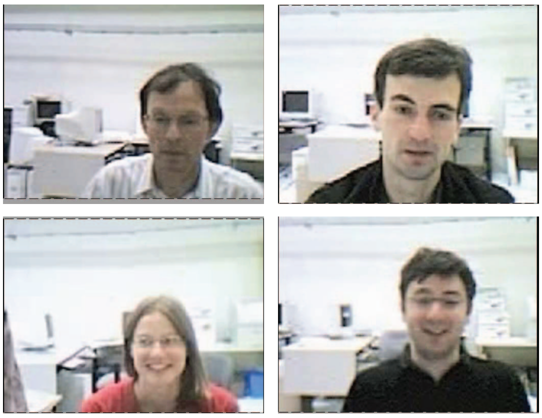

Fig. 2. Image samples of four video sets used in the experiments.

For comparing the four methods, two parameters are evaluated, i.e. the true detection rate and the frame rate. The true detection rate is defined as the number of successful detected face images divided by the total number of images. It is noted here that for PM, the detected face image is considered as successful detection if one or both bounding boxes generated by the algorithm are the face images. The frame rate is the number of captured images in one second, and expressed as frame per second (*fps*).

The experimental results of the detection rates and the frame rates are listed in Table 1 and Table 2 respectively. From Table 1, it is obtained that the highest true face detection rate of 98.3% is achieved by our proposed method. While the lowest face detection rate of 67.1% is achieved by the Viola-Jones method and the CamShift tracking. It is clearly shown from the results that our strategy to combine three methods and to generate two bounding boxes works effectively for detecting the face.

The frame rate of our proposed method is 7.09 fps. It is lower compared to the CamShift method. However, by considering both the detection rate and the frame rate, our proposed method is superior compared to the others.

TABLE I. RESULT OF TRUE DETECTION RATES

Video No.	Methods			
	VJ	*VJ-CS*	*VJ-KF*	*PM*
1	76.1%	82.4%	79.8%	99.3%
2	100%	81.0%	92.3%	100%
3	87.3%	21.6%	89.0%	93.8%
4	95.8%	83.3%	95.7%	100%
Average	**89.8%**	**67.1%**	**89.2%**	**98.3%**

TABLE II. RESULT OF FRAME RATES

Video No.	Methods			
	VJ	*VJ-CS*	*VJ-KF*	*PM*
1	3.05 fps	14.47 fps	5.56 fps	6.77 fps
2	3.00 fps	13.70 fps	6.55 fps	7.17 fps
3	2.95 fps	14.54 fps	6.87 fps	6.70 fps
4	3.01 fps	14.30 fps	7.24 fps	7.72 fps
Average	**3.00 fps**	**14.25 fps**	**6.56 fps**	**7.09 fps**

Some image sequences of the detection results are shown in Fig. 3. The detection results of **VJ**, **VJ-CS**, **VJ-KF**, and **PM** are shown in Fig. 3(a), Fig. 3(b), Fig. 3(c), and Fig. 3(d) respectively. The image sequences are taken from video set-1 during the time where the man moves his hands to his face then opens the hands over the face. Since the frame rate of every method is different, the image sequences and the frame numbers are not always the same.

In the figures, the green rectangle represents the bounding box of detected face obtained by the Viola-Jones method. The red rectangle represents the bounding box of predicted face obtained by the Kalman Filter tracking. While the red ellipse represents the detected face obtained by the CamShift method.

From Fig. 3(a), the face is detected in 15th frame. However in 17th frame, the face is occluded by the hand, and the face could not be detected. In 19th frame when the hands move out from the face, the face is detected properly. It is clear that the Viola-Jones method fails to detect the face under occlusions.

From Fig. 3(b), the CamShift method could detect the face in 76th frame. Since the CamShift method works by tracking the color of the face, the neck whose color is similar to the face will be detected as shown in the figure. In 79th frame, the detected face region includes the hands covering the face. In this case, the detected region is still closely to the face region. Thus it is considered as the true face detection. However in 83th frame, the detected region becomes very large following the hand position. It is considered as the false face detection.

From Fig. 3(c), the bounding boxes of predicted face in 31st, 33th, and 34th frames are located properly in the face region. By observing the images, it is clearly shown that the Kalman Filter tracking uses the position and velocity of the face for prediction. In the figures, the man moves his hands but does not move his face. Thus the predicted face is not disturbed by the movement of the hands.

From Fig. 3(d), in 34th frame, both detected and predicted faces are located closely. In 35th frame, since the hands occlude the face, the Viola-Jones method fails to detect the face. Thus the face detection is performed by the CamShift method that yields a larger face region due to the appearing of the hands. Fortunately, the predicted face obtained by the Kalman filter tracking is located on the face region closely. In 38th frame, the predicted face is located wrongly. This wrong tracking is caused by the movement of hands that are now considered in the Kalman Filter tracking due to the fact that in the previous frame (35th frame), the tracked object changes from the face region detected by the Viola-Jones method to the face region detected by the CamShift method. Fortunately, the Viola-Jones method could detect the face properly in 38th frame. Thus according to our strategy, all three images are considered as the true face detection.

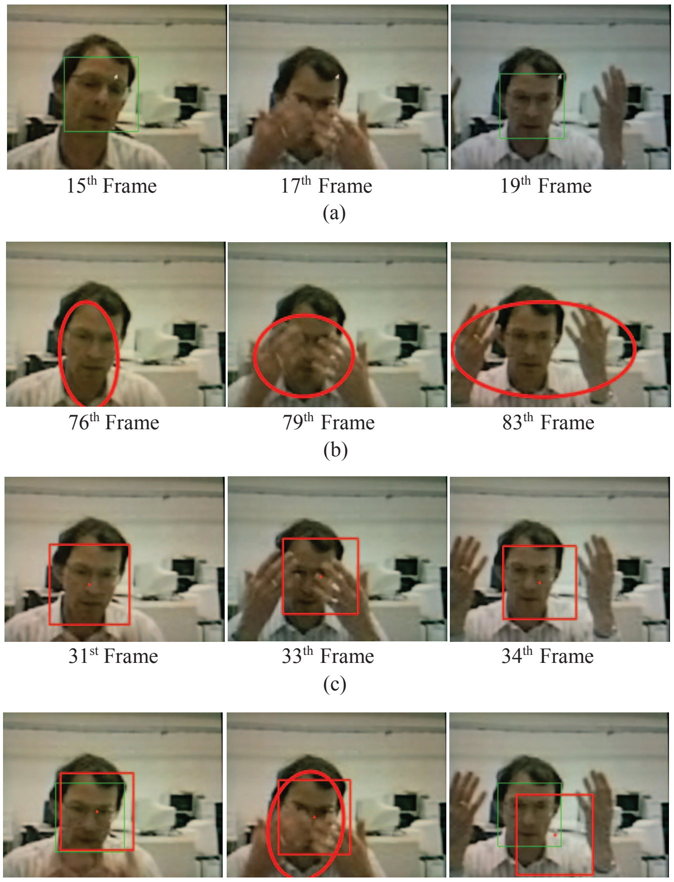
15th Frame 17th Frame 19th Frame
(a)
76th Frame 79th Frame 83th Frame
(b)
31st Frame 33th Frame 34th Frame
(c)
34th Frame 35th Frame 38th Frame
(d)

Fig. 3. Some image sequences obtained during the experiments.

IV. CONCLUSION

The real-time face detection and tracking system is proposed. To achieve the high detection rate, the fusion technique is employed. The method is evaluated and compared to the existing techniques, namely the Viola-Jones method, the combination of the Viola-Jones method and the CamShift method, and the combination of the Viola-Jones method and the Kalman Filter tracking. The detection rate of the proposed method is the best, while the frame rate is the second best.

In future, the method will be extended to improve the execution time. Further, the fusion techniques could be explored more to increase the detection efficiency.

ACKNOWLEDGMENT

This work is supported by the Research Grant 2016, Competence-Based Research scheme from Directorate General of Higher Education, Ministry of Research and Technology and Higher Education, Republic of Indonesia (SP/ DIPA. 023.04.1.673.453/2016).

REFERENCES

[1] M.H. Sigari, M.R. Pourshahabi, M. Soryani, and M. Fathy, “A Review on Driver Face Monitoring Systems for Fatigue and Distraction Detection,” International Journal of Advanced Science and Technology, Vol. 64, pp. 73-100, 2014.

[2] H,M., and M.H. Xu, “Design and Implementation of Embedded Driver Fatigue Monitor System,” Proceedings of International Conference on Artificial Intelligence and Industrial Engineering, Phuket, Thailand, 2015.

[3] I. Garcia, S. Bronte, L.M. Bergasa, J. Almazan, and J. Yebes, “Vision-based drowsiness detector for Real Driving Conditions,” Proceedings of Intelligent Vehicles Symposium, Alcala de Henares, Spain, 2012.

[4] V.E. Dahiphale, and Sathyanarayana, “Real-Time Computer Vision System for Continuous Face Detection and Tracking,” International Journal of Computer Applications, Vol. 122, No. 18, pp. 1-5, 2015.

[5] D. Sarkar, and A. Chowdhury, “A Real Time Embedded System Application for Driver Drowsiness and Alcoholic Intoxication,” International Journal of Engineering Trends and Technology, Vol. 10, No. 8, 2014.

[6] A. Soetedjo, and K. Yamada, “Skin Color Segmentation Using Coarse-to-Fine Region on Normalized RGB Chromaticity Diagram for Face Detection,” IEICE Transactions on Information and Systems, Vol. E91-D, No. 10, pp. 2493-2502, 2008.

[7] Z.W. Zhang, M.H. Wang, Z.H. Lu, and Y. Zhang, “A Skin Color Model Based on Modified GLHS Space for Face Detection,” Journal of Information Hiding and Multimedia Signal Processing, Vol. 5, No. 2, pp. 144-151, 2014.

[8] L. Man, W. Xiao-Yu, and M. Hui-ling, “Face Automatic Detection based on Elliptic Skin Model and Improved Adaboost Algorithm,” International Journal of Signal Processing, Image Processing and Pattern Recognition, Vol. 8, No. 2, pp. 227-234, 2015.

[9] S.V. Tathe, and S.P. Narote, “Mean Shift and Kalman Filter based Human Face Tracking,” Proceedings of International Conference on Advances in Signal Processing and Communication, Lucknow, India, 2013.

[10] K.W.B. Ghazali, J. Ma, and R. Xiao, “Driver’s Face Tracking Based on Improved CAMShift,” International Journal of Image, Graphics and Signal Processing, Vol. 5, No. 1, pp.1-7, 2013.

[11] J. Foytik, P. Sankaran, and V. Asari, “Tracking and Recognizing Multiple Faces Using Kalman Filter and Modular PCA,” Procedia Computer Science, Vol. 6, pp. 256-261, 2011.

[12] P. Viola, and M.J. Jones, “Robust Real-Time Face Detection,” International Journal of Computer Vision, Vol. 57, No. 2, pp. 137-154, 2004.

[13] G. Bradski, “Computer vision face tracking for use in a perceptual user interface,” Intel Technology Journal, Vol. 2, 1998.

[14] X. Chen, H. Wu, X. Li, X. Luo, and T. Qiu, “Real-time Visual Object Tracking via CamShift-Based Robust Framework,” International Journal of Fuzzy Systems, Vol. 14, No. 2, pp. 262-269, 2012.

[15] D.O. Gorodnichy, “Video-based framework for face recognition in video,” Proceedings of Second Canadian Conference on Computer and Robot Vision (CRV'05), pp. 330-338, Victoria, BC, Canada, 2005.