Using Margin-based Region of Interest Technique with Multi-Task Convolutional Neural Network and Template Matching for Robust Face Detection and Tracking System

Bacha Rehman, Ong Wee Hong, Abby Tan Chee Hong
Faculty of Science
Universiti Brunei Darussalam
Brunei Darussalam
e-mail: bachapk@gmail.com, weehong.ong@ubd.edu.bn, abby.tan@ubd.edu.bn

***Abstract*—Real-time face detection and tracking systems suffer from low accuracy and slow processing speed that lead to poor robustness. This problem is vital in real-time setups including human robot interactions (HRI) and video analysis systems. This paper presents margin-based region of interest (MROI) approach to speed up the processing time. Further a hybrid approach is also presented that combines Multi-task Convolutional Neural Networks (MTCNN) with template matching to improve face detection accuracy. The MROI approach which is responsible to speed up the processing time is presented in two variants with fixed and dynamic margin concepts. Dataset used in this work comprises of twenty RGB video files. Each video file is fifteen seconds long and been created from real-life videos containing a person in lecture delivering environment. Each video file contains a person in which the person moves, turns head and the camera also moves. The highest face detection and tracking accuracy achieved in this paper is 99.31% with a processing time of 14.93 milliseconds per frame. The proposed hybrid algorithm significantly improves the ability to detect and track faces in sideway orientation apart from frontal face. The proposed algorithm has the ability to process above 65 frames per second (FPS). The system presented has increased FPS processing ability by more than 400% as well as given boost to the accuracy if compared to the conventional MTCNN full frame scanning techniques.**

Keywords-face detection; face tracking; MTCNN; template matching; convolution neural network; region of interest; MROI

I. INTRODUCTION

Face detection in image processing and computer vision is considered to be an important field of research [1], [2]. It is also the primary step for any face recognition system. Many researches have been carried out on automatic face detection [1]. The development in face detector has progressed from simple features [3] to the use of complex features [4] with different approaches to deal with the challenges in face detection such as complex background, variation in illumination, occlusions as well as variation in the orientation of the face. Among the works we have surveyed, Multi-task Convolutional Neural Network (MTCNN) [4] has shown state-of-the-art performance. The complex features used in MTCNN has achieved high accuracy, however, at the expense of extra computation.

This paper presents a hybrid face detector and tracker based on the MTCNN and template matching with margin-based region of interest (MROI) to improve both processing speed and face detection/tracking accuracy. It is observed that MTCNN [4] algorithm is sometimes not successful to perform fast and fails to deal with non-frontal face orientation in a real-time or video stream environment. To address this problem, we have used the template matching (TM) algorithm [5] with the MTCNN in a hybrid system. Template matching can find the resemblance between the input images and the template images. In the hybrid system, the MTCNN algorithm is the main detector routine and when it fails, the system switches to the template matching as the escape routine. Template matching as the escape routine is responsible to determine the most prospective face position based on the face detected in the previous frame. The main advantage of TM method is that any template image regardless if it contains frontal or non-frontal face orientation can be used in it. This hybrid approach improves the proposed face detector and tracker to be robust and reliable for real-time surveillance, Human Robot Interactions (HRI) and video analysis systems.

The concepts of fixed and dynamic margin-based region of interest (MROI) are used to improve the processing speed of the hybrid system. Face tracking provides the required information for the margin-based algorithms in order to process subsequent frames. If the face is found in the first place, its position is stored for the computation of region of interest (ROI) for the subsequent frame in the video or real-time data to avoid having to scan the full frame. This work presents two variations for the MROI calculations i.e. fixed margin (FM) and dynamic margin (DM). Fixed percentage x extra pixels added around the face area for the face that was detected in initial frame in the FM approach. On the other hand, DM is calculated with a fixed percentage of pixels and with added pixels corresponding to the rate of change in the face position from the detected position in the previous frames. The observations presented in this paper suggest significant improvement in processing speeds in the face detection process.

A total of six algorithms have been developed and implemented in this work and tested on the dataset FDTV-20 [6], which consists of 20 videos of 15 seconds each. In FDTV-20, all the videos contain a person who moves around

and sometimes turning head in a lecture delivering situation. The main contributions of this paper are:

1. Proposed hybrid face detection and tracking approach incorporating MROI and TM with MTCNN to achieve higher accuracy and faster speed that could not be achieved by using a single algorithm and with full frame scanning for face detection in real-time.

2. Proposed DM based MROI which takes the face movement into consideration while calculating ROI for the subsequent frame.

3. Used TM as an escape routine to help detect face that the main routine failed to detect.

4. Implemented six variations of the proposed hybrid approach and evaluated their performance on real-life videos.

II. Related Work

Numerous face detectors [1] have been proposed and applied from time to time. The most popular and seminal face detection algorithm was presented by Viola *et al*. [7], [3], which has become the most popular detector in many software tools e.g. OpenCV [8]. However due to the simple nature of the features used in this algorithm, it faces noticeable issues with processing speed and detecting non-frontal faces in images [9]. To some extent these issues have been addressed [9] by presenting variation in simple features. This effort [9] improved the non-frontal faces detection to some extent. Adding more complex features encouraged researchers to look into more complex and sophisticated techniques like convolutional neural network (CNN) [10]. However, there were still issues regarding processing speed and to eliminate the handcraft features which limits its performances. In this regard deep CNN has been used [11] for facial feature detection to achieve high response in regions of face which further yield candidate windows of faces. But due to the complex structure of CNN, the approach was observed to be time costly for real-time systems. Faster R-CNN has been proposed [12], to improve both detection accuracy and processing time that should enable it to be used for real time system. However, the single task approach makes it less effective [4]. For this purpose multi-task approach is proposed [4] for face detection using CNN.

Template matching (TM) algorithm for face localization has been presented in [13], while a few variations of template matching techniques are presented in [14]. As template matching algorithm is fast by its nature [5], it can be used together with the Multi-task Convolutional Neural Networks (MTCNN) in an intelligent way for effective face detection and tracking task.

Considering different tracking algorithms, a centered correlation filter based tracking systems have been discussed in [15]. Haar wavelet and edge orientation based feature have been used in region of interest (ROI) grouping and classification for the purpose of pedestrian detection were discussed in [16]. Template matching based approach addressed issues regarding shape, color and motion was discussed in [17]. The works mentioned above helped understand the various concepts to develop the hybrid system proposed in this paper with improved accuracy and processing time.

The work presented in this paper is an alternative approach to the face detector presented in [18] with improvement in the performance. In [18], the proposed system combined the seminal Haar cascade detector with template matching and MROI. In this paper, we adapt the proposed hybrid system with the state of the art MTCNN to show the effectiveness of the hybrid approach and MROI to improve the performance of a state of the art face detector.

III. Proposed System

The proposed algorithms developed follow a hybrid approach using Multi-task Convolutional Neural Networks (MTCNN) and template matching (TM). MTCNN is used as the main routine for face detection. However, the system switches to the TM i.e. escape routine, if the main routine fails. In addition, the fixed margin (FM) or dynamic margin (DM) based region of interest (MROI) is applied to achieve fast face detection and tracking. The system switches back from the escape routine to the main routine after a predetermined number of frames. The system switches to full frame processing when there is not enough margin around the region of interest.

A Convolutional Neural Network (CNN) is a multi-layered neural network in which all layers are not fully connected [10]. The input layer is responsible to receive the normalized and identical size images. In the input layer, the convolution kernel processes a set of units in a small neighborhood to form a single unit in the feature map of the convolutional layer.

As the main routine task involves face detection and alignment, therefore, a multi-task convolutional neural network is used. The MTCNN main routine [4] uses a three-stage cascaded CNN framework. The stage one consists of a full CNN, called Proposal Network (P-Net) that provides the initial prediction of the candidate face windows. The P-Net uses a network size of 12. The stage two is called Refine Network (R-Net) and is responsible for further rejection of large number of false candidates. R-Net has a network size of 24. Stage three is called Output Network (O-Net) and is responsible to produce final face bounding box and facial landmarks position. O-Net uses a network size of 48.

We can express the application of MTCNN on video, i.e. sequence of frames, as in (1). The summation symbol in (1) and the other equations in this paper represents the iteration.

$$F_n(z) = \sum_{m=\frac{z}{10}}^{m=\frac{z}{2}} MTCNN(m) \quad (1)$$

F_n represents the application of MTCNN [4] main routine on a given frame z. *n* denotes the normal face detection and tracking method. m is the scaled window starting at a minimum size of $\frac{z}{10}$ and the maximum size is half of the image frame, i.e. $\frac{z}{2}$.

Template matching (TM) can detect a given template image in the input RBG frame on the basis of best match

procedure using sliding. Normally the predefined template and source image are assumed to be given in the RGB format. As shown in Figure 1, the full frame source image is defined by a matrix of $X \times Y$. The template image is defined by the matrix of $x \times y$. In the main full frame searching area, the template can be matched at matrix (a, b) area. The template matching process takes the template image and search in the provided image frame to decide whether a face exists in the image using sliding procedure. The template matching procedure can be described as locating the best location (a, b) for a match with the template image template $x \times y$ so that the match between template matrix $(1 : y, 1 : x)$ and test matrix $(a : (a + y - 1), b : (b + x - 1))$ is maximized.

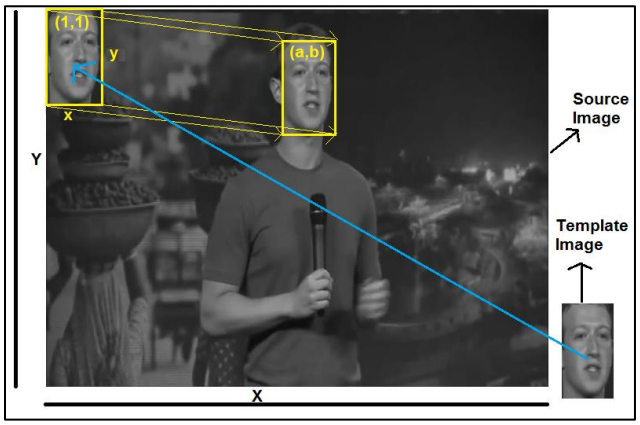

Figure 1. Template matching schema diagram.

Equation (2) represents the template matching algorithm using normalized sum of squared difference.

$$TM(x,y) = \left(\frac{\sum_{x',y'} \left(T(x',y') - I(x+x',y+y') \right)^2}{\sqrt{\sum_{x',y'} T(x',y')^2 \cdot \sum_{x',y'} I(x+x',y+y')^2}} \right) \quad (2)$$

$TM(x,y)$ is the equation for the template matching algorithm. I denotes the input image, T is the template, and TM is the result matrix. Each location (x,y) in TM contains the match metric. (x', y') represents the location within the template.

The proposed fixed and dynamic margin-based detectors are expressed mathematically in (3) and (5) respectively.

$$F_{fm}(z) = \sum_{m=r*\frac{1}{3}}^{m=r*\frac{6}{5}} MTCNN(m) \quad (3)$$

where

$$r = (1 + \Delta b)b \quad (4)$$

F_{fm} is face detection method with fixed MROI approach. In (3), r represents the region of interest area which is determined from the face bounding box extracted from the frame z. m is the scaled window starting at a minimum size of $\frac{6r}{5}$ and the maximum size of $\frac{r}{3}$. In fixed MROI, extra pixels around the face area, b, are taken at fixed percentage of Δb, set at 25%, on each side. Equation (4) gives the calculation for the region of interest in fixed margin approach. In the implementation in this paper, Δb is set at 25% for fixed margin approach.

$$F_{dm}(z) = \sum_{m=r*\frac{1}{3}}^{m=r*\frac{6}{5}} MTCNN(m) \quad (5)$$

where

$$r = (1 + \Delta b)b + \Delta p \quad (6)$$

F_{dm} is the equation for dynamic margin approach. It is similar to F_{fm} except that the region of interest r is determined with dynamic extra pixels, Δp, proportional to the movement of the face position in the previous two frames. Equation (6) gives the calculation for the region of interest in dynamic margin approach. In the implementations in this paper, Δb is set at 20% here. Figure 2 illustrate the MROI algorithm used for this work.

MROI algorithm

FOR each frame
1. Obtain the MROI (if any) from the previous frame
2. Search for the face in the given MROI with the selected algorithm
3. IF face detected within MROI
   - 3.1. Get face box
   - 3.2. If fixed margin then add Δb around the face box as MROI
   - 3.3. Else If dynamic margin then add Δb and Δp around the face as MROI
   - 3.4. Else take the whole frame as MROI
   - 3.5. End If

   END IF

END FOR

Figure 2. MROI Algorithm.

Based on the mathematical models presented in (1) to (6), six algorithms have been derived and implemented. These algorithms are expressed mathematically in (7) to (12). Equations (8), (9), (11), and (12) use the MROI approaches presented in this work. In which (8) and (11) presents the fixed margin while (9) and (12) present dynamic margin approaches. The algorithms expressed from (7) to (12) are named as below:

1. Normal Face Tracking (NT)
2. Fixed Margin Face Tracking (FMT)
3. Dynamic Margin Face Tracking (DMT)
4. Normal Template Matching Face Tracking (NTMT)
5. Fixed Margin with Template Matching Face Tracking (FMTMT)
6. Dynamic Margin with Template Matching Face Tracking (DMTMT)

$$F_{NT}(z) = F_n(z) \quad (7)$$

$$F_{FMT}(z) = \begin{cases} F_n(z) & z = 1 \parallel F_n(z-1) = 0 \parallel F_{fm}(z-1) = 0 \\ F_{fm}(z) & otherwise \end{cases} \quad (8)$$

$$F_{DMT}(z) = \begin{cases} F_n(z) & z = 1 \parallel F_n(z-1) = 0 \parallel F_{dm}(z-1) = 0 \\ F_{dm}(z) & otherwise \end{cases} \quad (9)$$

$$F_{NTMT}(z) = \begin{cases} F_n(z) & z = 1 \parallel T = null \parallel cnt = 10 \\ TM(x,y)_{z,cnt} & F_n(z) = 0 \end{cases} \quad (10)$$

$$F_{FMTMT}(z) = \begin{cases} F_n(z) & z = 1 \parallel T = null \parallel cnt = 10 \\ F_{fm}(z) & F_n(z-1) = 1 \parallel F_{fm}(z-1) = 1 \\ TM(x,y)_{z,cnt} & F_n(z) = 0 \parallel F_{fm}(z) = 0 \end{cases} \quad (11)$$

$$F_{DMTMT}(z) = \begin{cases} F_n(z) & z = 1 \parallel T = null \parallel cnt = 10 \\ F_{dm}(z) & F_n(z-1) = 1 \parallel F_{dm}(z-1) = 1 \\ TM(x,y)_{z,cnt} & F_n(z) = 0 \parallel F_{dm}(z) = 0 \end{cases} \quad (12)$$

For algorithms involving template matching, T is the template image and cnt=10 means that whenever the routine is switched to the escape routine i.e. TM, it will process the next 10 frames before switching back to the main routine i.e. MTCNN.

In addition, a distance variable is used to measure the distance between the previous and current frame face position. If the distance exceeds a certain threshold i.e. 30 pixels in this case, it is considered as a wrong detection and the face tracking switches from MTCNN to TM for continuation of face detection and tracking. MTCNN detector sometimes has false detection; therefore, such situations are minimized by introducing the distance filter with an escape routine.

IV. Experimental Setup

A. Dataset

A dataset comprised of 20 videos been created for this work and is made available for general public. The videos were obtained from YouTube and resized to a resolution of 640x480. Each video is 15 seconds each which roughly having 450 frames and contains single face in a lecture delivery environment. There are 13 videos of male and 7 videos of female.

B. Experiment

All the algorithms were tested on the dataset FDTV-20 [6] containing 20 video files to evaluate the performance of each algorithm in terms of accuracy (correct, incorrect, and not detected), average time taken per frame in milliseconds, and the ability of the whole system to process the number of frames per second. Each algorithm was executed ten times on each video file. From the results obtained, the accuracy, execution time and FPS performance are calculated by taking the average of all obtained results.

C. Development Enviroment

Hardware used for all experiments used is Intel® Core™ i5 CPU 650 @ 3.20 GHz with 8 GB RAM. The software tools used including visual C++, OPENCV and Caffe on windows operating system.

V. Results

The accuracy threshold is taken as 10 pixels. The face position is taken as the center of the face rectangle. If the distance between the face position detected and ground truth position is less than threshold value, then it is considered as correct detection. If the above distance is greater than the threshold then it is incorrect detection. If a face is not detected, then it is considered as Not Detected.

Figure 3 shows the average face tracking accuracy for each of the algorithms. It can be seen from the results that the incorporation of template matching has significantly improved the accuracy of the MTCNN based tracking. The result shows that the dynamic margin is more robust than the fixed margin in detecting faces.

In Figure 3, the average accuracy of the algorithm DMTMT is the highest i.e. 99.31%. In this particular algorithm the concept of the dynamic margin is used. For both the MTCNN classifier and TM, it scans the dynamic MROI from the face detection once the face is detected within the frame. The conventional NT algorithm has the average accuracy of 97.64%. The FMTMT and DMTMT algorithms successfully achieved an accuracy of 99.28% and 99.31%.

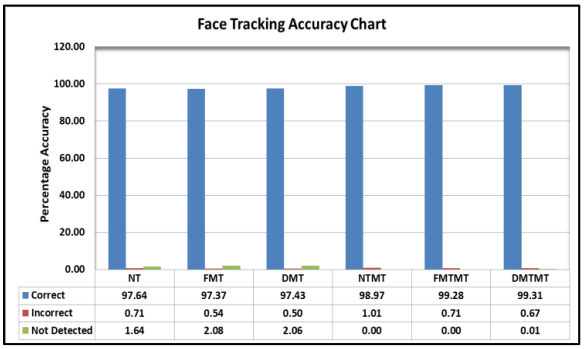

Figure 3. Average face tracking accuracy.

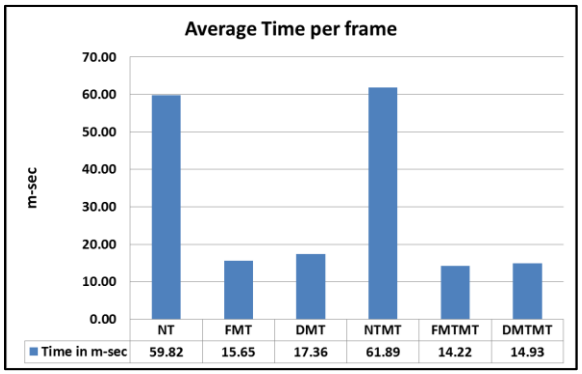

Figure 4. Average time per frame.

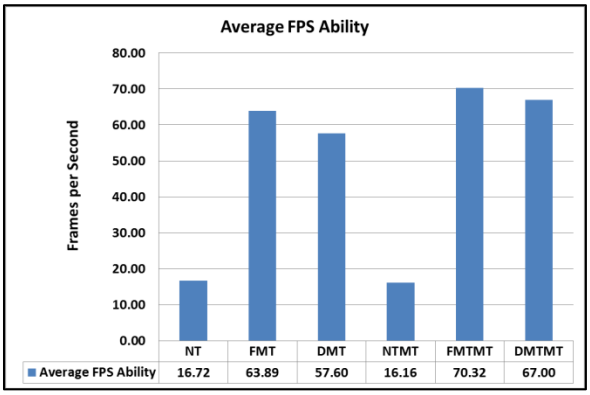

Figure 5. Average frames per second (FPS).

The main research problem while taking up this task is to set a tradeoff between accuracy and processing time, therefore, combined results with accuracy and processing time need to be achieved. Figure 4 shows the processing time details for each algorithm in milliseconds.

Figure 5 shows the average ability of each algorithm in processing the frames per second. It can be seen that FMTMT was the fastest approach and DMTMT came next. Both margin-based approaches could easily handle a frame rate of more than 65fps.

It can be seen from the results that the margin-based approach significantly improves the speed. For the MTCNN detector, the fixed margin (FMT) has reduced the time per frame from 59.82ms to 15.65ms and the dynamic margin (DMT) has achieved a time per frame of 17.37ms, bringing the processing time down by roughly 400% of the non-margin-based algorithm.

Similarly, the time per frame for the hybrid MTCNN and template matching approach (NTMT) has been significantly improved from 61.89ms to 14.22ms with fixed margin (FMTMT) and 14.93ms with dynamic margin (DMTMT). The longer time in the dynamic margin is due to the extra pixels taken proportional to the movement in the face position.

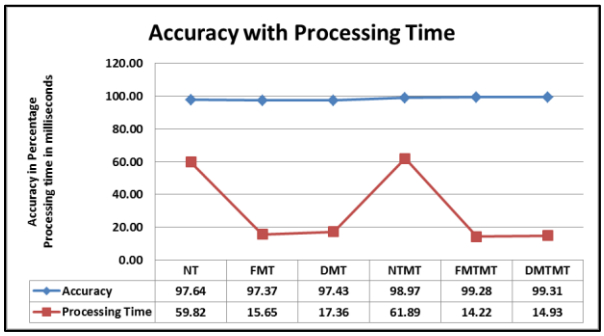

Figure 6. Accuracy with processing time comparison chart.

Figure 6 summarize both performance parameters i.e. processing time in milliseconds and average accuracy achieved in percentage for each algorithm. It shows that the FMTMT and DMTMT show significant performance improvement in terms of both accuracy and processing time.

VI. Conclusion

This work proposes the use of MROI with the hybrid MTCNN and TM face detectors in order to improve processing time as well as the face detection accuracy. Various experiments were performed with six combinations of different components of the algorithm. It allowed us to observe the impact of each component separately on the speed and accuracy performance of the resulted algorithm. The introduction of TM as an escape routine has boosted the accuracy of the MTCNN detector from 97.64% to 99.31%. Moreover, the MROI helped to boost the processing speed by 400%, i.e. from 59.82ms to 14.22ms per frame. The DM based approach achieved the highest accuracy. However, the FM is slightly faster than the DM. It is observed that in case of significant face movement the DM is a better choice. Further study can be made by evaluating these algorithms on videos that involve fast movement.

References

[1] S. Zafeiriou, C. Zhang, and Z. Zhang, "A survey on face detection in the wild: Past, present and future," Comput. Vis. Image Underst., vol. 138, pp. 1–24, 2015.

[2] C. Zhang and Z. Zhang, "A Survey of Recent Advances in Face Detection," Microsoft Res., no. June, p. 17, 2010.

[3] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," Comput. Vis. Pattern Recognit., vol. 1, p. I--511--I--518, 2001.

[4] K. Zhang, Z. Zhang, Z. Li, S. Member, Y. Qiao, and S. Member, "Joint Face Detection and Alignment using Multi-task Cascaded Convolutional Networks," IEEE Signal Process. Lett., vol. 23, no. 10, pp. 1499–1503, 2016.

[5] L.-Y. Wei and M. Levoy, "Fast texture synthesis using tree-structured vector quantization," Proc. 27th Annu. Conf. Comput. Graph. Interact. Tech. - SIGGRAPH '00, pp. 479–488, 2000.

[6] "http://ailab.space/projects/multimodal-human-intention-perception/," Data/Code Section, 2017.

[7] P. Viola and M. Jones, "Robust real-time face detection," Int. J. Comput. Vis., vol. 57, no. 2, pp. 137–154, 2004.

[8] G. Bradski, "The OpenCV Library.," Dr. Dobb's J. Softw. Tools Prof. Program., vol. 25, no. 11, pp. 120–123, 2000.

[9] D. Chen, S. Ren, Y. Wei, X. Cao, and J. Sun, "Joint cascade face detection and alignment," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 8694 LNCS, no. PART 6, pp. 109–122, 2014.

[10] G. Li, Haoxiang and Lin, Zhe and Shen, Xiaohui and Brandt, Jonathan and Hua, "A Convolutional Neural Network Approach for Face Detection," in The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 5325–5334.

[11] S. Yang, P. Luo, C.-C. Loy, and X. Tang, "From Facial Parts Responses to Face Detection: A Deep Learning Approach," 2015 IEEE Int. Conf. Comput. Vis., no. 3, pp. 3676–3684, 2015.

[12] H. Jiang and E. Learned-Miller, "Face Detection with the Faster R-CNN," in 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017), 2017, pp. 650–657.

[13] N. N. Dawoud, B. B. Samir, and J. Janier, "Fast Template Matching Method Based Optimized Sum of Absolute Difference Algorithm for Face Localization," Int. J. Comput. Appl., vol. 18, no. 8, pp. 975–8887, 2011.

[14] T. K. T. T. K. Tan, C. S. B. C. S. Boon, and Y. S. Y. Suzuki, "Intra Prediction by Template Matching," 2006 Int. Conf. Image Process., no. September, pp. 1–4, 2006.

[15] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista, "High-Speed Tracking with Kernelized Correlation Filters," IEEE Trans. Pattern Anal. Mach. Intell., vol. 37, no. 3, pp. 583–596, 2015.

[16] D. Gerónimo, A. D. Sappa, D. Ponsa, and A. M. López, "2D-3D-based on-board pedestrian detection system," Comput. Vis. Image Underst., vol. 114, no. 5, pp. 583–595, 2010.

[17] D. Held, J. Levinson, S. Thrun, and S. Savarese, "Robust real-time tracking combining 3D shape, color, and motion," Int. J. Rob. Res., vol. 35, no. 1–3, pp. 1–28, 2015.

[18] B. Rehman, O. W. Hong, A. Tan, and C. Hong, "Hybrid Model with Margin-Based Real-Time Face Detection and Tracking," in The 11th Multi-disciplinary International Workshop on Artificial Intelligence (MIWAI). Lecture Notes in Computer Science, 2017, vol. 10607, pp. 360–369.