Hybrid Model with Margin-Based Real-Time Face Detection and Tracking

Bacha Rehman(✉), Ong Wee Hong, and Abby Tan Chee Hong

Faculty of Science, Universiti Brunei Darussalam,
Bandar Seri Begawan, Brunei Darussalam
bachapk@gmail.com, {weehong.ong, abby.tan}@ubd.edu.bn

Abstract. Face detection and tracking algorithms mainly suffer from low accuracy, slow processing speed, and poor robustness when meet with real-time setup. The problem becomes crucial in real-time situations such as in human robot interactions (HRI) or video analysis. A margin-based region of interest (ROI) hybrid approach that combines Haar cascade and template matching for face detection and tracking is proposed in this paper to improve the detection accuracy and processing speed. To speed up the processing time, region of interests (ROIs) with fixed and dynamic margin concepts are used. A dataset comprising of ten RGB video streams of fifteen seconds have been created from real-life videos containing a person in lecture delivering environment. In each video, there exists person's movement, face turning and camera movements. An accuracy of 97.96% with processing time of 10.76 ms per frame has been achieved. The proposed algorithm can detect and track faces in sideway orientation apart from frontal face. The proposed approach can process the video streams at the speed above 90 frames per second (FPS). The proposed approach reduces processing time by ten times and with a boost to accuracy in comparison to the conventional full frame scanning techniques.

Keywords: Face detection · Face tracking · Haar cascade · Template matching · Dynamic margin · ROI

1 Introduction

Face detection within an image is an important field of research in human computer interaction and computer vision [1, 2]. It is also a necessary step in face recognition. Several researches on automatic face detection have been carried out. The inspiring work of Viola and Jones [3] has recognized the two basic principles in face detection for applied solutions as simple features and boosted cascade structure. Majority of the academia and industrial real-time face detector applications are based on the said two principles. Such face detection applications work quite good for nearly frontal faces under usual conditions. However, they lack the effectiveness for faces with non-frontal orientation in addition to challenges under rough real world situations, such as expression, lighting, and occlusion. The reason behind this ineffectiveness is because the simple Haar-based features are not sufficient to detect large variations in the face orientation and other facial and ambient properties. This technique also involves huge

S. Phon-Amnuaisuk et al. (Eds.): MIWAI 2017, LNAI 10607, pp. 360–369, 2017.
https://doi.org/10.1007/978-3-319-69456-6_30

computational processing which leads to high computational time, not suitable for real-time system involving face detection [4].

In this paper, a combined face detector and tracker with margin-based ROI to improve both the speed and accuracy is presented. It combines both Haar cascade and template matching principles to improve accuracy. Haar cascade [3] algorithm requires some fine tuning to perform really fast and to deal with non-frontal face orientation. To overcome these deficiencies, template matching method [5] is used that finds the resemblance between the input images and the template images.

Template matching method can use the relationship between the input images and stored standard pattern of face features, to detect the existence of a face in an image. The benefit of this method is that any template image can be used regardless whether it is frontal or otherwise. Moreover, based on the correlation values i.e. corresponding to the template and input image common pattern, it is very easy to apply the algorithm. Further it can easily determine the face location as well as eyes, nose, mouth and other features of a face. The method can also be applied on various variations of the images.

In addition, the concepts of fixed and dynamic margin are used to improve the processing speed. Face tracking is used to provide the necessary information for the margin based algorithms to process subsequent frames. If face is found in the first place, the face position is stored for the computation of region of interest (ROI) in the next frame. In the ROI calculations, two variations are presented, i.e. fixed margin and dynamic margin. In fixed margin, fixed percentage extra pixels is added around the face area of the detection face in the initial frame. In the dynamic margin calculation, the margin corresponds to the change in the face position in the previous frames. This procedure significantly speeds up the detection process. As the algorithm is developed keeping in mind if Haar cascades fail, the template matching algorithm calculates the most prospective face position based on the face detected in previous frames. This variation of the detection algorithm makes it robust and reliable.

Six algorithms are implemented and compared in this paper. The detector and tracker presented takes around 10 ms on average which is 10 times faster than the conventional algorithm [3]. It also achieves high detection accuracy on the dataset [6] of 10 videos of 15 s. The videos contain a person who is moving around and occasionally turning his/her head in a normal lecture delivering environment. The main contributions of this paper are:

1. Proposed the face detection and tracking approach incorporating margin-based and template matching with Haar cascade detector to achieve high accuracy and fast speed face detection in real-time.
2. Implemented six variations of the proposed hybrid approach and evaluated their performance on real-life videos.

2 Related Work

Different face detection algorithms [1] have been developed and applied from time to time. One of the most popular algorithm regarding face detection is presented in [3, 7], which has become a benchmark in many software packages e.g. OpenCV [8]. However these conventional algorithms have some serious issues with speed and confronting non-frontal face images [4]. To address these issues, the work done in [4] presented variation of the conventional Haar cascade algorithm [3, 7]. This effort [4] significantly improves the processing time and reduce it to 28 m-sec per frame for a 40 × 40 frame size, and also improved the detection of non-frontal faces to some extent. However there is still a need to improve the detection technique and also the processing time to enable it to be used for real time system e.g. HRI. Nevertheless, algorithm presented in [4] can be considered as a good alternative of the conventional Haar cascade algorithm [3] for future reference. Having said that, there are other works going on currently to improve the conventional Haar cascade method and to be used for various purposes e.g. a cascade based Deep Neural Networks (DNNs) is proposed [9] to obtain pose estimation results with high precision. Similarly [10–12] adapted multi-task cascaded convolutional networks for face detecting using DNNs.

Template matching algorithm for face localization has been presented in [13], while few variations of template matching techniques are presented in [14]. As template matching algorithm is fast by its nature [5], therefore, it is worth a try to combine it with the conventional cascade algorithm in an intelligent way for effective face detection and tracking task.

Centered correlation filters based tracking systems have been discussed in [15]. Haar wavelet and edge orientation based feature are used to ROI grouping and classification for the purpose of pedestrian detection were discussed in [16]. Re-registration and dynamic template based approach has been developed for head motion recovery in [17]. Template matching based approach addressed issues regarding shape, color and motion was discussed in [18]. The works mentioned above helped understanding various concepts to develop the hybrid system with improved accuracy and processing time.

3 Proposed Margin-Based Hybrid Approach

The algorithm developed as a result of this work is a hybrid approach using Haar cascades and template matching. The main face detector is the Haar cascade, however, in case the Haar cascade fails, the system switches to the template matching detector under certain stopping scenario including time and edge detections. In addition, fixed or dynamic margin-based region of interest (ROI) is used to achieve fast face tracking.

Once a face is detected, its position, face region and template are stored and the ROI is calculated around it. In fixed margin, extra pixels around the face area are taken at fixed percentage i.e. 25% on each side. In the dynamic margin calculation, fixed margin is taken as initial margin with additional pixels corresponding to the change in the face position in previous frames. The detector, either Haar cascade or template matching is applied with this margin-based ROI (MROI) to achieve reliable tracking. Figure 1 describes the general algorithm of the margin-based ROI approach for face detection and tracking. The downscaling of frame size is taken as half. In this work the frame size is downsized to 320 × 240 from the original size of 640 × 480.

```
MROI Algorithm
FOR each frame in the video stream DO
1.  Get the next frame f from the video stream source
2.  Downscale frame size keeping the aspect ratio
3.  Obtain the ROI based on previous frame information
4.  Apply algorithm (Select from equation (5) to (10))
5.  IF face found
    5.1. Update the face position
    5.2. Update the ROI with the given margin m
    5.3. Update the face template
    5.4. Update the face area
END FOR
```

Fig. 1. Margin-based region of interest (MROI) algorithm.

Mathematically, the Haar cascade and template matching detector are expressed in Eqs. (1) and (2) respectively.

$$F(n)_z = \sum_{(0,0)}^{(x,y)} \sum\nolimits_{\frac{m}{10}}^{\frac{m}{2}} (Multiscale_{HC_m})_z \qquad (1)$$

$F(n)$ represents the conventional Haar cascade [3] where, z is the frame number and m is pixels scale window. The minimum size of the window is $\frac{m}{10}$ while the maximum is half of the image frame, i.e. $\frac{m}{2}$. The points represented by $(0,0)$ till (x,y) denote the whole frame scanning.

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

$TM(x,y)$ represents the equation for the template matching algorithm. I denotes the input image, T is the template, and TM is the result.

The proposed fixed and dynamic margin-based detectors are expressed mathematically in Eqs. (3) and (4) respectively.

$$F(fm)_z = \sum_{(x1,y1)}^{(x2,y2)} \sum\nolimits_{r*\frac{1}{3}}^{r*\frac{6}{5}} (Multiscale_{HC_r})_z \qquad (3)$$

$F(fm)$ represents face detection method in fixed margin-based ROI approach. In Eq. (3), r represents the region of interest area which is from the points $(x1,y1)$ to $(x2,y2)$. In fixed margin, extra pixels around the face area are taken at fixed percentage i.e. 25% on each side. The scale of the windows are taken as related to the ROI with the minimum size as one third of the ROI area and the maximum size is *20%* extra from the ROI.

$$F(dm)_z = \sum_{(x1,y1)-Dy_r}^{(x2,y2)+Dy_r} \sum\nolimits_{r*\frac{1}{3}}^{r*\frac{6}{5}} (Multiscale_{HC_r})_z \tag{4}$$

$F(dm)$ represents the equation for dynamic margin approach. It is quite similar to $F(fm)$ with an addition of Dyr, which represents the dynamic extra pixels taken which is proportional to the face movement in previous two frames.

Based on the algorithms in Eqs. (1) to (4), six further algorithms have been implemented and they are expressed mathematically in Eqs. (5) to (10). The six algorithms are labeled 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) without margin-based ROI
5. Fixed Margin with Template Matching Face Tracking (FMTMT)
6. Dynamic Margin with Template Matching Face Tracking (DMTMT)

$$F(NT)_z = F(n)_z \tag{5}$$

$$(FMT)_z = \begin{cases} F(n)_z & F(n)_{z-1} = 0 \parallel F(fm)_{z-1} = 0 \parallel z = 1 \\ F(fm)_z & otherwise \end{cases} \tag{6}$$

$$F(DMT)_z = \begin{cases} F(n)_z & F(n)_{z-1} = 0 \parallel F(dm)_{z-1} = 0 \parallel z = 1 \\ F(dm)_z & otherwise \end{cases} \tag{7}$$

$$F(NTMT)_z = \begin{cases} F(n)_z & n \geq 10 \parallel z = 1 \\ \sum_{n=1}^{10} TM(x,y)_z & F(n)_{z-1} = 0 \end{cases} \tag{8}$$

$$F(FMTMT)_z = \begin{cases} F(n)_z & F(n)_{z-1} = 0 \parallel z = 1 \\ \sum_{n=1}^{10} TM(x,y)_z & F(fm)_{z-1} = 0 \\ F(fm)_z & n \geq 10 \end{cases} \tag{9}$$

$$F(DMTMT)_z = \begin{cases} F(n)_z & F(n)_{z-1} = 0 \parallel z = 1 \\ \sum_{n=1}^{10} TM(x,y)_z & F(dm)_{z-1} = 0 \\ F(dm)_z & n \geq 10 \end{cases} \tag{10}$$

For algorithms involving template matching, $n = 1\ to\ 10$ means that whenever the routine is switched to the template matching, then it will process the next *10* frames before switching back to the Haar detector. In addition to the variables specified in the above equations, a distance variable is used to record the difference between the previous frame face position and the current frame face position. If the distance exceeds a certain threshold then it is considered as a wrong detections and the face tracking switches from Haar detector to template matching for continuation of face detection and tracking. Haar cascade detector sometimes has false detection. Therefore, by introducing the distance filter such situations are minimized.

4 Experimental Setup

4.1 Dataset

A database of 10 videos have been created in this work and are made available at [6]. The videos are extracted from open domain sources in YouTube. The videos in the database have the following properties:

1. All videos contain one face and there are changes in the face orientation.
2. The length of each video is about 15 s each, roughly 450 frames with resolution 640 × 480.
3. The videos contain some popular personalities in a lecture delivering environment.
4. Both the camera and the person in the video are moving.
5. 7 videos of male and 3 videos of female (Fig. 2).

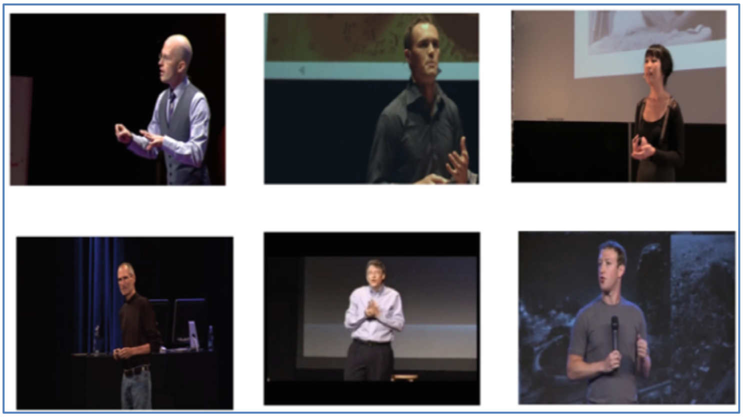


Fig. 2. Screen shots of some of the videos in the database.

4.2 Experiment

Six algorithms have been developed and compared for face detecting and tracking. They have been described in Sect. 3. The conventional algorithm (NT) gives the base idea about accuracy and speed of the algorithm.

All these algorithms are tested on the dataset [6] to reach the conclusions in terms of accuracy (correct, incorrect, and not detected), average time taken per frame, and ability to process the number of frames per second.

Each algorithm is executed ten times on each video file. The accuracy, processing time and FPS performance are calculated as the average of the ten test results (Table 1).

4.3 Development Environment

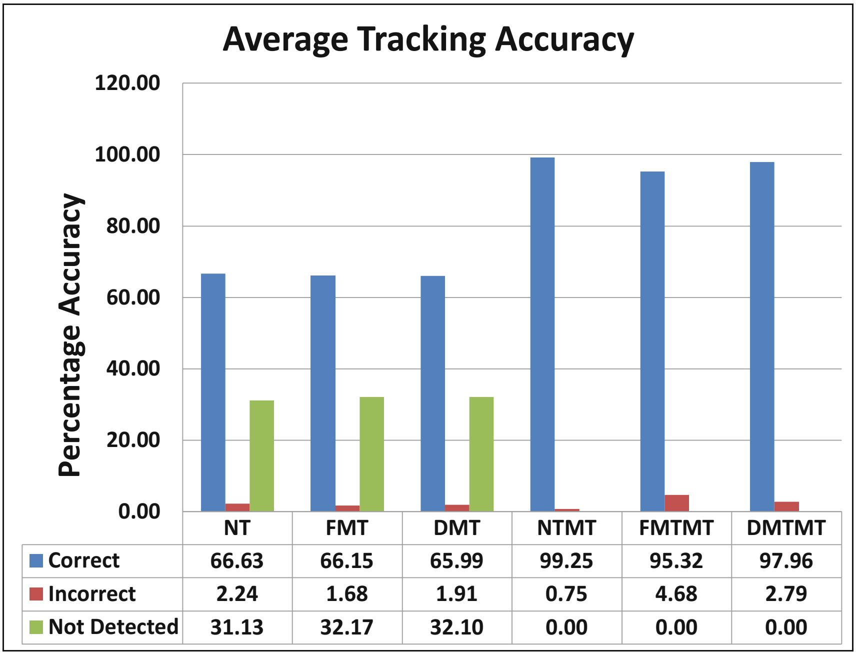
Table 1. Hardware and software used in the development

Hardware	CPU	Intel® Core™ i5 CPU 650 @ 3.20 GHz
	RAM	8 GB
Software	OS	Widows 8.1 pro 64 bits
	Language	Microsoft Visual C++ community 2015
	Tool	OPENCV 3.1

5 Results

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

Figure 3 shows the average face tracking accuracy for each of the algorithms. It can be seen that the incorporation of template matching has significantly improved the accuracy of the Haar-based tracking from 66.63% (NT) to 99.25% (NTMT). The introduction of margin-based approach did not help in improving the accuracy as can be seen from the results for FMT, DMT, FMTMT and DMTMT. There appears to be slight decline in the accuracy when the margin-based approach is used. However, the

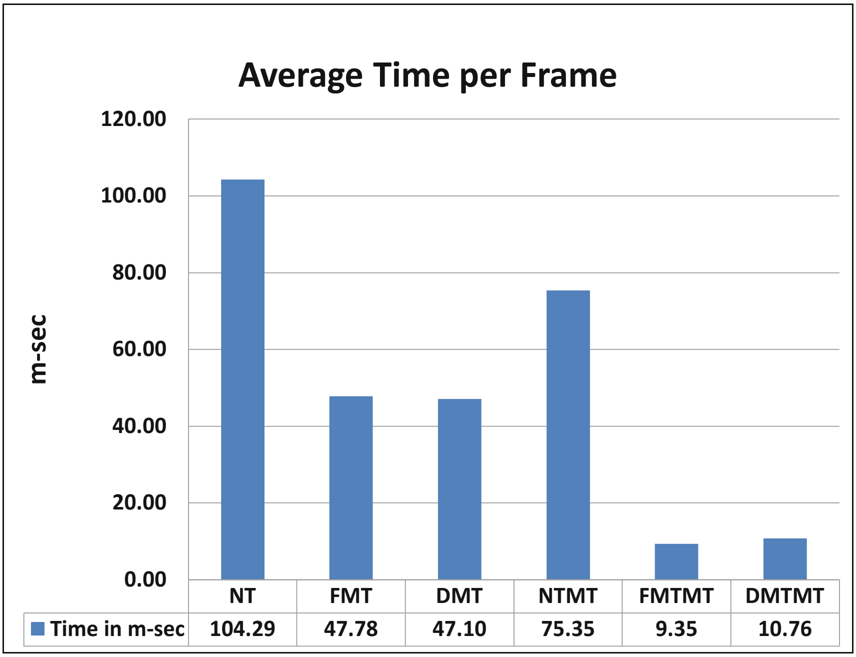


	NT	FMT	DMT	NTMT	FMTMT	DMTMT
🟦 Correct	66.63	66.15	65.99	99.25	95.32	97.96
🟥 Incorrect	2.24	1.68	1.91	0.75	4.68	2.79
🟩 Not Detected	31.13	32.17	32.10	0.00	0.00	0.00

Fig. 3. Average face tracking accuracy.

decline is within 2% and DMTMT has achieved an accuracy of 97.96%. The result shows that the dynamic margin is more robust than the fixed margin in detecting faces.

Turning to the results shown in Fig. 4 will shade light on the merits of the proposed margin-based approach. Figure 4 shows the average time to process or detect a face in each frame. The lower the value the faster is the algorithm. From the result, it can be seen that the margin-based approach significantly improves the speed of the algorithm. For the Haar cascade detector, the fixed margin (FMT) has reduced the time per frame from 104.29 ms to 47.78 ms and the dynamic margin (DMT) has achieved a time per frame of 47.10 ms bringing the processing time down to 45% of the non-margin based algorithm.



	NT	FMT	DMT	NTMT	FMTMT	DMTMT
■ Time in m-sec	104.29	47.78	47.10	75.35	9.35	10.76

Fig. 4. Average time per frame.

Likewise, the time per frame for the hybrid Haar cascade and template matching approach (NTMT) was significantly improved from 75.35 ms to 9.35 ms with fixed margin (MFTMT) and 10.76 ms with dynamic margin (DMTMT). The longer time in the dynamic margin is due to the extra pixels taken proportional to the changes in the face position.

Figure 5 shows the average ability of each algorithm in processing the frames per second. It can be seen that FMTMT is the fastest approach and DMTMT the second. Both margin-based approaches can easily handle a frame rate of more than 60 fps.

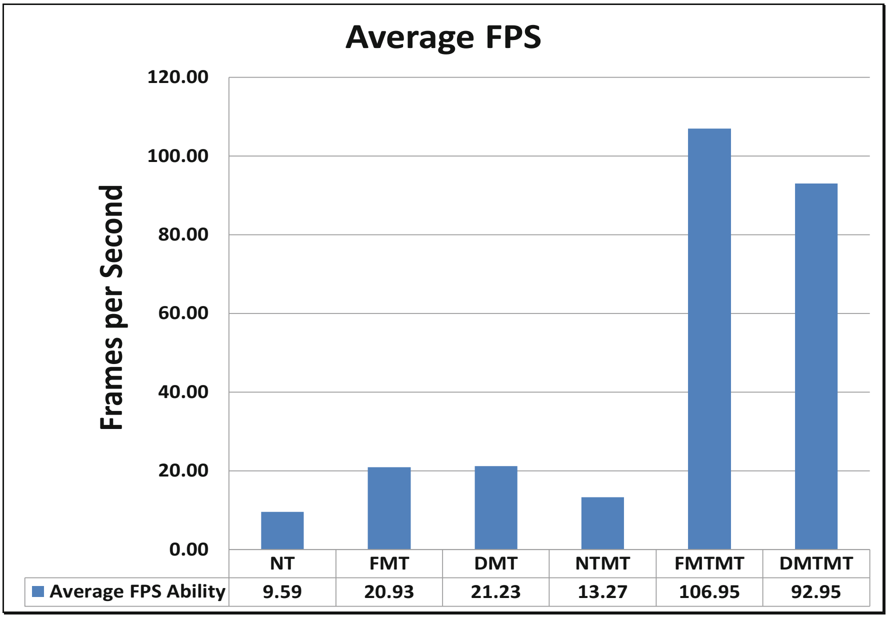


Fig. 5. Average frames per second (FPS).

6 Conclusion

This paper has proposed the use of margin-based ROI with the hybrid Haar cascade and template matching face detector to improve processing speed while achieving high accuracy. Experiments are conducted with six different combinations of the different components of the algorithm to observe the impact of each component. The incorporation of template matching has boosted the accuracy of the Haar cascade detector from 66.63% to 99.25%. On the other hand, the margin-based ROI has speed up the algorithm by 55%, i.e. from 104.29 ms to 47.10 ms per frame. The dynamic margin has achieved higher accuracy than the fixed margin, however the fixed margin is faster than the dynamic margin. If significant movement is involved, the dynamic margin will be a good choice. Further study can be made by evaluating these algorithms on videos that involve fast movement.

References

1. Yang, M.-H., Kriegman, D.J., Ahuja, N.: Detecting faces in image: a survey. IEEE Trans. Pattern Anal. Mach. Intell. **24**, 34–58 (2002)
2. Zhang, C., Zhang, Z.: A Survey of Recent Advances in Face Detection. Microsoft Research (2010)
3. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. Comput. Vis. Pattern Recognit. **1**, I–511–I–518 (2001)

4. Chen, D., Ren, S., Wei, Y., Cao, X., Sun, J.: Joint cascade face detection and alignment. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8694, pp. 109–122. Springer, Cham (2014). doi:10.1007/978-3-319-10599-4_8
5. Wei, L.-Y., Levoy, M.: Fast texture synthesis using tree-structured vector quantization. In: Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques - SIGGRAPH 2000, pp. 479–488 (2000)
6. http://ailab.space/projects/multimodal-human-intention-perception/
7. Viola, P., Jones, M.: Robust real-time face detection. Int. J. Comput. Vis. **57**, 137–154 (2004)
8. Bradski, G.: The OpenCV library. Dr. Dobb's J. Softw. Tools Prof. Program. **25**, 120–123 (2000)
9. Toshev, A., Szegedy, C.: DeepPose: human pose estimation via deep neural networks. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1653–1660 (2014)
10. Zhang, K., Zhang, Z., Li, Z., Member, S., Qiao, Y., Member, S.: Joint face detection and alignment using multi - task cascaded convolutional networks. IEEE Sig. Process. Lett. **23**, 1499–1503 (2016)
11. Ranjan, R., Sankaranarayanan, S., Castillo, C.D., Chellappa, R.: An all-in-one convolutional neural network for face analysis. In: 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017), pp. 17–24 (2017)
12. Jiang, H., Learned-Miller, E.: Face detection with the faster R-CNN. In: 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017), pp. 650–657 (2017)
13. Dawoud, N.N., Samir, B.B., Janier, J.: Fast template matching method based optimized sum of absolute difference algorithm for face localization. Int. J. Comput. Appl. **18**, 975–8887 (2011)
14. Tan, T.K.T.T.K., Boon, C.S.B.C.S., Suzuki, Y.S.Y.: Intra prediction by template matching. In: 2006 International Conference on Image Processing, pp. 1–4 (2006)
15. Henriques, J.F., Caseiro, R., Martins, P., Batista, J.: High-speed tracking with kernelized correlation filters. IEEE Trans. Pattern Anal. Mach. Intell. **37**, 583–596 (2015)
16. Gerónimo, D., Sappa, A.D., Ponsa, D., López, A.M.: 2D-3D-based on-board pedestrian detection system. Comput. Vis. Image Underst. **114**, 583–595 (2010)
17. Xiao, J., Kanade, T., Cohn, J.F.: Robust full-motion recovery of head by dynamic templates and re-registration techniques. In: Proceedings of 5th IEEE International Conference on Automatic Face Gesture Recognition, FGR 2002, pp. 163–169 (2002)
18. Held, D., Levinson, J., Thrun, S., Savarese, S.: Robust real-time tracking combining 3D shape, color, and motion. Int. J. Rob. Res. **35**, 1–28 (2015)