Heliyon 6 (2020) e05107

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

Multi-audience tracking with RGB-D camera on digital signage

Chuan-Chuan Low [a], Lee-Yeng Ong [a,*], Voon-Chet Koo [b], Meng-Chew Leow [a]

[a] Faculty of Information Science & Technology, Multimedia University, Jalan Ayer Keroh Lama, Melaka, 75450, Malaysia
[b] Faculty of Engineering & Technology, Multimedia University, Jalan Ayer Keroh Lama, Melaka, 75450, Malaysia

ARTICLE INFO

Keywords:
Computer science
Face detection
Face tracking
Depth camera
Digital signage

ABSTRACT

Digital signage is widely utilized in digital-out-of-home (DOOH) advertising for marketing and business. Recently, the combination of the digital camera and digital signage enables the advertiser to gather the audience demographic for audience measurement. Audience measurement is useful for the advertiser to understand the audience's behavior and improve their business strategies. When an audience is facing the digital display, the vision-based DOOH system will process the audience's face and broadcast a personalized advertisement. Most of the digital signage is available in an uncontrolled environment of public areas. Thus, it poses two main challenges for the vision-based DOOH system to track the audience's movement, which are multiple adjacent faces and occlusion by passer-by. In this paper, a new framework is proposed to combine the digital signage with a depth camera for tracking multi-face in the three-dimensional (3D) environment. The proposed framework extracts the audience's face centroid position (x, y) and depth information (z) and plots into the aerial map to simulate the audience's movement that is corresponding to the real-world environment. The advertiser can further measure the advertising effectiveness through the audience's behavior.

1. Introduction

In recent years, digital signage is widely developed for DOOH advertising. Digital signage can be found in public areas such as shopping malls, airports, stations, and retail shops [1, 2, 3]. The digital signage has an advantage over the traditional signage, which is able to show different multimedia content as requested by the advertiser. Most of the digital signage are designed with the display screen and computer equipment to deliver the message to the public [3, 4]. Instead of using static images, a variety of changeable content, such as video, audio, and animation are normally used to attract the audience's attention [5, 6].

Many types of research about vision-based DOOH systems have been studied and implemented to collect the audience measurement with digital cameras [1, 7, 8, 9]. The audience measurement metric includes the number of audiences, time duration, distance, and height. It helps the advertiser to understand the audience's behavior and measure the effectiveness of advertising. Thus, the vision-based DOOH system is equipped with a digital camera for collecting the ambient information in real-time. The ambient information includes the people that pass-by or stays in front of the display. Whenever the vision-based DOOH system detects an audience's face, it requires to show a personalized advertisement in the shortest time [10]. Therefore, the computational power for face detection and tracking algorithms indirectly affect the responsiveness of the vision-based DOOH system for real-time targeted advertising [11].

In such an uncontrolled environment of the public areas, there are two main challenges that need to be handled by the vision-based DOOH system. The phenomenon where people generally walking alone or together in a group shows the single or multi-person interaction in an uncontrolled environment. When a group of audiences is facing the digital display, the faces that are nearby each other may merge together and are generally treated as a single audience in the counting and tracking procedure [12, 13, 14]. Since the aforementioned digital signage is available in an uncontrolled environment, the passer-by may occasionally occlude the audience who is watching the advertisement. The presence of passer-by will create an obstacle for face detection and tracking algorithms to process the audience's face. Figure 1 demonstrates the sequence of scenes for both of the aforementioned challenges in an uncontrolled environment.

This paper presents a new framework of vision-based DOOH system that utilizes a depth camera to address the challenges of multiple adjacent faces and passer-by occlusion in the uncontrolled environment. DOOH advertising system is commonly equipped with a screen display and limited performance of processor hardware. The integration of the

* Corresponding author.
E-mail address: lyong@mmu.edu.my (L.-Y. Ong).

https://doi.org/10.1016/j.heliyon.2020.e05107
Received 2 October 2019; Received in revised form 30 April 2020; Accepted 25 September 2020

Figure 1. The sequence of the video frames in an uncontrolled environment for (a) multiple adjacent audiences and (b) occlusion by a passer-by.

depth camera is specially designed to detect the audience's face within the region-of-interest (ROI) before determining a personalized advertisement for the audience. The computational power for the vision-based DOOH system indirectly affects the responsiveness of targeted advertising to the presence of the audience in ROI.

Figure 2 shows the bounded ROI used in the proposed framework to discover the presence of the audience. The area of ROI is restricted within 0.5 meter (m) of minimum distance and 2.5m of maximum distance from the digital signage. The vision-based DOOH system detects the audience's face and extracts the depth information (distance) within the ROI. The face information generally contains the two-dimensional (2D) position (x, y). The integration of depth information (z) with the 2D-position is determined as 3D-position (x, y, z) for face tracking.

Figure 3 illustrates the processing flow of the proposed 3D face tracker. During the advertisement broadcasting, the face of the audience who stands within the ROI is detected by the depth camera. Each detected audience in the same video frame is labeled and assigned to an independent tracker path. The proposed 3D face tracker processes the audience's face based on the 3D-position (x, y, z). When each advertisement broadcast is completed, the audience trajectory is recorded into the aerial map based on the tracker path. The aerial map simulates the movement of the audience when he/she is watching the advertisement. Consequently, the advertiser can measure the targeting effect of advertisements based on the audience's behavior.

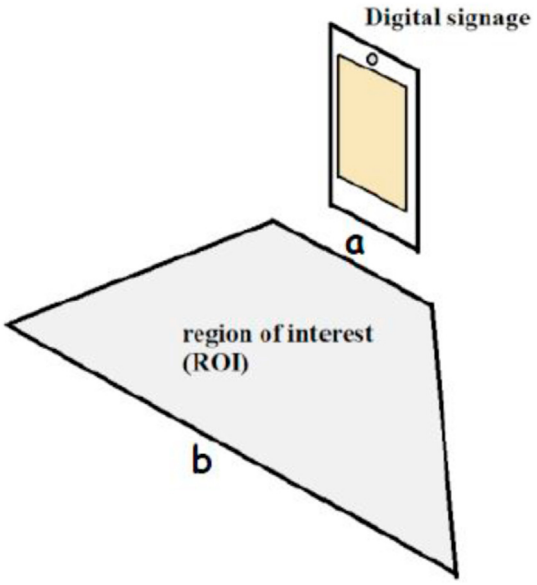

Figure 2. The ROI of the proposed framework is bounded within (a) 0.5m of minimum distance, and (b) 2.5m of maximum distance from the digital signage.

The contributions in this paper are summarized as the following: 1) a proposed framework for DOOH system that utilized the RGB-D camera to detect and track multiple audiences; 2) a new 3D face tracker that integrated the depth (z) with xy-position; 3) a new mapping method to simulate the audience's real-world movement in the aerial map. In this paper, Section 2 describes the related work about the method used by the existing vision-based DOOH system. Section 3 describes the proposed framework with the face tracking algorithm and aerial map mapping method. The experimental setup and results are recorded in Section 4. Lastly, the conclusion is remarked in Section 5.

2. Related works

The DOOH system usually installed in the public area with a high population of audiences. Most of the companies are focusing on digital advertising to promote their product to the anonymous audience. However, the companies unable to verify whether the message of the product is delivered to the targeted audience. The vision-based DOOH system which incorporates a digital camera can provide various information about the audience and ambient information. It allows the companies to analyze the advertising effectiveness towards business improvement.

Several vision-based DOOH systems had been studied for collecting the ambient information in the public space. Ravnik and Solina proposed to process the spatial localization of the audience using a single monocular camera [1, 15]. The audience temporal, spatial, and demographic including dwell time, distance, and gender are extracted. The distance of the audience is determined by measuring the interpupillary distance on the audience's face. The measurement of the audience's interpupillary distance required a more complex algorithm to improve the accuracy of the audience's location. Thus, it needs more computational power to execute the algorithm in real-time processing.

Due to the technology enhancement of the digital camera, a depth sensor is integrated to provide the depth information (z). Hyun et al. utilized the Kinect Sensor to extract the audience's depth information, height, and moving direction [8]. OpenCV is applied to process the face information for gender and age classification. However, the number of the detected face is restricted because the Kinect Sensor is only able to support the height and depth extraction for a maximum of six people.

In a vision-based DOOH system, face detection is the initial step used for gathering the audience measurement. Face detection is basically used to locate and detect the presence of an audience's face in the video frame. Face detection algorithms can be separated into two categories, which are handcrafted feature-based and learned based methods [16, 17]. The

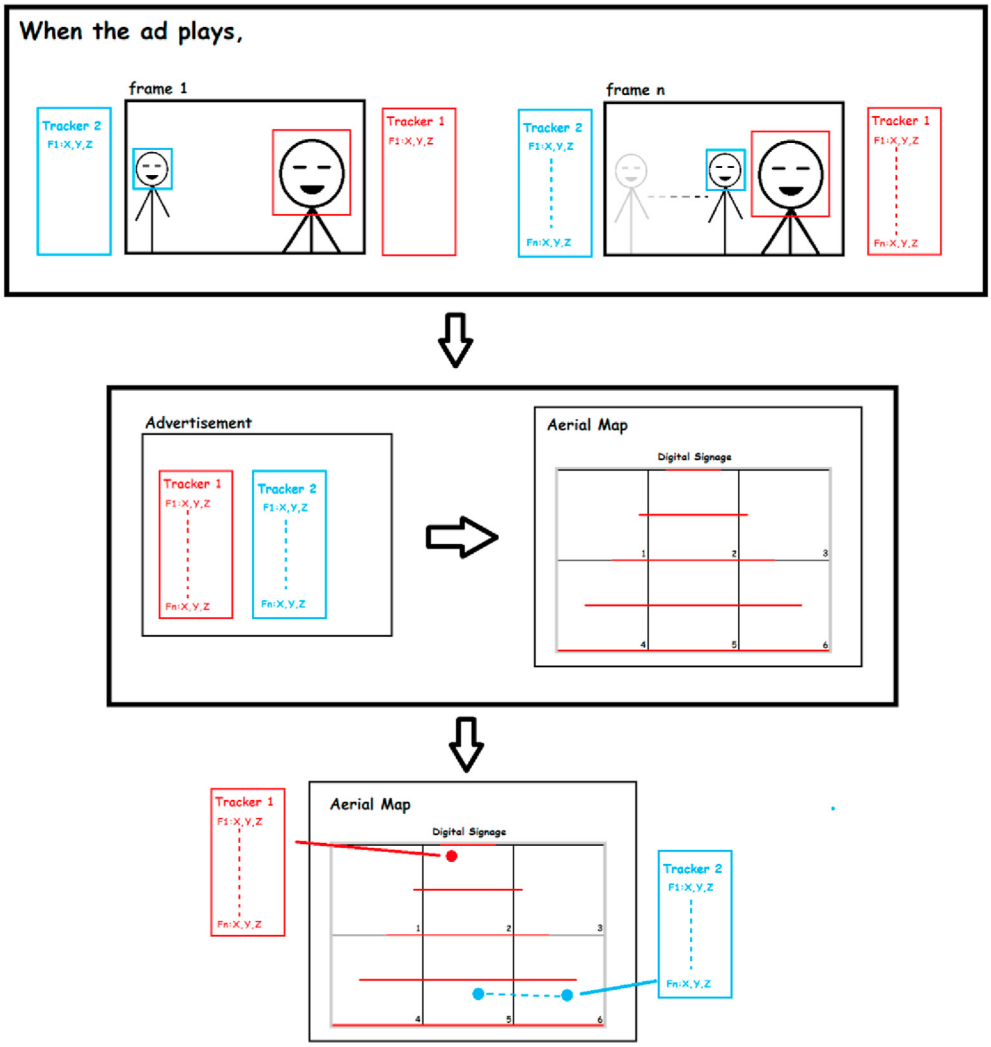

Figure 3. The processing flow of the proposed 3D face tracker.

handcrafted feature-based method detects the face based on the basic features present in the facial region. Skin color detection is one of the commonly-used handcrafted features [11, 16]. On the other hand, the learned-based method is computed using a deep learning algorithm with a number of training samples. The learned-based method is more robust in handling multiple challenges but requires more computational time as compared to the handcrafted feature-based method [17].

Over the past decade, deep learning algorithm becomes a trend in the computer vision research field which achieves a high accuracy result as compared to the handcrafted feature-based method [34, 35, 36]. The deep learning algorithm extracts the image features by using a neural network to classify the features into several classification layers. It allows the computer to learn the unique feature of the face to distinguish or recognize the person's face. P. Nithin et al. [37] proposed an interactive face tracking robot with face recognition through a deep learning algorithm. The system performs face detection using Caffe deep learning framework and applied the local binary pattern (LBP) histogram algorithm for face recognition. It performs the face tracking if the detected person is recognized and matched to the face in the dataset. The various handcrafted feature-based methods such as LBP and Viola-jones are compared with the deep learning algorithm to evaluate the face detection performance. The deep learning algorithm achieves a better accuracy but it requires more computational power and time as compared with other handcrafted feature-based methods. The result shows that the Viola-Jones algorithm with Haar-like feature achieves a lower value on memory usage, latency, and processing delay which enables it to process more frames as compared to the deep learning algorithm.

The Viola-Jones algorithm has been proven to apply in real-time due to the rapid computation and simplicity of face feature extraction with the best feature selection [17]. It is introduced by Paul Viola and Michael Jones in the year 2001 for robust human face detection. It combines the Haar-like feature using the integral image, Adaboost, and cascade classifier [18, 19]. Dang described that the Viola-Jones algorithm achieves better performance in the aspect of precision and recall results as compared to other face detection methods such as Successive Mean Quantitative Transform (SMQT), neural network and support vector machine (SVM) [20]. Although the Viola-Jones algorithm achieved a high detection rate, the challenges including head pose, occlusion by other objects, and lighting conditions are still reducing the positive detection rate [11,21,22].

The skin color feature is incorporated to face detection because skin color is invariant to face size, pose, and expression [23, 24]. Therefore, the skin color feature increased the positive detection rate. Lucena et al. applied the skin color feature in the post-processing method to improve the performance of face detection [25]. Subsequently, Nusirwan et al. [26] combine several color tones of human skin to achieve a higher skin detection rate. However, the performance is affected when the background color is having a similar color region with the skin color. This issue is also highlighted in our recent paper, where the depth camera is used to verify the detected human face within the ROI [27]. The detection rate and computational time of the skin color processing with face

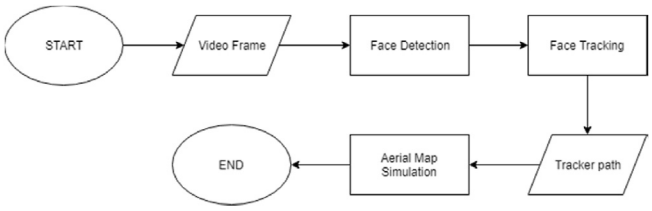

Figure 4. The combination of modules for the proposed framework.

detection are tested with single and multiple person scenes. The RGB-H-CbCr color model achieved 88% detection rate and computational time as short as 15 ms for multi-face processing.

Once the audience's face is verified, the face tracking algorithm is implemented to record the audience trajectory within the ROI. The mean shift is one of the famous algorithms used for face tracking [28]. It is implemented by continually computing the new *xy*-position for the target in the search window until there is no significant position shift in the next frame [29]. The size of the search window is constant no matter the target is far from or near to the camera [28, 30]. However, it will lose the target if the target's size changes.

Q.Cao et al. mentioned that Camshift is a robust algorithm for color-based tracking in real-time applications [29, 31]. The Camshift algorithm is the adaptation of the mean shift algorithm for face tracking [32]. This algorithm updates the search window size and effectively tracks the target with different sizes and shapes. Camshift is sensitive to the skin color which affects the tracking output when the target is occluded or having a similar background color. Cao et al. mentioned that the difficulty of multi-face tracking happened when there are interactions and occlusion between multiple faces [33]. The occlusion issue poses a complication for the face tracking to distinguish the movement of a targeted audience in the 2D-plane, as shown in Figure 1 (b).

3. Methodology

The proposed framework of the vision-based DOOH system extracts the face centroid position and depth information to create a 3D-position *(x, y, z)*. Figure 4 shows the proposed framework. Firstly, the depth camera captures the video frame of the real-world environment within the ROI. Then, the Viola-Jones algorithm detects the audience's frontal face in the video frame. The skin color feature is implemented to filter the false positive of the face images. Once the human face is verified, the centroid position *(x, y)* of the face image is calculated. At the same time, the depth camera gathers the depth information/distance *(z)* between the detected audiences and digital signage. Subsequently, the 3D-position is created for the proposed 3D face tracker, where the audience trajectory is recorded into the tracker path. Lastly, the tracker path is simulated in the aerial map. The details of the methodology are presented in the following subsections.

3.1. Face detection using the Viola-Jones algorithm with skin color

The proposed framework is designed with the Viola-Jones algorithm to detect the presence of the audience from the video frame [1]. However, the detection rate of the Viola-Jones algorithm is affected by the face pose, obstacles from other object and illumination [11, 16, 21, 22, 38]. In order to reduce the false detection rate, the skin color feature is added. Static skin color models are attractive for real-time applications due to the simplicity of implementation and fast performance without requiring any training procedure for skin color features [41].

The combination of several static color tones which is RGB-H-CbCr is applied to define the skin color [26]. This color combination achieves a high detection rate for Asian ethnicity. Three algorithms are formulated to implement the RGB-H-CbCr. Algorithm 1 presents the methodology of processing the skin color to filter every image pixel in the face image. The image pixel that fulfills Algorithm 2, 3, and 4 are finalized as human skin color [26, 27]. Algorithm 2 presents the RGB color space based on two illumination situations, which are normal uniform daylight or the skin color under daylight (outdoor) and flashlight lateral [39]. Algorithm 2 based on a uniform daylight illumination situation achieves better performance for bright skinned persons [40]. Algorithm 3 combines the YCbCr color spaces while Algorithm 4 filters the skin color with the Hue color from the HSV color space. Hue color exhibit a significant difference between skin and non-skin color region [26].

Algorithm 1 Skin Color Detection Processing

Input: The set of face images

Description: The images are processed for human skin color verification

Step 1: Get the pixel of the face image

Step 2: Compute each pixel into three different color tone function

Step 2.1: Compute pixel to Algorithm 2 to obtain the RGB result (R_{GB})

Step 2.2: Compute pixel to Algorithm 3 to obtain the YCbCr result (Y_{CbCr})

Step 2.3: Compute pixel to Algorithm 4 to obtain the HSV result (H_{SV})

Step 2.4: Compare the skin result (S_{color}) from Algorithm 2–4

$S_{color} = (R_{GB}$ && Y_{CbCr} && $H_{SV})$

Step 2.5: If S_{color} = true, filter the pixel into white color. Otherwise, filter it with black color

Step 3: Repeat step 1 to 2 until all the pixels are completed

Step 4: Count the white pixels in of the face images

Step 5: Compare the number of white pixels with the skin threshold

Output: The skin color status of the face image

Algorithm 2 RGB color space for daylight and flashlight laterals

Input: Image pixel from Algorithm 1-Step 2

Step 1: Split the image pixel into different color components like Red (R), Green (G), and Blue (B)

Step 2: Compute the E_1 and E_2 result

If ((R > 95) && (G > 40) && (B > 20) && (Max (R, G, B) – Min (R, G, B) > 15)) && (|R – G| ≥ 15) && (R > B) && (G > B) **Then**

E_1 = True

Else

E_1 = False

End If

If ((R > 220) && (G > 210) && (b > 170) && (|R – G| ≤ 15) && (R > B) && (G > B)) **Then**

E_2 = True

Else

E_2 = False

End if

Step 3: Define the final result

R_{gb} result = E_1 ||E_2

Output: R_{gb} result of the color tone

Algorithm 3 YCbCr color space for skin color detection

Input: Image pixel from Algorithm 1-Step 2

Step 1: Split the image pixel into different color components like Y, Cb, and Cr

Step 2: Compute the A, B, C, D, and E result

A = Cr ≤ 1.5862* Cb + 20

B = Cr ≥ 0.3448 * Cb + 76.2069

C = Cr ≥ - 4.5652 * Cb + 234.5652

D = Cr ≤ - 1.15 × Cb + 301.75

E = Cr ≤ - 2.2857 * Cb + 432.85

Step 3: Define the Y_{cbcr} result

Y_{cbcr} result = (A && B && C && D && E)

Output: Y_{cbcr} result of the color tone.

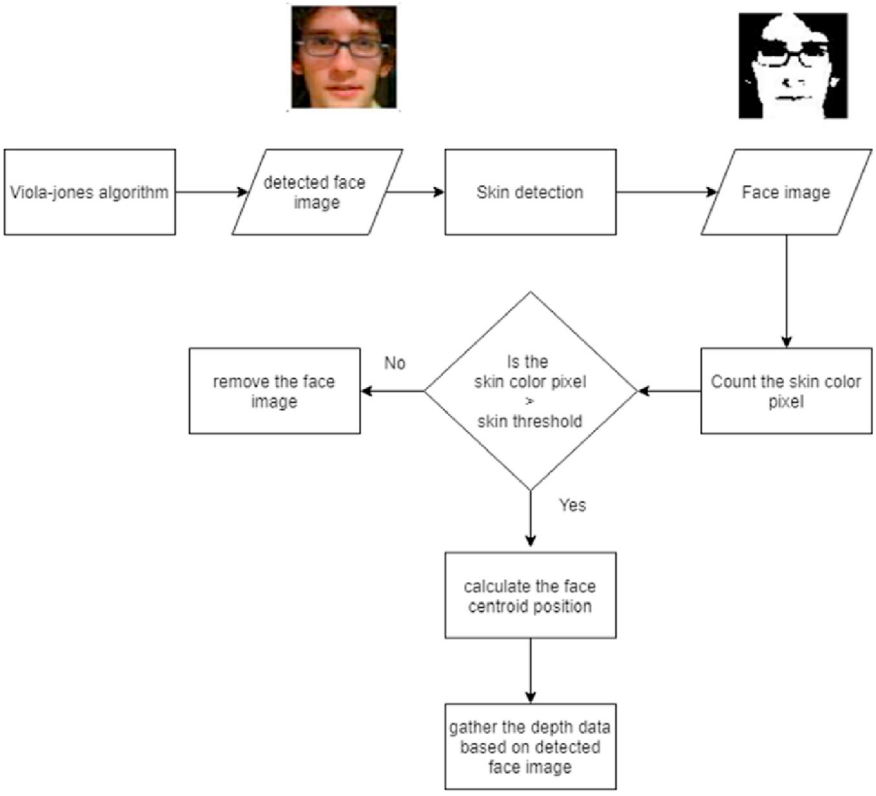

Figure 5. The processes involved in the proposed face detection module.

Algorithm 4 HSV color space for skin color detection

Input: Image pixel from Algorithm 1-Step 2

Step 1: Split the image pixel into different color components like Hue (H), saturation (S), and value (V)

Step 2: Define the H_{sv} result

If (H < 25) || (H > 230) **Then**

H_{sv} = True

Else

H_{sv} = False

End if

Output: H_{sv} result of the color tone

Figure 5 shows the flowchart of the proposed face detection module. The Viola-Jones algorithm detects the audience's face in the video frame sequence. The skin color detection is used to filter the face image pixel with the aforementioned skin color tone. The image pixel that is finalized as skin color is indicated with white color. Otherwise, the black color is used for non-skin representation. Subsequently, the total number of white pixels in the face image is calculated and compared with the skin threshold. The skin threshold is focusing on Malaysian ethnicity skin with the range from white to black-brown skin color. The white pixels region that is greater than the skin threshold is verified as a human face. The face image is excluded if the number of white pixels is less than the skin threshold.

Once the face image is verified as a human face, the selected face centroid position is extracted based on Eqs. (1) and (2). The calculated face centroid position is declared as the 2D-position *(x, y)* before integrating with the depth information. Based on Eqs. (1) and (2), (FC_X, FC_Y) is the *x*- and *y*-positions of the face centroid, and the $(face_x, face_y)$ is the *x*- and *y*-position for the face region. The width and height of the face region is represented with $face_{width}$ and $face_{height}$.

$$FC_X = face_x + \frac{face_{width}}{2} \quad (1)$$

Figure 6. RGB-D camera named Intel Real Sense D430.

Figure 7. Video frame captured from the RGB-D camera.

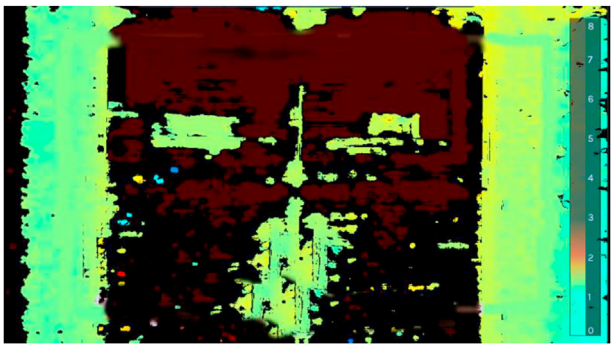

Figure 8. Depth map captured from the RGB-D camera.

$$FC_y = face_y + \frac{face_{height}}{2} \quad (2)$$

3.2. Integration with depth information

Comparing to the digital camera, a depth camera provides extra depth information to estimate the distance from an object to the camera. Figure 6 shows the RGB-D camera named Intel Real Sense D430. Figures 7 and 8 display the RGB video frame and depth map captured by the depth camera. The comparison views are highlighted with green, red, and yellow boxes in Figure 9 (e). The green box represents the depth view region, which occupies 80% of the depth frame while the yellow box only covers 60% of the depth frame. Both depth view regions are obtained from the depth quality tool software that is provided by Intel Realsense [42]. The red box in Figure 9 (e) represents the video frame with an undefined view region with respect to the depth frame.

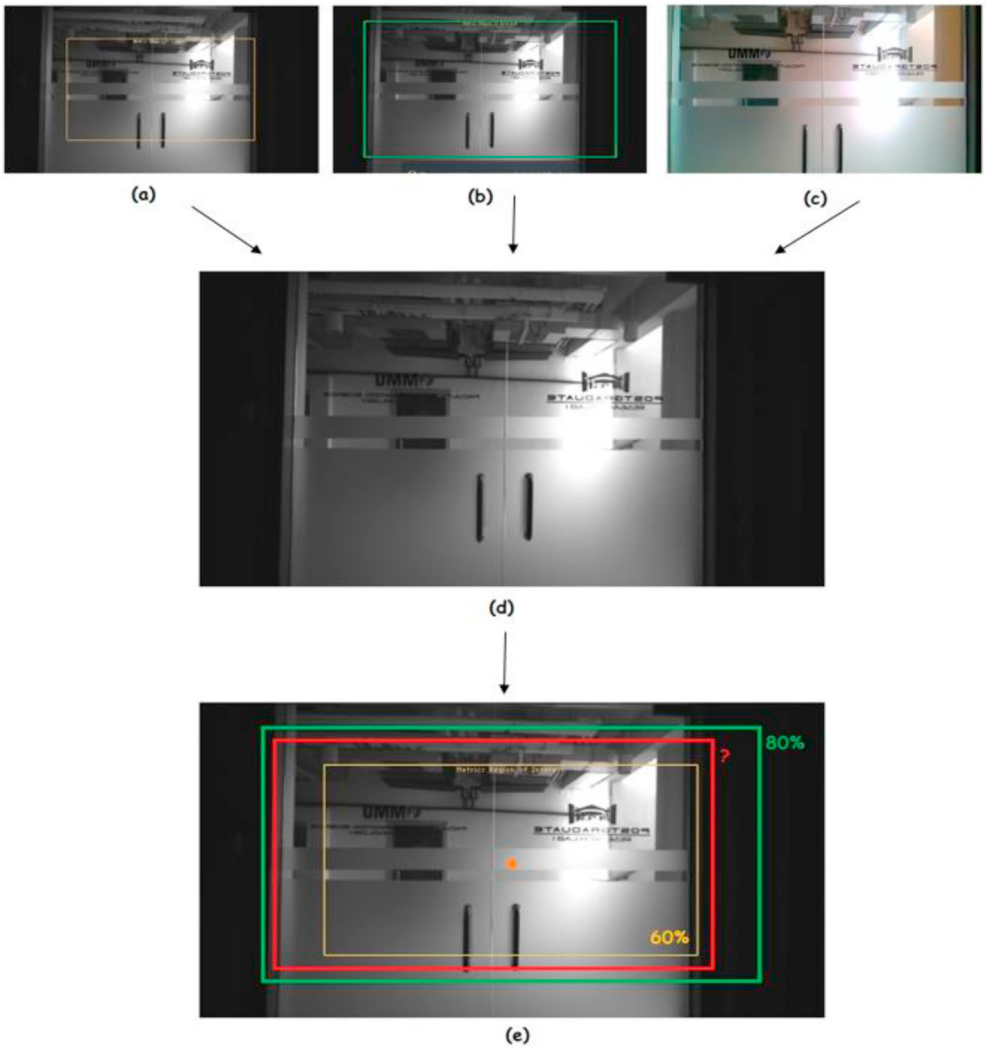

Figure 9. Depth frame with different view regions. (a) 60% of depth view (yellow box), (b) 80% of depth view (green box), (c) Video frame (red box), and (d) 100% of depth frame.

Table 1. The size information (in pixels) for both depth view regions of 60%, and 80% with respect to the depth frame.

View region (box color)	x-position	W	$x + W$	Width ratio (%)	y- position	H	$y + H$	Height Ratio (%)
60% (Yellow)	249	771	1020	**12.8**	137	433	570	**7.2**
80% (Green)	124	1025	1149		64	574	638	

Table 2. The size information (in pixels) for video frame with respect to the depth frame.

View region (box color)	*x*-position	*W*	$x + W$	Width ratio (%)	*y*- position	*H*	$y + H$	Height Ratio (%)
71% (Red)	171	908	1079	**12.8**	112	511	523	**7.2**

The centroid position of the depth frame (orange circle) in Figure 9 (e) is highlighted as the starter point to estimate the ratio between three view regions (a), (b), and (c). In order to establish the relationship between the video frame and depth frame, the ratio of the width and height for different view regions are calculated. Based on Table 1, the width and height ratios for the depth view regions that occupied 60% and 80% are determined with Eqs. (3) and (4). Next, the ratios from the width and height from Table 1 are used to calculate the percentage of the red box (video frame) with respect to the depth frame, as recorded in Table 2. The video frame (red box) occupied up to 71% of the view region with respect to the depth frame.

$$\text{Width ratio} = \frac{W(\text{px})}{\text{depth view region percentage } (\%)} \tag{3}$$

$$\text{Height ratio} = \frac{H(\text{px})}{\text{depth view region percentage } (\%)} \tag{4}$$

where Width and Height ratios are the size ratio for each pixel while *W* and *H* are the width and height of the view region.

The 2D-position of the audience's face in the video frame is calculated to find the adjustment parameter for 2D-position with respect to the depth frame. Based on the close observation in Figure 9 (e), the video frame (red box) is slightly placed to the left side of the depth frame due to the positions of the RGB camera and depth sensor as indicated in Figure 6. Thus, Eqs. (5) and (6) are deduced to adjust the audience's face position in the video frame with respect to the depth frame for *x*- and *y*-positions.

$$x2 = (x1 \times 0.71) + x \text{ position of the video frame (red box) in depth frame} \tag{5}$$

$$y2 = (y1 \times 0.71) + y \text{ position of the video frame (red box) in depth frame} \tag{6}$$

where (*x1, y1*) is the *xy*-position in the video frame, and (*x2, y2*) is the adjusted *xy*-position in the depth frame.

3.3. Face image ratio in different distances

Once the audience's face is detected during the advertisement broadcast, the face features including the position and size are extracted. When multiple faces are detected in a nearby range, the 2D-position (*x, y*) is insufficient to label the faces to the right person. Even though multiple faces are nearby each other, the proposed framework is able to distinguish the audience standing in the front of or behind another audience based on the 3D-position (*x, y, z*). The 2D-position only simulates the audience movement from left to right in the video frame. Meanwhile, the 3D-position is used to simulate the audience's movement in the *x, y,* and *z* directions, which represents the forward, backward, or diagonal movements in the ROI. When the audience is standing nearer to the depth camera, the size of the detected face increases but the number of detected faces reduces. For example, the video frame covers a maximum of eight frontal faces at a distance of 2.5m from the depth camera, but the distance of 0.5m is only sufficient to occupy three frontal faces.

The smiley face in Figure 10 is used to investigate the face's width ratio at different distances. The size of the smiley face is 18 cm × 18 cm. Figure 11 shows the scenario of comparing the smiley face at every distance of 0.5m from the depth camera until it reaches 2.5m. Figure 12 shows the size of the smiley faces that are detected by the Viola-Jones algorithm at every distance of 0.5m. The comparison of the face's width with respect to different distances is sketched in Figure 13. The size of the smiley face increases whenever the smiley face is getting closer to the depth camera. Eq. (7) is deduced for the proposed framework to fine-tune the 3D-position in the video frame to correspond with the audience's actual position in the real-world environment.

$$\text{Ratio} = \text{Smiley face width (px)} \times \frac{\text{smiley face depth (m)}}{\text{ROI maximum depth (m)}} \tag{7}$$

where face width is the width of the smiley face region, and the face depth represents the distance between face and camera. The ROI maximum depth is the distance from 0.5m to 2.5m.

3.4. Proposed 3D face tracker with the search window

The tracker path is created to record the audience's movement in the ROI. Each independent tracker path represents an active audience who is watching the advertisement. Since the audience's movement is predictable within a range, a search window is created to estimate the movement in the next frame. The radius of the search window is estimated based on the width of the detected face from the previous frame. The radius of the search window is calculated every time the tracker path is updated with a new audience 3D-position. Once the advertisement broadcast is completed, the tracker path provides the audience's position on the aerial map.

Figure 14 shows the tracker path with the search window in the proposed 3D face tracker module. The face A1 belongs to an audience's tracker path from the previous frame. The green circle represents the search window for the tracker path A1. When face A and face B are detected in the next frame, the proposed 3D face tracker only considers the face image appearing within the A1 search window. The face image with the shortest distance to face A1 is chosen to update the tracker path for the next tracking procedure. Figure 15 shows the procedure of the proposed 3D face tracker and steps are listed below.

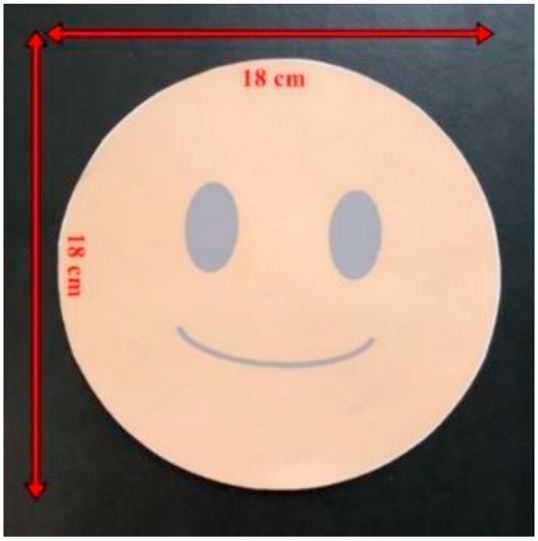

Figure 10. Smiley face.

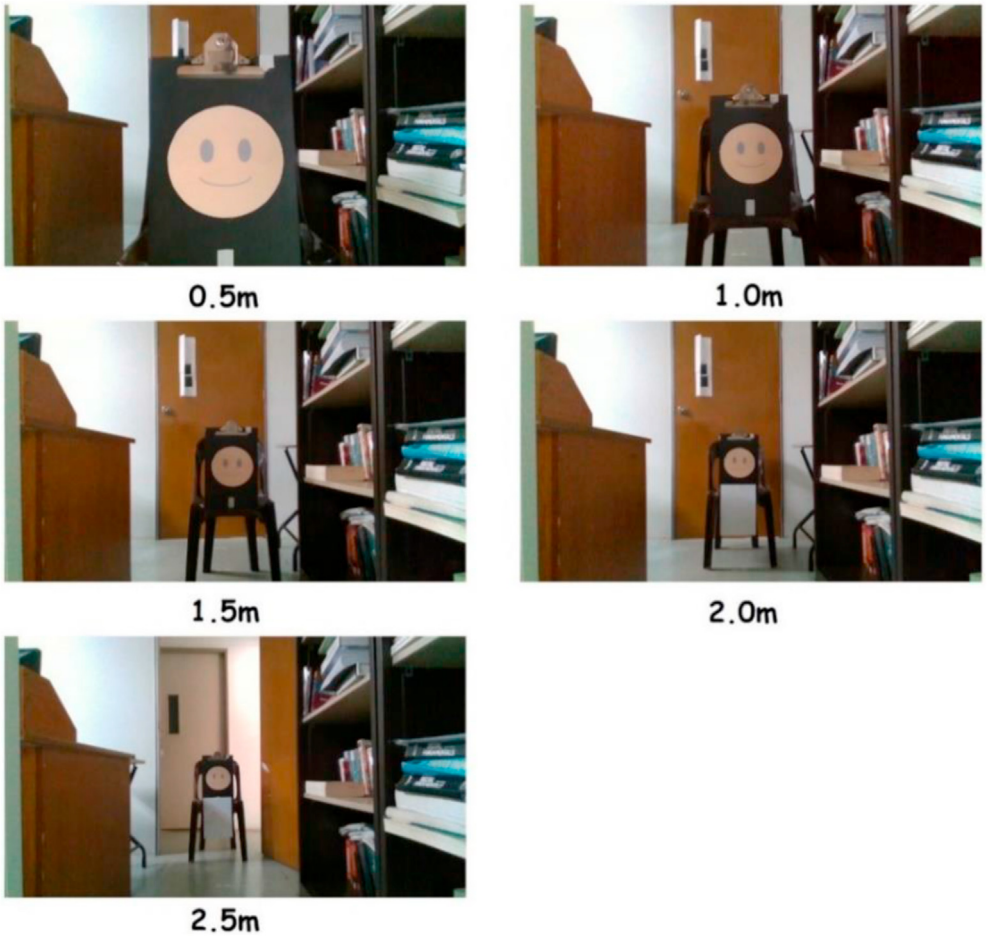

Figure 11. The smiley faces are located at different distances.

1. Obtain the detected audience 3D-position.
2. Check if the audience 3D-position is within any search window of the tracker path. If the audience 3D-position is unable to locate into the tracker path, the new tracker path is created for the audience and then goes to step 5.
3. Calculate the tracker distance between the audience 3D-position with the selected tracker path.
4. The audience 3D-position with the shortest tracker distance is chosen to update the selected tracker path. The unselected audience 3D-position is excluded from the selected tracker path.
5. The proposed 3D face tracker module continues with other audience 3D-position in the same frame. Then go to step 1.

The tracker path for a specific advertisement represents the audience's movement within the ROI. The tracker path is recorded in the aerial map to simulate the audience movement in a real-world environment. Figure 16 shows the design of the aerial map with six regions and the red line indicates the camera view from 0.5m to 2.5m.

The proposed framework is designed with direct and proposed methods for aerial map mapping. The direct mapping method plots the audience's position without adjusting the 3D-position. Eq. (7) is applied for the proposed mapping method to adjust the 3D-position before plotting into the aerial map. Figure 17 shows the smiley face's position in the ROI. The smiley face represents the audience face is arranged for

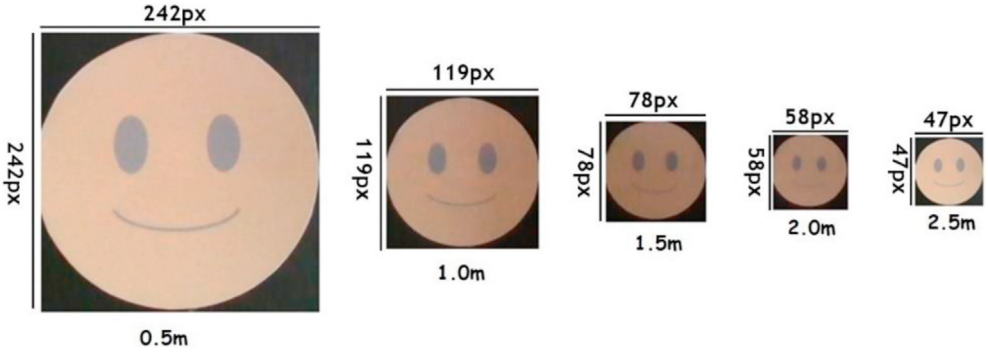

Figure 12. The sizes of smiley faces that are detected by Viola-Jones algorithm.

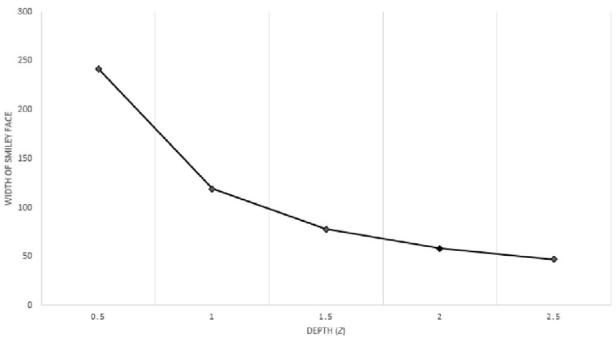

Figure 13. The comparison of smiley face width at different distances.

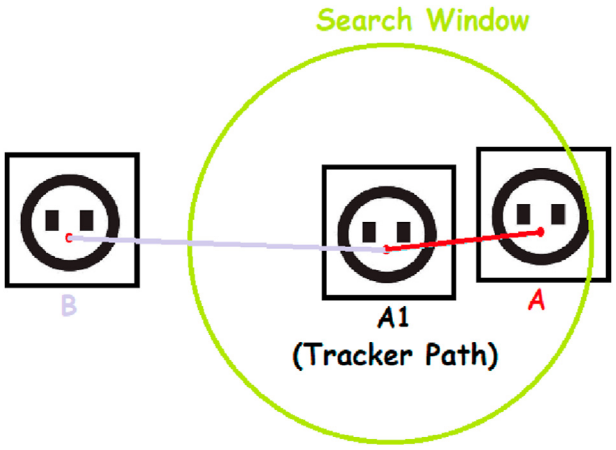

Figure 14. Proposed 3D face tracker module with a search window.

every 0.5m in a straight line. The position of the smiley face is on the left sides of the camera location with 0.4m distance range.

The aerial map from Figure 18 shows the simulation of the smiley face position that corresponds with the actual position in the real-world environment. The blue dot in the aerial map represents the smiley face's position in the real-world environment. In the direct mapping result, the smiley face at the distance of 0.5m is located out of the video frame. It shows that the simulation of the smiley face in the aerial map does not correspond to the video frame. Besides that, the sequence of all smiley face's position in direct mapping is not matching with the smiley face arrangement from Figure 17. Compare to direct mapping, the proposed mapping method locates the smiley face in the aerial map with a similar simulation of the actual position in the real-world environment. It shows a straight-line arrangement of smiley faces at different distances, which is similar to Figure 17.

4. Experiment results and discussion

In this section, two experiments are carried out to evaluate the proposed framework performance in handling two main challenges, which are multiple adjacent faces and occlusion. A depth camera, Intel Real sense D430 is mounted on a tripod of 1.7m height for the experiment. Meanwhile, every 0.5m of distance from the depth camera is measured and marked on the floor within the ROI. These mark points are used as the ground truth position for evaluating the tracking result of the proposed framework. All the experimental results are processed with

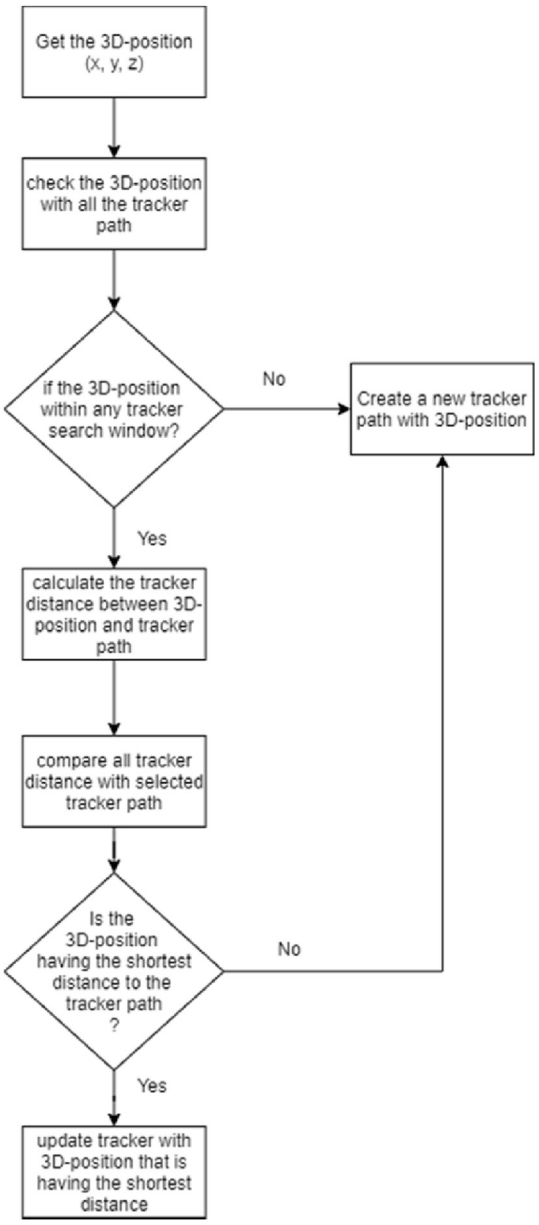

Figure 15. Proposed 3D face tracker module.

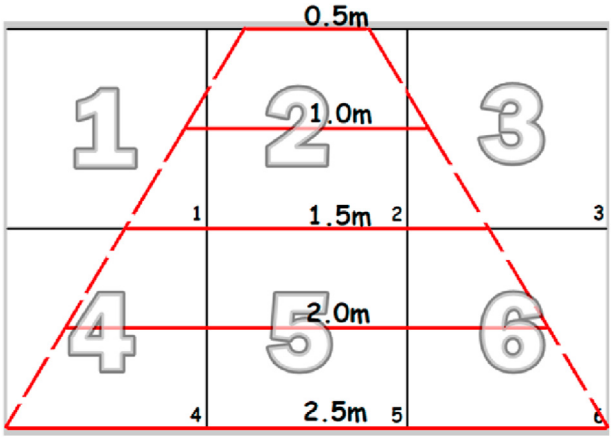

Figure 16. The design of an aerial map with 6 small regions.

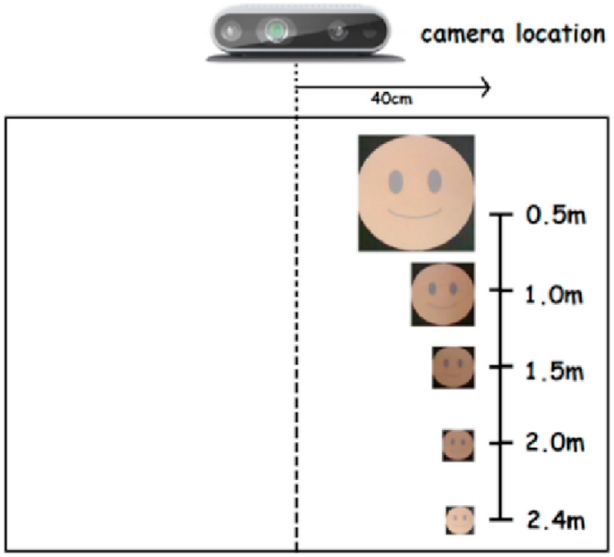

Figure 17. Smiley face position in the ROI.

Microsoft Visual Studio 2017 on the Intel I5 processor 3.30GHz and 16GB RAM. The details of the experiment preparation are listed in Figure 20.

Two of the challenges are simulated with scenarios based on the common situations of the uncontrolled environment while the audience is watching the advertisement. The performance of the proposed 3D face tracker is compared with the existing Camshift tracker algorithm. The Camshift tracker is a two-dimensional color-based tracking. In order to simulate the audience position in the aerial map, the *xy*-position and depth information are required. Therefore, the depth information is provided for Camshift tracker to simulate the audience's position in the aerial map to overcome the shortcoming of the existing 2D face tracking algorithm. In the experiment result, the root means square error (RMSE) in Eq. (8) is applied to measure the performance of the mapping method for each detected audience. The N value in Eq. (8) represents the frame number of the captured video. The $Predicted_i$ represents the audience position in the aerial map, the GT_i is the ground truth position.

$$\text{RMSE for each audience} = \sqrt{\frac{\sum_{I=1}^{N}(Predicted_i - GT_i)^2}{N}} \tag{8}$$

In addition, the T-test evaluation criterion is applied for the performance comparison between the Camshift tracker and the proposed 3D face tracker. The paired-samples T-test is conducted to analyze the significant difference between the Camshift tracker and 3D face tracker for the simulation of audience position in the aerial map. Two hypotheses are formulated to describe the tracking result of the proposed framework:

- The null hypothesis $\boldsymbol{H_0}$ – Two different face trackers are having similar results in face tracking performance.
- The alternative hypothesis $\boldsymbol{H_a}$ – The tracking result for the proposed 3D face tracker is having different tracking results and fewer errors as compared with the result of the Camshift tracker.

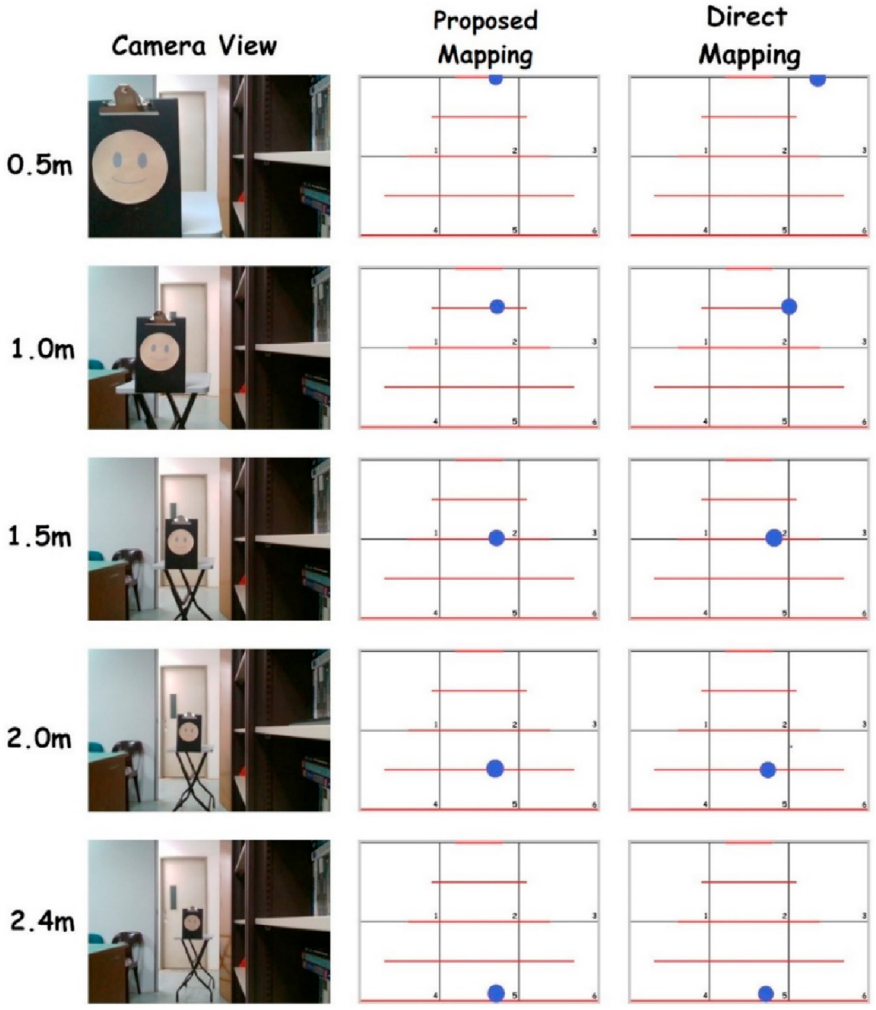

Figure 18. Aerial map result for proposed and direct mapping.

Figure 19. RGB-D camera setting.

Figure 20. Mark point in the region.

Scenario 1: Multiple adjacent audiences

This experiment demonstrates the scenario of multiple audiences that are watching the advertisement. This scenario is designated to prove the robustness of the proposed framework in dealing with multiple faces that are nearby each other. The audience position is recorded on the aerial map with different colors to distinguish various the audience's position in the ROI. Figure 21 shows the experimental procedure for Scenario 1 and the steps are listed below.

1. In this scenario as shown in Figure 22, the smiley face represents audience A is located at a distance of 0.5m while audience B is standing at a distance of 2.0m.
2. The audience B is moving to the right in the horizontal direction for 0.5m, enters the search window range of audience A.
3. The aerial map result is recorded to determine if the Camshift tracker and proposed 3D face tracker are able to distinguish the audience.

The tracker path result is recorded in the aerial map to simulate the audience's position in the ROI. Figure 23 shows the aerial map based on different face tracking algorithms. By referring to Figure 23(a), the aerial map with Camshift tracker shows the position of audience A at 0.5m is not corresponding to ground truth position and resulting a high error value of 0.15m in Table 3. As mentioned earlier in Section 2, the Camshift tracker is sensitive to the color similarity between skin and background. When the audience B is entering the search window of audience A, the Camshift tracker mistakenly estimated audience B as audience A and continuously track audience B in the subsequent frames. Although depth information of audience B has been provided to Camshift tracker, it still suffers the color similarity problem when multiple faces are close to each other. Thus, the RMSE of Camshift tracker for audience A that given in Figure 24 is increased drastically from Frame 20 as highlighted with the blue line.

In order to solve the shortcoming of the existing 2D face tracking algorithm, the search window of audience A is implemented with the depth information (distance) to compare the 3D-position of audience B. The 3D face tracker takes into account the distance between audience B and audience A. If the distance is greater than the depth threshold of 0.25m, audience B is excluded from the consideration to be audience A. In this scenario, audience A maintains at the same position without movement. The 3D face tracker demonstrated a low error value of less than 0.025m throughout the video frames, as highlighted with the red line in Figure 24.

Figure 25 shows the mapping result for the movement of audience B. In this scenario, the audience B is moving horizontally towards the center of the ROI. When he moves into the door area, there is a color similarity between the skin and door. Since audience B is moving during the scenario, the detected facesize is inconsistent due to the changing face pose

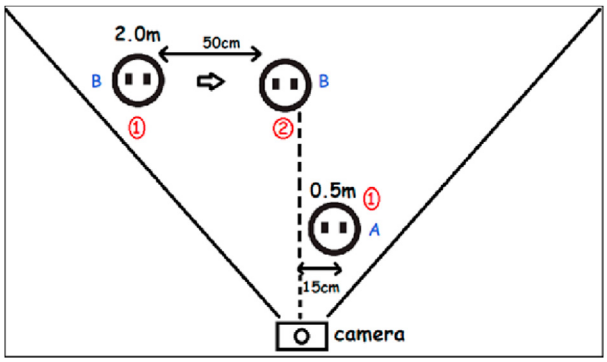

Figure 21. The procedure of Scenario 1, which indicates (1) is the audience A, (2) and (3) is the audience B.

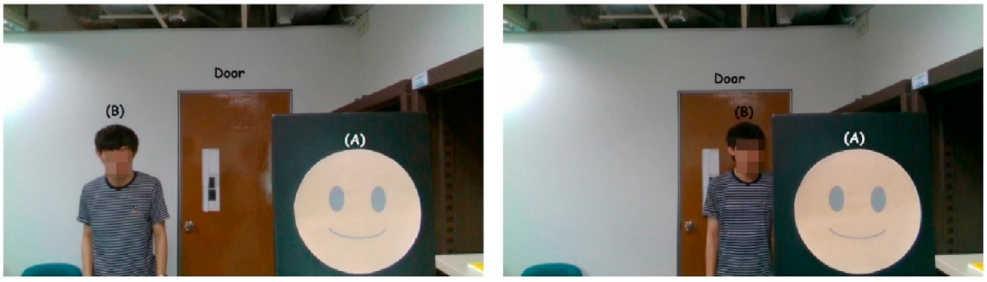

Figure 22. Video frame sample for Scenario 1.

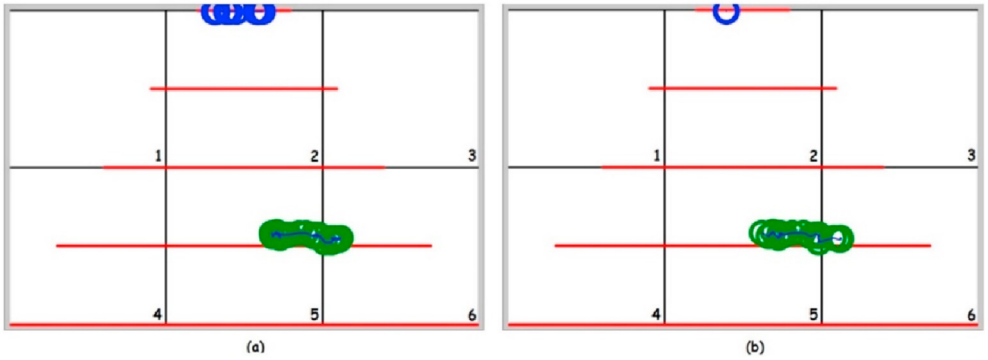

Figure 23. Aerial map result for Scenario 1 with (a) Camshif tracker and (b) 3D face tracker.

Table 3. Average RMSE (meter) result of Camshift and 3D face tracker in Scenario 1.

Audience	A	B	Average Time (ms)
Distance	0.5 m	2.0 m	
Camshift tracker	0.15	0.10	53
3D face tracker	0.019	0.062	50

in the uncontrolled environment. Therefore, the graphs of both face trackers in Figure 25 have unevenly deviated from the ground truth as compared to audience A. For the 3D face tracker result in Figure 25, it illustrates a lower error rate as compared to Camshift tracker. Meanwhile, the average RMSE of the 3D face tracker from Table 3 also shows a smaller error value. Both face trackers are having an average of low computational time that is less than 60milliseconds (ms) to process the face detection and tracking processes for each frame. Thus, the proposed 3D face tracker is proven to simulate the audience movement that is

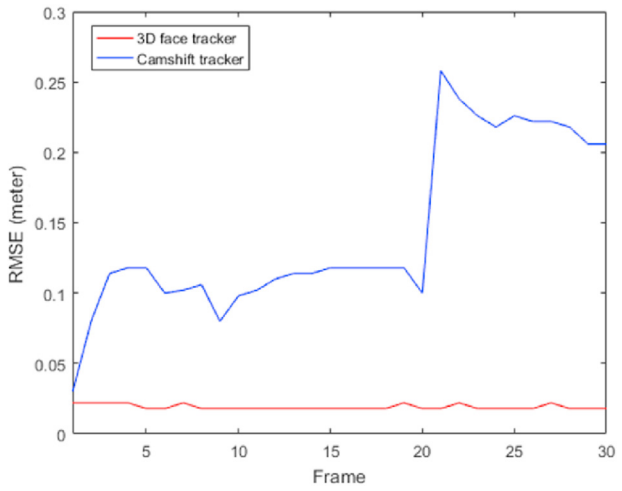

Figure 24. The RMSE result of audience A for the Camshift and 3D face tracker.

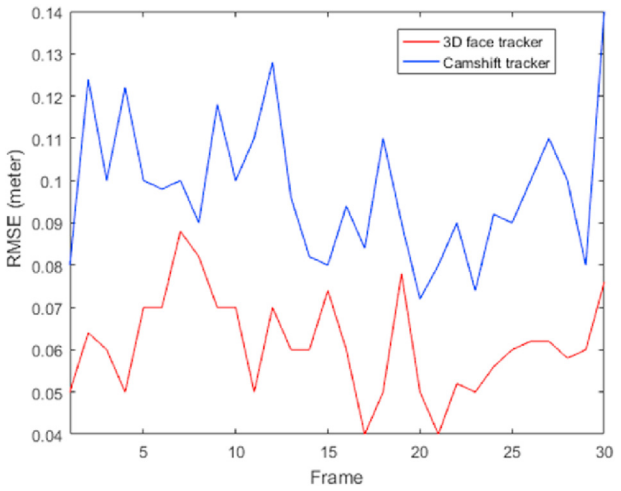

Figure 25. The RMSE result of audience B for the Camshift and 3D face tracker.

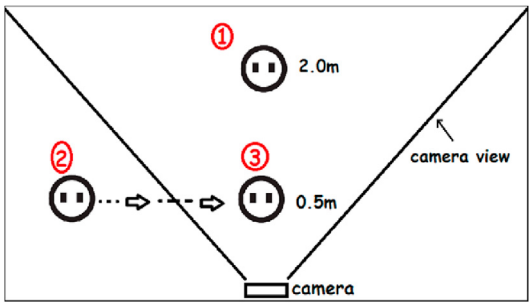

Figure 26. The procedure of Scenario 2, which indicates (1) is the audience A, (2) and (3) is audience B.

comparable with the ground truth position of the real-world environment.

Scenario 2: Passer-by occlusion

The second scenario describes the incident of the passer-by occlusion while the audience is watching the advertisement. The audience B acts as a passer-by will eventually block the face of audience A and stayed for a few seconds to watch the advertisement. Figure 26 shows the experimental procedure for Scenario 2 and steps are listed below.

1. The audience A is standing at the center of the video frame at a distance of 2.0m while watching the advertisement.
2. The audience B is walking into the ROI and stop at the center of the video frame at a distance of 0.5m to view watch the advertisement.
3. The presence of the audience B blocks audience A who is standing at a distance of 2.0m.

Figure 27 shows the sequence of video frames that are captured during the experiment. The smiley face in Figure 27 (a) displays the audience A is standing at a distance of 2.0m. While the audience A is watching the advertisement, the audience B walks into the region and appears in the video frame. Subsequently, the audience B stops in front of the audience A at 0.5m and facing to the digital display. The presence of audience B will create an obstacle for face detection and tracking algorithms to process the face of audience A.

All the detected audience is recorded and mapping into the aerial map in Figure 28. The green dot represents the audience B position and blue dot displays the audience A in the aerial map. Figure 28(a) shows the aerial map with Camshift tracker that simulates the position of audience A in the ROI, but the audience A (smiley face) is located on static position

Figure 27. The sequence of video frames captured during the experiment.

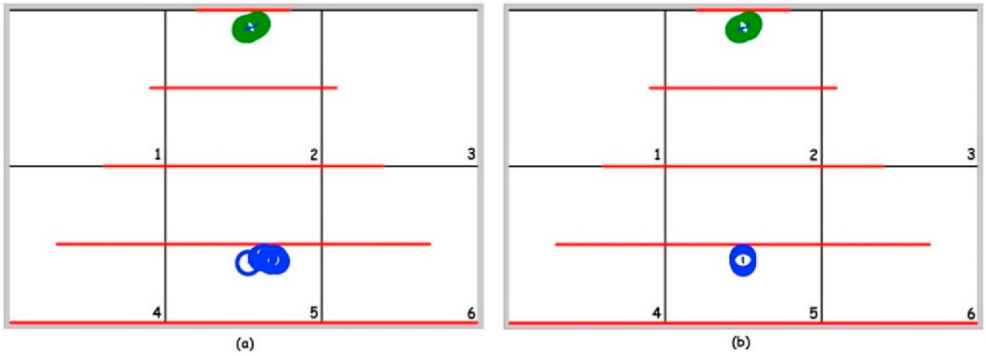

Figure 28. Aerial map result for Scenario 2 experiment (a) Camshift, and (b) 3D face tracker.

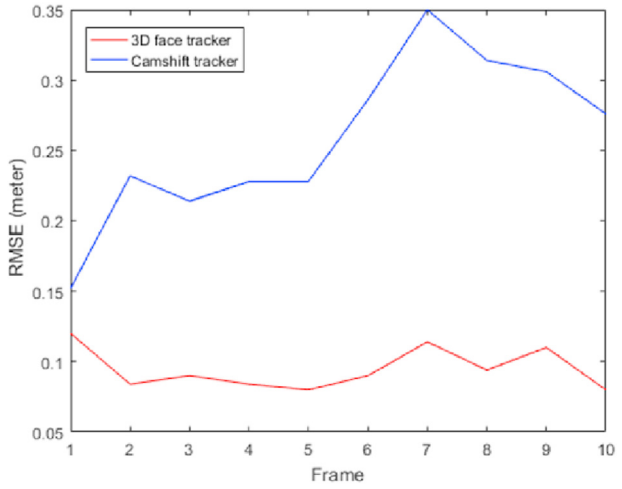

Figure 29. The RMSE result of audience A for the Camshift and 3D face tracker.

without movement in this scenario. Therefore, the error value of Camshift tracker result given by Figure 29 is slightly increased at Frame 2 onwards due to the color similarity between the background and skin of audience A. On the contrary, the aerial map in Figure 28(b) simulates the position of an audience A with a low error value of less than 0.125m as recorded in Table 4 and illustrated in Figure 29. It signifies that the simulation of audience A using the 3D face tracker is more corresponding to the ground truth position.

When audience B walks into the ROI and blocks the audience A, it creates an occlusion on the presence of audience A in the video frame. Figure 30 illustrates the result of audience B for the subsequent frames after the occlusion happens. Figure 30 shows the blue line graph of Camshift tracker is increased at Frame 18, which can be observed in the sample of the video frame shown in Figure 27 (e). The color similarity problem also happened to audience B, because the color space of the door and skin is highly correlated. Therefore, the Camshift tracker result achieves a slightly higher average RMSE value of 0.16m for audience B as recorded in Table 4.

As aforementioned, the presence of the audience B creates the occlusion for the audience A. The 3D face tracker is able to distinguish the 3D-position of the audience B from the search window of audience A. As a result, audience B is excluded from the tracker path of the audience A because the distance (z) is greater than the depth threshold of 0.25m. The 3D face tracker labels the audience B into a new independent tracker path to distinguish the detected audience. Table 4 shows that the 3D face tracker achieves an average of low RMSE value for the position of the audience as compared to the Camshift tracker. Thus, the aerial map of the 3D face tracker simulates the audience's position that is corresponding to the actual position in the real-world environment.

Lastly, the T-test statistics result is compiled in Table 5 for the comparison of the Camshift and 3D face trackers in both scenarios which are multiple adjacent audiences and passer-by occlusion. Based on the result in Table 5, the p-value of the Camshift and 3D face trackers are less than the statistical analysis critical *p*-value of 0.05. Therefore, the *p*-value shows that the null hypothesis $\boldsymbol{H_0}$ can be rejected and accept the alternative hypothesis $\boldsymbol{H_a}$. Thus, it can be concluded there is a statistically significant difference between the Camshift tracker and the proposed 3D face tracker.

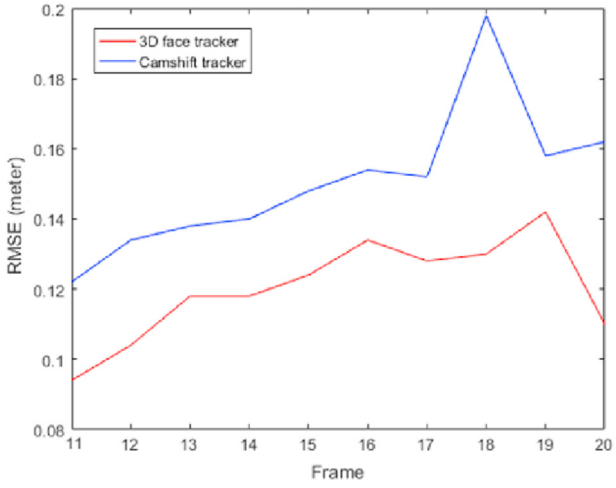

Figure 30. The RMSE result of audience B for the Camshift and 3D face tracker.

Table 4. Average RMSE (meter) result of Camshift and 3D face tracker for Scenario 2.

Audience	A	B	Average Time (ms)
Distance	2.0 m	0.5 m	
Camshift tracker	0.26	0.16	48
3D face tracker	0.09	0.11	42

Table 5. T-test result of Camshift and 3D face trackers for both scenarios.

Audience	Face Trackers	Mean (m)	Standard Deviation	*p*-value
Scenario 1 – Audience A	Camshift	0.144	0.061	4.712E-12 (0.0000000000047)
	Proposed 3D Face	0.019	0.001	
Scenario 1 – Audience B	Camshift	0.098	0.017	2.023E-12 (0.000000000002)
	Proposed 3D Face	0.061	0.012	
Scenario 2 – Audience A	Camshift	0.259	0.058	1.035E-05 (0.0000103)
	Proposed 3D Face	0.095	0.015	
Scenario 2 – Audience B	Camshift tracker	0.151	0.021	0.0002
	3D face tracker	0.12	0.145	

5. Conclusion

The proposed framework has shown its significant contribution towards handling two main challenges of the uncontrolled environment for vision-based DOOH advertising. Two scenarios, such as multiple adjacent faces and occlusion are simulated to investigate the performance of the proposed framework over another existing approach. Our proposed framework yields better performance in the aspect of accuracy and robustness to locate the audience in an uncontrolled real-world environment. The audience face centroid position (x, y) is integrated with depth information (z) to build the 3D-position (x, y, z) for 3D face tracker. The 3D face tracker utilized the 3D-position to distinguish multiple tracker paths with the detected audience in the search window. While building the aerial map, the face's centroid position will be adjusted based on the depth information. The experiment results showed that aerial map simulation with the 3D face tracker is more closely corresponding to the ground truth position as compared to the Camshift tracker. The aerial map records the audience's movement in the ROI during the advertisement broadcast, which is beneficial for the advertiser to measure the advertisement's effectiveness.

Although the static skin color model is popular for the real-time application, the performance of skin detection is still affected by the various illumination conditions in the uncontrolled environment. In future work, an adaptive skin color detection method can be added to overcome the weakness of the static skin color model to further improve the overall performance of face detection which indirectly enhances the accuracy rate.

Declarations

Author contribution statement

Ong Lee-Yeng: Conceived and designed the experiments; Analyzed and interpreted the data.

Low Chuan-Chuan: Performed the experiments; Wrote the paper.

Koo Voon-Chet: Analyzed and interpreted the data.

Leow Meng-Chew: Contributed reagents, materials, analysis tools or data.

Funding statement

This work was supported by Telekom Malaysia Berhad RDTC/191001 (MMUE/190086).

Competing interest statement

The authors declare no conflict of interest.

Additional information

No additional information is available for this paper.

References

[1] R. Ravnik, F. Solina, Interactive and audience adaptive digital signage using real-time computer vision, Int. J. Adv. Rob. Syst. 10 (2013).

[2] M. Wisotzki, K. Sandkuhl, A. Smirnov, A. Kashevnik, N. Shilov, Digital Signage and Targeted Advertisement Based on Personal Preferences and Digital Business Models, in: Conference Of Open Innovation Association, FRUCT, 2018, pp. 374–381.

[3] N. Turov, N. Shilov, N. Teslya, Digital signage personalization through analysis of the visual information about viewers, Conf. Open Innov. Assoc. Fruct 2019 (2019) 444–450.

[4] K. Mishima, T. Sakurada, Y. Hagiwara, Low-cost managed digital signage system with signage device using small-sized and low-cost information device, in: 2017 14th IEEE Annu. Consum. Commun. Netw. Conf. CCNC 2017, 2017, pp. 573–575.

[5] S. Kim, E. Park, S. Hong, Y. Cho, A.P. Pobil, Designing digital signage for better wayfinding performance new visitors ' navigating campus of university, in: 4th Int. Conf. Interact. Sci., 2008, pp. 35–40.

[6] Z. Koto, Y. Bandung, Interactive Digital Signage architecture to improve user interaction on tourism information services, in: 2016 Int. Symp. Electron. Smart Devices, ISESD 2016, 2017, pp. 380–385.

[7] S.C. Kuo, C.J. Lin, C.C. Peng, Using adaboost method for face detection and pedestrian-flow evaluation of digital signage, in: Proc. - 2014 Int. Symp. Comput. Consum. Control. IS3C 2014, 2014, pp. 90–93.

[8] W. Hyun, M.Y. Huh, S.H. Kim, S.G. Kang, Study on design and implementation of audience measurement functionalities for digital signage service using Kinect camera, Int. Conf. Adv. Commun. Technol. ICACT (2014) 597–600.

[9] K.C. Yin, H.C. Wang, D.L. Yang, J. Wu, A study on the effectiveness of digital signage advertisement, in: Proc. - 2012 Int. Symp. Comput. Consum. Control. IS3C 2012, 2012, pp. 169–172.

[10] T. Ogi, Y. Tateyama, Y. Matsuda, Push type digital signage system that displays personalized information, in: Proc. - 2014 Int. Conf. Network-Based Inf. Syst. NBiS 2014, 2014, pp. 411–415.

[11] A. Sharifara, M.S. Mohd Rahim, Y. Anisi, A general review of human face detection including a study of neural networks and Haar feature-based cascade classifier in face detection, in: Proc. - 2014 Int. Symp. Biometrics Secur. Technol. ISBAST 2014, 2015, pp. 73–78.

[12] C.T. Hsieh, H.C. Wang, Y.K. Wu, L.C. Chang, T.K. Kuo, "A Kinect-based people-flow counting system," *ISPACS 2012, IEEE Int. Symp. Intell. Signal Process. Commun. Syst.*, no. Ispacs (2012) 146–150.

[13] T.Y. Chen, C.H. Chen, D.J. Wang, Y.L. Kuo, A people counting system based on face-detection, in: Proc. - 4th Int. Conf. Genet. Evol. Comput. ICGEC 2010, 2010, pp. 699–702.

[14] X. Liu, P.H. Tu, J. Rittscher, A. Perera, N. Krahnstoever, Detecting and counting people in surveillance applications, in: IEEE Int. Conf. Adv. Video Signal Based Surveill. - Proc. AVSS 2005 2005, 2005, pp. 306–311.

[15] R. Ravnik, F. Solina, Audience measurement of digital signage: Quantitative study in real-world environment using computer vision, Interact. Comput. 25 (3) (2013) 218–228.

[16] M.N. Chaudhari, M. Deshmukh, G. Ramrakhiani, R. Parvatikar, Face detection using Viola jones algorithm and neural networks, in: Proc. - 2018 4th Int. Conf. Comput. Commun. Control Autom. ICCUBEA 2018, 2018, pp. 1–6.

[17] E. Alionte, C. Lazar, "A practical implementation of face detection by using Matlab cascade object detector," 2015, in: 19th Int. Conf. Syst. Theory, Control Comput. ICSTCC 2015 - Jt. Conf. SINTES 19, SACCS 15, SIMSIS 19, 2015, pp. 785–790.

[18] P. Viola, M. Jones, Rapid object detection using a boosted cascade of simple features, Proc. 2001 IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognition. CVPR 1 (2001) I-511–I–518, 2001.

[19] P. Viola, M. Jones, Robust real-time face detection, Int. J. Comput. Vis. 57 (2) (2004) 137–154.

[20] K. Dang, S. Sharma, Review and comparison of face detection techniques, in: 2017 7th Int. Conf. Cloud Comput. Data Sci. Eng. - Conflu., 2017, pp. 629–633.

[21] M.-H. Yang, D.J. Kriegman, N. Ahuja, "Detecting faces in image: a survey,", IEEE Trans. Pattern Anal. Mach. Intell. 24 (1) (2002) 34–58.

[22] T. Kinebuchi, H. Arai, I. Miyagawa, Image processing techniques for measuring advertising effectiveness of digital signage, NTT Tech. Rev. 7 (12) (2009) 1–6.

[23] C.E. Erdem, S. Ulukaya, A. Karaali, A.T. Erdem, "Combining Haar feature and skin color based classifiers for face detection," *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, no. March 2014 (2011) 1497–1500.

[24] S. Kolkur, D. Kalbande, P. Shimpi, C. Bapat, J. Jatakia, Human skin detection using RGB, HSV and YCbCr color models 137, 2016, pp. 324–332.
[25] O. Lucena, I.D.P. Oliveira, L. Veloso, E. Pereira, Improving face detection performance by skin detection post-processing, in: Proc. - 30th Conf. Graph. Patterns Images, SIBGRAPI 2017, 2017, pp. 300–307.
[26] N.A. bin Abdul Rahman, K.C. Wei, J. See, RGB-H-CbCr skin colour model for human face detection, in: Proc. MMU Int. Symp. Inf. Commun. Technol. (M2USIC 2006), 2006, pp. 90–96.
[27] C.C. Low, L.Y. Ong, V.C. Koo, "Experimental study on multiple face detection with depth and skin color," *ISCAIE 2019 - 2019 IEEE Symp*, Comput. Appl. Ind. Electron. (2019) 114–119.
[28] C. Xiu, X. Su, X. Pan, Improved target tracking algorithm based on Camshift, in: Proc. 30th Chinese Control Decis. Conf. CCDC 2018, 2018, pp. 4449–4454.
[29] Q. Cao, R. Liu, Real-time face tracking and replacement, 2014, pp. 1–10.
[30] J.G. Allen, R.Y.D. Xu, J.S. Jin, Object tracking using CamShift algorithm and multiple quantized feature spaces, Reproduction 36 (2006) 3–7.
[31] K.E.B. Ahmed, R.A. Saeed, R.A. Mokhtar, Real time CAMSHIFT tracking algorithm using TMS320DM6437 EVM, in: Proceedings - 2017 International Conference on Communication, Control, Computing and Electronics Engineering, ICCCCEE 2017, 2017, pp. 1–6, 1.
[32] M. Harahap, A. Manurung, Priya, A. Prakoso, M.F. Tambunan, Face tracking with camshift algorithm for detecting student movement in a class, J. Phys. Conf. Ser. 1230 (1) (2019).
[33] F.D. by Weihua Cao, Guangzhu Xu, Bangjun Lei, Panlong Yin, A Multiple Face Detection and Tracking System Based on TLD, *Icimcs*, 2013, pp. 386–389.
[34] Y. Liu, P. Wang, H. Wang, Target Tracking Algorithm Based on Deep Learning and Multi-Video Monitoring, in: 2018 5th Int. Conf. Syst. Informatics, ICSAI 2018, Icsai, 2019, pp. 440–444.
[35] D.T. Nguyen, T.D. Pham, N.R. Baek, K.R. Park, Combining deep and handcrafted image features for presentation attack detection in face recognition systems using visible-light camera sensors, Sensors 18 (3) (2018).
[36] G. Antipov, S.A. Berrani, N. Ruchaud, J.L. Dugelay, Learned vs hand-crafted features for pedestrian gender recognition, in: MM 2015 - Proc. 2015 ACM Multimed. Conf., 2015, pp. 1263–1266.
[37] P.B. Nithin, A. Francis, A.J. Chemmanam, B.A. Jose, J. Mathew, Face tracking robot testbed for performance assessment of machine learning techniques, in: 2019 7th Int. Conf. Smart Comput. Commun. ICSCC 2019, 2019, pp. 1–5.
[38] A. Pal, Multicues face detection in complex background for frontal faces, in: Proc. - IMVIP 2008, 2008 Int. Mach. Vis. Image Process. Conf., 2008, pp. 57–62.
[39] J. Kovac, P. Peer, F. Solina, "Human skin color clustering for face detection," *EUROCON 2003. Comput. as a Tool*, IEEE Reg 2 (2003) 144–148, 8, vol. 2.
[40] M.A.A. Akash, M.A.H. Akhand, N. Siddique, Robust face detection using hybrid skin color matching under different illuminations, in: 2nd Int. Conf. Electr. Comput. Commun. Eng. ECCE 2019, 2019, pp. 1–6.
[41] M.J. Taylor, T. Morris, "Adaptive skin segmentation via feature-based face detection," *Real-Time Image Video Process* 9139 (2014), 91390P, 2014.
[42] I. Corporation, "Intel ® RealSense ™ Camera: Depth Testing Methodology," *New Technol. Group*, Intel Corp., 2018, pp. 1–18.