MJSAT

Malaysian Journal of Science
and
Advanced Technology

Journal homepage: https://mjsat.com.my/

Robust, Resilient Enhanced CAMSHIFT Model: Advancing Face Detection and Tracking Stability in Challenging Environments

Alex Kok Bin See*[1], and Keith Han Wee Goh[2]

[1] *School of Engineering, Ngee Ann Polytechnic, Singapore.*
[2] *School of Electrical Engineering and Computer Science, University of Newcastle, Australia.*

KEYWORDS

Robust, Resilient CAMSHIFT
Perceptual Grouping
Connected Components
Weighted Adaptive Histogram
Selective Adaptation

ARTICLE HISTORY

Received 19 December 2023
Received in revised form 18 January 2024
Accepted 19 January 2024
Available online 27 January 2024

ABSTRACT

This paper presents the development of a robust CAMSHIFT model for theoretical face detection and tracking. The proposed model integrates innovative techniques such as Perceptual Grouping, three Connected Component Operators, Weighted Adaptive Colour Histogram, and Selective Adaptation. Experimental results highlight its superior performance across scenarios like occlusions, varying illumination, near/far face tracking, skin-like background tracking, and disturbance from multiple faces. The normalized log-likelihood index serves as a robust indicator for face tracking analysis. Connected Component operations provide strong markers for error detection in video sequences. The enhanced CAMSHIFT algorithm exhibits resilience and stability, even in the presence of occlusions. Comparisons with the original CAMSHIFT reveal the enhanced model's superiority, extending tracking range to 500 cm, a calculated enhancement of 42.9 percent improvement. The study consistently favours the robust and resilient CAMSHIFT model in tracking against skin-like backgrounds and disturbances. Despite webcam convenience in used for algorithm development, the benefits of high-performance camera systems are envisioned for future research. This model is a significant advancement in face detection methods, promising improved adaptability and tracking capabilities.

1. INTRODUCTION

Over the decades, significant advancements have been made for developing face detection and tracking algorithms in the computer vision and engineering research communities. Deep learning and Convolutional Neural Networks (CNNs) continued to play a crucial role in face detection and tracking. Improvements in model architectures and training techniques, such as the use of deeper networks and transfer learning, were likely to have been explored. Real-time performance has been a focus on improving the real-time performance of face detection algorithms, making them more efficient and suitable for applications like video surveillance, facial recognition systems, and augmented reality. Advancement in 3D face recognition have made significant progress that involves capturing and analysing facial features in three dimensions. This technology has potential applications in security and authentication systems.

Robustness to occlusion and pose variations were some highlights by researchers who were working on making face detection algorithms more robust to challenges such as occlusion (partial obstruction of the face) and pose variations, ensuring better performance in diverse real-world scenarios. In edge computing and deployment research, there has been a trend towards developing lightweight models suitable for deployment on edge devices, enabling face detection and tracking capabilities in resource-constrained environments. Privacy and ethical consideration in mind, there is an increased of research in this areas where the use of facial recognition technology for biometrics, there has been a growing emphasis on addressing privacy concerns and ethical considerations. Researchers and developers have been working on solutions to mitigate biases and protect individuals' privacy.

Robust face localization/detection and tracking has become an important task in many biometrics, video surveillance systems and has been successfully implemented using the

*Corresponding author:
E-mail address: Alex Kok Bin See <alex_see@np.edu.sg>
https://doi.org/10.56532/mjsat.v4i1.238
2785-8901/

CAMSHIFT (Continuous Adaptive Mean-Shift) model. At the time of this writing and reviewing the research articles on CAMSHIFT algorithm, as in [1], Du has presented a research paper on the use of CAMSHIFT-based tracking system with Kalman Filter as a low-cost system. He used a hybrid system involving Kalman Filter to predict the tracking window of the CAMSHIFT in the next frame. Both the original CAMSHIFT and his hybrid system was tested by LaSOT (Large-Scale Single Object Tracking) benchmark, reported elsewhere. His benchmark results show that his hybrid system features better success rate on certain categories of objects and benchmark attributes.

As in [2], Jiang et al reported and discussed about a remote sensing image target recognition system specifically designed for tennis sports in China. It utilizes the CAMSHIFT (Continuously Adaptive Mean Shift) algorithm as a central component for moving target detection and tracking. The focus is on image processing and recognition within the context of tennis-related scenarios. Kalman filtering was reported and used as a prediction mechanism to estimate the approximate range of the moving target in the next frame of the image. Another application of motion detection and tracking research paper on vehicle flow was presented by Sun et al as in [3]. They presented the developed vehicle flow statistics system in video surveillance based on CAMSHIFT and Kalman Filter. The system utilised traffic pavement camera system to detect traffic flow in a specific period. The system calculates traffic flow changes and provides intuitive feedback to users. Further this paper discusses the combination of CAMSHIFT and Kalman filter algorithms for multi-target tracking and proposes an improved algorithm for vehicle tracking. The paper concludes that the proposed system is efficient and suitable for surveillance system design, and can overcome the challenges of moving object detection, tracking, and tracking in occlusion. As in [4], Wang et al reported their research paper on CAMSHIFT algorithm based on feature matching and prediction mechanism. Their paper also discusses the shortcomings of the mean shift algorithm and how the CAMSHIFT algorithm was created to resolve them. The CAMSHIFT algorithm uses the mean shift object tracking algorithm for each frame of the video channel, and the tracking result of each image determines the initial value of the tracking iteration in the next frame. Furthermore, this paper reported target feature matching algorithm based on Oriented FAST and Rotated BRIEF (ORB). It was developed to address some limitations of existing feature detection algorithms like SIFT (Scale-Invariant Feature Transform) and SURF (Speeded-Up Robust Features). Further details of this algorithm are reported elsewhere. One of the strengths of ORB is its computational efficiency. The binary nature of the descriptors allows for fast matching, and the algorithm is designed to be computationally lightweight compared to some alternatives. ORB is particularly useful in real-time applications such as object recognition, image stitching, and augmented reality due to its speed and reliability in various conditions. It strikes a balance between accuracy and computational efficiency, making it suitable for resource-constrained environments. This paper concludes that the improved algorithm can more accurately detect the position and size of the moving object compared to the traditional CAMSHIFT algorithm.

In [5], Nguyen et al presented a paper real-time target human tracking using CAMSHIFT and LucasKanade optical flow algorithm. In their approach, the authors combined the CAMSHIFT algorithm and Lucas-Kanade Optical Flow Algorithm (LK-OFA) to create a more accurate and efficient tracking system. The Lucas-Kanade Optical Flow Algorithm uses ORB features to track reliable keypoints, reducing the effects of illumination or displacement on human targets. The CAMSHIFT algorithm is used to determine the area of the human target in the frame more precisely to aid and maintain the tracking accuracy even in cases of similar objects being detected. The Kalman Filter is used to predict the state of the human target, improving the tracking performance. The proposed method has several advantages, including rapid calculations in implementation, high accuracy in cases of similar objects detection, and the ability to deploy easily on mobile devices. The effectiveness of the proposed tracking algorithm is demonstrated through experimental results, which show certain level of accuracy and robustness under various scenarios. The study's findings and the proposed tracking system's performance may be limited to the specific conditions and environments in which it was tested, and its generalizability to other settings or scenarios may be a potential limitation.

As in [6], the researchers, Guo et al reported on a target tracking system for an amphibious robot based on improved CAMSHIFT algorithm. The paper discusses the development of an amphibious robot for various applications, including navigation and terrain tracking. Further, the proposed robot was designed to be modular and capable of moving in both especially marine water and on land environments. The proposed improved CAMSHIFT algorithm for target tracking in amphibious environments, included features using ORB, Kalman filter, morphological processing, neural network and improved matching scores was used to track targets in complex environments, such as those with background interference and occlusion. Neural network was utilised for target image recognition from a host board using Raspberry PI, which communicates with another microcontroller, STM32F4 with a much lower computing power. The STM32F4 microcontroller resides inside the amphibious robot performs the target tracking and motion decision processing on the collected target information. Although this paper did not explicitly mention about the research's limitations, however, their experimental results did not suggest authentic and fair comparison for target tracking and recognition between original versus the claimed enhancement CAMSHIFT algorithm.

In [7], another Chinese research group, Hu and Huang presented face detection based on SSD and CAMSHIFT algorithm. They proposed a face detection system that combines the Single Shot MultiBox Detector (SSD) algorithm and the CAMSHIFT tracking algorithm. The SSD algorithm is a fast and accurate deep learning object-detection method that predicts the object and bounding box in the same architecture. It is used to detect faces in the proposed masked-face dataset. This is particularly useful for applications such as fatigue driving detection. In this work, however, prior training is required for the SSD model with drivers' face images were collected for the construction of the model. A two-step approach was seen in this research, where improved SSD initially detects the face area followed by transferring the information to CAMSHIFT with Kalman filtering. Similar to the earlier research paper, this paper did not explicitly mention about the research's limitations, however, their experimental results did not suggest authentic and fair comparison for target tracking and recognition between original versus the claimed enhancement CAMSHIFT algorithm. In [8], this paper discusses the challenges of detecting and tracking human motion targets in video images and proposes a connected region search method to select the appropriate moving target and calibrate the target area. The CAMSHIFT algorithm is used to track the human motion targets in video images, and the paper demonstrates the effectiveness of the proposed method through experimental results, which showed high accuracy and robustness in various scenarios. The paper contributes to the development of efficient and accurate human motion target tracking systems, which have applications in surveillance, robotics, and human-computer interaction.

Roy et al [9] reported in their paper entitled "An Efficient Sign Language Recognition (SLR) System Using CAMSHIFT Tracker and Hidden Markov Model (HMM)". The authors proposed an end-to-end SLR system for American Sign Language (ASL) gestures. The system consists of four modules: hand and face detection, hand tracking, feature extraction, and gesture recognition. The authors used skin colour segmentation to detect hands, CAMSHIFT tracker to track hand motion, and Hidden Markov Model (HMM) based sequence classification to recognize gestures. Further this paper also proposes a novel approach to differentiate between double hand and single hand gestures, and new features from skin region and hand trajectories to improve gesture classification performance. The system is tested on the dataset proposed by American Sign Language Linguistic Research Project (ASLLRP), which consists of isolated signs, and the experiment results are encouraging.

In [10], Zhang and Zhang reported in their paper "Optimization of Face Tracking Based on KCF and CAMSHIFT" discusses the improvement of face tracking algorithms by combining the KCF (Kernel Correlation Filter) and CAMSHIFT algorithms. The KCF algorithm is a method for visual tracking that utilizes a linear kernel to learn a correlation filter for locating the target in the subsequent frames of a video. It is known for its high-speed performance and accuracy in tracking various objects, including faces. The algorithm works by learning the correlation filter from the training samples of the target object and then using it to detect the object in the subsequent frames of the video. The KCF algorithm has been widely used in visual tracking applications due to its efficiency and effectiveness In summary, the paper presents a method to optimize face tracking through the combination of the KCF and CAMSHIFT algorithms, resulting in improved tracking accuracy and reduced failure rates compared to the KCF algorithm alone. The detail mathematical treatment of the KCF algorithm and CAMSHIFT method are both reported in that paper, and no further explanation is required for this work here.

Human skin has been utilized and proven to be an effective feature in certain applications from face detection to hand tracking. Although different races of people have different skin colors, for example Asians, Africans, Caucasians etc, several studies have shown that the major difference lies largely between their intensity rather than their chrominance. Many color spaces such as RGB, normalized RGB, HSV, YCrCb etc, are being used to label pixels as skin.

In the seminal work originally reported by Bradski [11] on CAMSHIFT, human faces are tracked by projecting the face color distribution model onto the colour frame and moving the search window to the mode (peak) of the probability distributions by climbing density gradients. The author of this original work developed this novel algorithm based on a robust non-parametric technique for climbing density gradients to find the mode (peak) of probability distributions called mean shift algorithm. Bradski modified the mean shift algorithm to deal with dynamically changing colour probability distributions derived from video frame sequences. This modified algorithm is called Continuously Adaptive Mean Shift (CAMSHIFT) algorithm. Hue Saturation Value (HSV) colour system is used to correspond to projecting standard Red, Green, Blue (RGB) colour space along its principal diagonal from white to black Tracking of non-rigid objects is done through finding the most probable target position by minimizing the metric based on Bhattacharyya coefficient between the target model and the target candidates as in [12]. Bhattacharyya coefficient is a popular method that colour histogram to correlate images.

The field of computer vision dealing with face detection, tracking algorithm development and analysis presents researchers with intricate challenges arising from the complex nature of these diverse environmental conditions. Conditions such as the following:

- Occlusions
- Tracking under varying illumination environment
- Tracking of a near versus far face
- Tracking with skin-like background or object
- Disturbance from multiple faces

This paper presents the discussion on face detection, tracking and localization in video images using a proposed robust & resilient enhanced CAMSHIFT model that we have developed. The improvised model must be able to track face efficiently with very little error such as occlusion. Since skin colour model is being utilized in implementing this algorithm, hand occlusion (i.e. hand covering the whole face) will prove to be a challenge. The algorithm will also further enhance the technique to obtain a better performance and stability in face localization and tracking system. Several experimental tests will be conducted to fairly compare its robustness and resilient with Bradski's original CAMSHIFT algorithm.

In this paper, the key motivation is to ensure that the proposed robust & resilient enhanced CAMSHIFT algorithm is presented with clarity to the reader. Mathematical treatment of the theory of the algorithms are presented. Experimental data and analysis are presented and discussed in this paper.

After initial literature scan, the authors postulated that an extensively developed robust & resilient CAMSHIFT model's capability with experimental results has not been reported in the literature to date, and at the time of this writing. It is believed that this proposed technique and implementation through the combined Perceptual Grouping method, three Connected Component Operators, application of Weighted adaptive colour histogram and Selective Adaptation technique have not been reported anywhere. Little is known about the specific effects combining these methodologies prompting the need for further investigation. Research setup and experimental results presented here will contribute to the body of knowledge to the existing research related to face detection and tracking of human subjects. Needless to say, that face detection and tracking has enormous potential applications in video surveillance, security, biometrics and many others. The remaining paper is structured as follows: with Section 2 on literature review, Section 3 describes the methodology used, Section 4 focuses on theoretical analysis and implementation. Section 5 presents the results and discussion, Section 6 concludes the paper.

2. Literature Review

There are many face detection, localisation and tracking techniques widely used by researchers from all over the world. In this section, the author will attempt to review some of the relevant journals and conference papers which cover mainly on these techniques. The attempt to review by the authors is by no means exhaustive in nature. As in [13], a review paper entitled "A review of recent advances methodologies for face Detection" published in the International Journal of Current Engineering and Technology in 2023 provides a good review discussion about recent advances in face detection methodologies. The paper discusses various face detection algorithms, including statistical and neural network approaches, and their advantages, limitations, and use in fields other than face detection. The authors conclude that there is no perfect algorithm for face detection, but the comparative comparisons in the paper will help to choose the algorithm. Further scanning of the available literature in [14], another review paper "Human Face Detection Techniques: A comprehensive review and future research directions" published in Electronics in 2021 provided a

face detection algorithms can work well in real-time applications, but their performance depends on the algorithm used, image quality, and hardware capabilities. These algorithms use mathematical and machine learning techniques to identify facial features in images or videos. For example, the Viola-Jones algorithm is popular for real-time face detection, but it may struggle with covered or improperly oriented faces. Deep learning-based algorithms, such as Multi-Task Cascaded Convolutional Neural Networks (MTCNN), have shown impressive speed in real-time face detection, but they can be affected by low-resolution images or image noise. Further research is needed to enhance the accuracy and robustness of face detection algorithms.

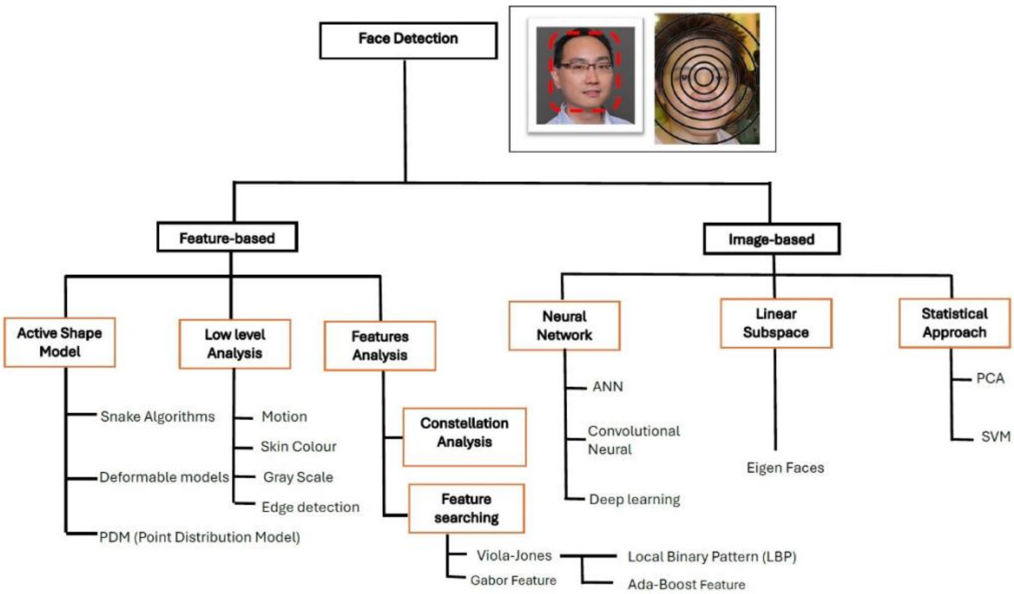

Fig. 1. depicts the general classification of face detection techniques based on literature review

comprehensive review of human face detection techniques. The paper explores different face detection algorithms in five parts, including history, working principle, advantages, limitations, and use in fields other than face detection. The authors review many face detection algorithms, such as different statistical and neural network approaches, which were neglected in earlier literature but gained popularity recently because of hardware advancements. The limitations of current face detection algorithms include:

- High variability of human faces: Human faces vary in terms of shape, size, pose, expression, illumination, occlusion, and makeup, making it challenging for algorithms to generalize and cope with these variations
- Poor image quality: The effectiveness of face detection algorithms is influenced by the quality of the images. Low-quality images can limit the performance of the algorithms.
- Different face angles: The relative angle of the target's face can significantly impact the recognition score, and algorithms may struggle with faces at angles other than frontal view.
- Partial occlusion and distortions: Face detection can fail when the face is partially hidden by objects like masks, hats, or hands, or when the face is tilted, rotated, or distorted, leading to reduced performance.

As in [15], the authors, Kumar et al reported a review paper entitled “Face detection techniques: a review”. The research paper provided a comprehensive review of face detection techniques, covering various approaches and their underlying principles. It discusses the use of mathematical algorithms, such as the Viola-Jones algorithm and deep learning-based methods, for real-time face detection. Furthermore, the paper also explores the challenges and limitations associated with face detection algorithms, including issues related to image quality, face variability, and different face angles. Additionally, it delves into the comparison of feature-based and image-based approaches for detecting faces in digital images, highlighting the strengths and weaknesses of each method. Furthermore, the paper examines the use of colour models and statistical shape models for face detection. Overall, it offers a detailed overview of the advancements, challenges, and future research directions in the field of face detection. Figure 1, depicts a general classification of face detection techniques summarised based on literature review.

The authors Alqahtani et al [16] reported a review paper entitled "3D face tracking using stereo cameras: A review". Applications of 3D face-tracking systems with a focus on stereo camera-based systems was postulated. Stereo cameras are less expensive than laser ranging systems, and they are widely available on devices such as smart phones and Microsoft Kinect 3D camera mainly for Xbox 360 game console. Further in that paper, the authors presented and summarised the five main techniques/methods for face tracking algorithms, namely as follows:

- Kanade-Lucas-Tomasi (KLT) method
- Particle filtering technique
- Trace learning detection (TLD)
- Probability hypothesis density (PHD) filtering technique
- Mean-Shift/CAMSHIFT methods

Numerous investigations into face tracking have employed methodologies like the Kanade-Lucas-Tomasi approach, particle filters, tracking-learning-detecting, probability hypothesis density, and mean shift/cam shift, among others. As constraints on imaging are eased, the complexity of facial tracking intensifies. From Alqahtani et al's review paper, they discussed and presented an overview outlines prevalent challenges encountered in face tracking, encompassing issues such as occlusion and clutter, variations in pose, alterations in facial resolution, fluctuations in illumination, and facial deformation. Detailed treatment of each technique including advantages, disadvantages and limitations of each method/technique has already been reported by authors Alqahtani et al in that paper, beyond the scope of this research paper, presented by the current authors. The authors of this paper will focus on the Mean-Shift technique but even more specifically at primarily on CAMSHIFT method for face detection and tracking.

The Mean-Shift plays a vital role in both image processing and cluster analysis within the realm of computer vision, specifically homing in on the peaks of a designated density function. Initially, Mean-Shift method was conceptualized by Fukunaga and Hostetler [17], this method or technique has solidified its standing as a noteworthy approach in the expansive domain of computer vision, and it has led to numerous developments and adaptations thereafter. In the sphere of clustering predicaments, the mean shift method operates on the premise that each given point denotes samples linked to a probability density function. In this scenario, regions exhibiting a heightened density function are construed as aligning with the local maxima of the distribution. The algorithm facilitates points attracting one another, mimicking a gravitational force of minimal magnitude. Consequently, points converge toward regions boasting greater density, leading to their amalgamation at diverse points and unveiling the local maxima situated in the vicinity of these convergent zones. The pivotal aspect of identifying local maxima lies in achieving optimal resolution, particularly in the context of 3D face tracking, as it facilitates the maximum dispersion of pixels. To encapsulate, the mean shift algorithm adeptly pinpoints local maxima by shifting a window toward zones boasting the highest density. Characterized as a "hill climbing algorithm", mean shift entails an iterative process of shifting kernels toward regions of augmented density until reaching convergence, encapsulating its role in the intricacies of clustering.

3. METHODOLOGY

In this section, the authors will provide an overview with an emphasis on the proposed and developed robust and resilient CAMSHIFT algorithm. The overall of this proposed CAMSHIFT face detection and tracking algorithm is divided into six (6) stages/processes, which will be described and discussed here. The hardware and computer in used for this work are described in subsection 3.1.

3.1 Hardware and Computer Specification usage

The convenient use of a webcam and computer were utilized for this purpose of this research project mainly to focus on the software algorithm development, data collection and analysis.

- Creative Live! WebCam, 640x480 VGA CCD sensors
- 640x480 resolution video up to 30 FPS(frames/sec)
- USB 2.0 Hi-Speed connectivity
- 4X digital Zoom
- Computer equipped with Intel Pentium M processor 1.73GHz and 1GB DDR2 RAM

3.2 Development Software

The development software and all the coding were developed using LabVIEW, C programming and National Instruments NI-IMAQ. The developed software codes are upward compatible with the latest version of the LabVIEW and NI-IMAQ toolkit.

One of the developed LabVIEW analysis software front panels is depicted in Figure 2. This front panel is shown in partial because there are several testing and data analysis programs which are known as LabVIEW subVIs or in layman's terms known as subroutines. The front panel has programming features where the user can interact with the software program through a dialogue box. The user has the option to load the video files (e.g. .AVI or other video formats) from the system file directory on the computer. User has the option to pause the video to analyse frame by frame sequence. The CAMSHIFT window, bounding rectangles can be seen overlaid onto the video frame. The probability distribution (PD) images are shown in Figure 2 as dark colour picture of the human subject. Figure 3 depicts one of the LabVIEW block diagrams for detecting and analysing the histogram (i.e. Hue, Saturation and Value).

Certain coding was performed in C language, however, the LabVIEW formulae node was invoked and used. The C codes were embedded into the LabVIEW formulae node to ensure that the developed software program was executing efficiently. As an example, the algorithm on selective adaptation and log likelihood was coded inside the LabVIEW formulae node as depicted in Figure 4.

3.3 Overview of Six (6) Stages of the Robust and Resilient Enhanced CAMSHIFT Algorithm

The entire process is divided into six main stages/process. With reference to the condensed version of the flow chart of this detection and tracking algorithm as depicted in Figure 5. The first stage requires acquisition of images with conversion of its colour space. The next stage ensures initializing of search window and obtaining the colour histogram. The third stage performs segmentation of skin and usual of perceptual grouping technique. This follows by application of connected component operators. Fifth stage focuses on the original CAMSHIFT algorithm. Finally, Weighted Adaptive Colour Histogram with Selective Adaptation is utilized. Each stage will be described in detail over the next few sections found in this paper.

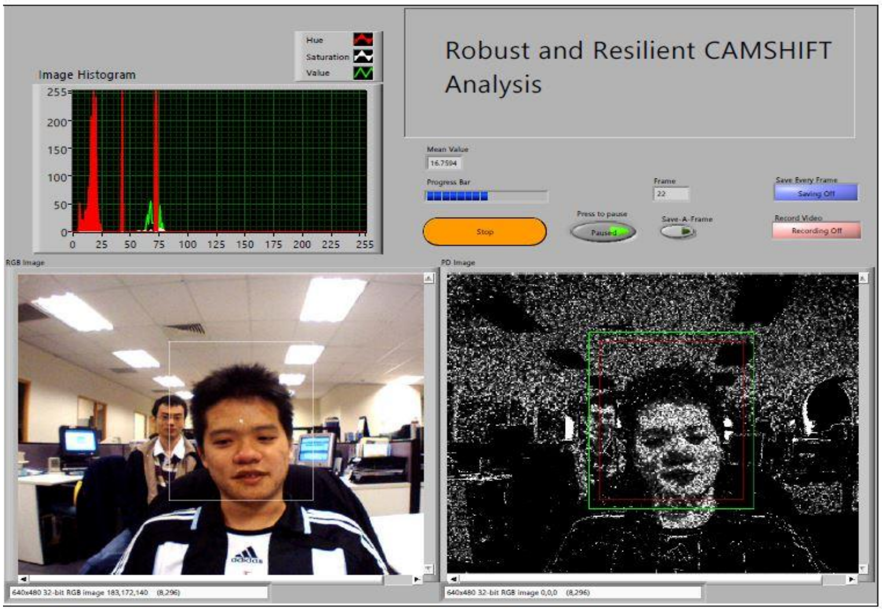

Fig. 2. Depicts one of the LabVIEW software front panels.

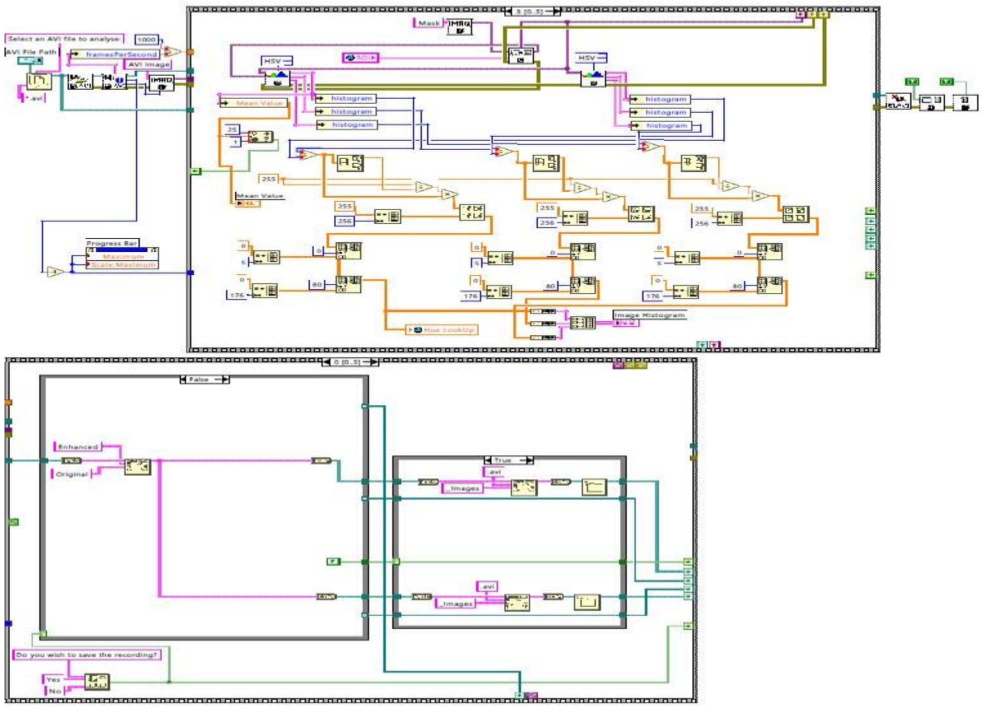

Fig. 3. Depicts a sample code of one of LabVIEW block diagrams

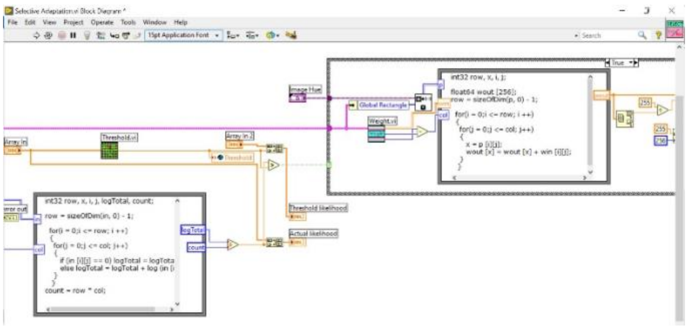

Fig. 4. Depicts LabVIEW coding with the use of formulae node, where selective adaption algorithm for log likehood was coded.

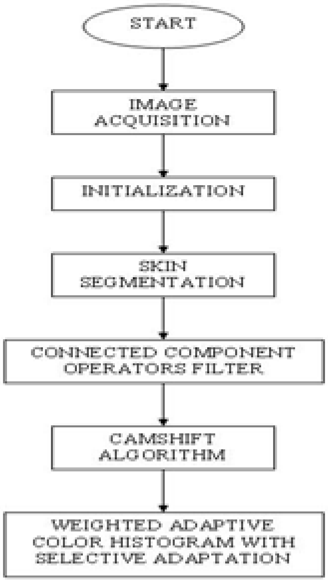

Fig. 5. Depicts a condensed version of the flow chart of the detection and tracking for this robust and resilient CAMSHIFT algorithm

3.4 Image Acquisition Stage

Most color cameras provide an RGB (Red, Green, Blue) signal [18]. It is found that RGB color space will be unreliable under changing illumination conditions. The pixels of a face image form a distribution in this color space which can be modeled by estimating a probability density function. However, intensity is distributed across the RGB values so that a face's distribution varies with scene brightness [19]. To achieve invariance under illumination conditions, the captured RGB signal has to be converted into HSL color space. HSL color space consists of hue, saturation and luminance. Human's face has a highly curved surface, the observed intensity of a face exhibits strong variation [20]. Hence it explains that why mostly face tracking application eliminates or ignores the luminance plane of color space. In other words, segmentation of skin color regions become robust only if the chrominance component is used in analysis. Chrominance colour helps to improve the tracking efficiency. And since chrominance of HSL color model can be considered as the hue and saturation, a distribution in 2-D HS space will provide a color model that is invariant to illuminations.

3.5 Initialization Stage

The face detection algorithm introduced in this project is mainly based on skin color. There is a need to obtain the skin sample of the user to be tracked. Two small windows, overlaying on the screen, are created for this purpose. One of them is known as the search window, which indicates the region of the tracked face. The other one, which is slightly larger in size than the search window is solely used for computation or calculation purposes only. The reason is to provide an accurate calculation of the color histogram. The search window will contain the detected face and track it accordingly. The author has proceeded to perform coding and testing of this stage. Technique for auto detection of user face may be considered as a future enhancement. The limitation of auto detection at start will need to be addressed for future work.

3.6 Skin Segmentation Stage

Hue derived are sampled and binned into a 1-D histogram. The skin color histogram is used to form a lookup table for the separating the skin color pixel values and the non-skin color pixel values. This lookup table consists of 256 elements with corresponding pixel value used for segmentation. Non-skin like pixel will be replaced with pixel value 0 based on the lookup table. A probability distribution image will be generated as a result of this skin segmentation. The result is a binary mask that marks the skin color area in that generated image [21]. Unfortunately, the probability distribution image generated may consist of background noise such as pixel form by skin-like objects. Furthermore, the CAMSHIFT algorithm proposed by Bradski et al relies on colour probability distribution image alone. There will be error in the tracking whereby illumination condition is too bright or dim. Perceptual grouping is introduced to filter the image. It is a process whereby a vision system organizes image regions into emergent boundary structures that aim to separate objects and scene background.

3.7 Connected Component Operators Filter Stage

It is insufficient to track a face based on skin color alone as postulated by Wang et al [22]. Relying solely on skin color for face tracking proves inadequate. This is due to the potential interference caused by skin-like pixels, for instance, the user's hand, which can lead to occlusion during the tracking process. Failing to identify such errors may result in the color model adjusting to image areas unrelated to the actual object. Consequently, the tracker is at risk of losing track of the intended object. A group of shape-oriented image processing connected component operators namely, Compactness, Solidity, and Orientation were employed on the remaining components to determine their representation as a face or otherwise. Threshold values for different connected operators are computed using a database comprising face and non-face components. Components falling below this specified threshold are excluded as non-face images, while those meeting the criteria are retained for subsequent analysis.

3.8 CAMSHIFT Algorithm Stage

Mean-Shift algorithm was first applied to the problem of mode seeking by Cheng [23]. In contrast to Mean-Shift, which caters to static distributions, CAMSHIFT (Continuously Adaptive Mean Shift) is tailored for distributions that undergo dynamic changes. The mean location, also referred to as the Centroid, is determined within the search window of the discrete probability image calculated in the earlier stage using moments. The Zeroth, First, and Second moments, all ascertainable within the Min-Max box, contribute to computing the Centroid of the search window. The Mean-Shift element involves iteratively recalculating new values for the window position derived in the previous frame until there is no substantial shift in position.

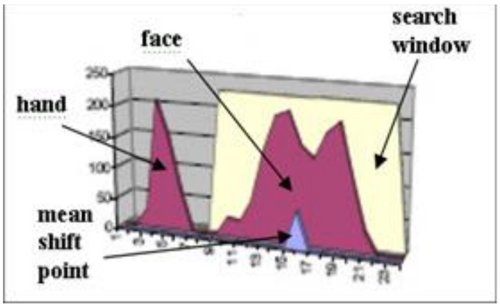

Fig. 6. Depicts an illustration on the adaptation of CAMSHIFT with dynamically changing Centroid position.

3.9 Weighted Adaptive Colour Histogram with Selective Adaptation Stage

The Mean-Shift algorithm, on its own, lacks effectiveness as a tracker. If the initially chosen region includes pixels beyond the facial area (such as background pixels), the 2-D probability distribution image can be influenced by their prevalence in the histogram back-projection. Pixels situated close to the boundary of the search window are typically less reliable and, therefore, should be excluded. Consequently, lower weight can be assigned to pixels farther from the center of the search window, with the center having the highest weights compared to more distant pixels as described by one of the current authors, See and Liaw [24]. In the earlier research work by one of the current authors, he has proposed and developed a weighted distribution or assignment mechanism that prioritize higher weighting for pixels closer to the region center, a weighted histogram may be employed to calculate the target histogram as postulated by Comaniciu et al [25]. Figure 7 illustrates the weight distribution within a search window of a captured image.

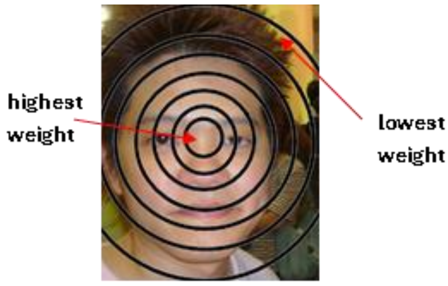

Fig. 7. Depicts the weight distribution of the human face within the search window.

A static skin color model is susceptible to significant impact from changes in illumination, leading to a loss of tracking. To address this issue, adapting skin color model becomes a crucial step in dealing with varying lighting conditions. However, a notable challenge when adjusting a color model during tracking is the absence of a reliable ground-truth. Any color-based tracker runs the risk of losing the object, particularly when it's occluded by other elements. Without error detection, the system may erroneously adapt to image regions that do not correspond to the intended target.

To alleviate this problem, the use of observed log-likelihood measurements can identify frames with errors. Colour data from these frames is excluded from adapting the object's color model. When the tracker loses the object, there is often a sudden, substantial drop in its value. Adaptation is then suspended until the object is successfully tracked with a sufficiently high likelihood. Selective adaptation is vital and it serves as the decision-maker determining whether the new skin model is allowed to adapt or not. Figure 8 depicts the detailed flowcharts with the combination of all the six (6) stages of this proposed, developed and implemented robust and resilient CAMSHIFT model.

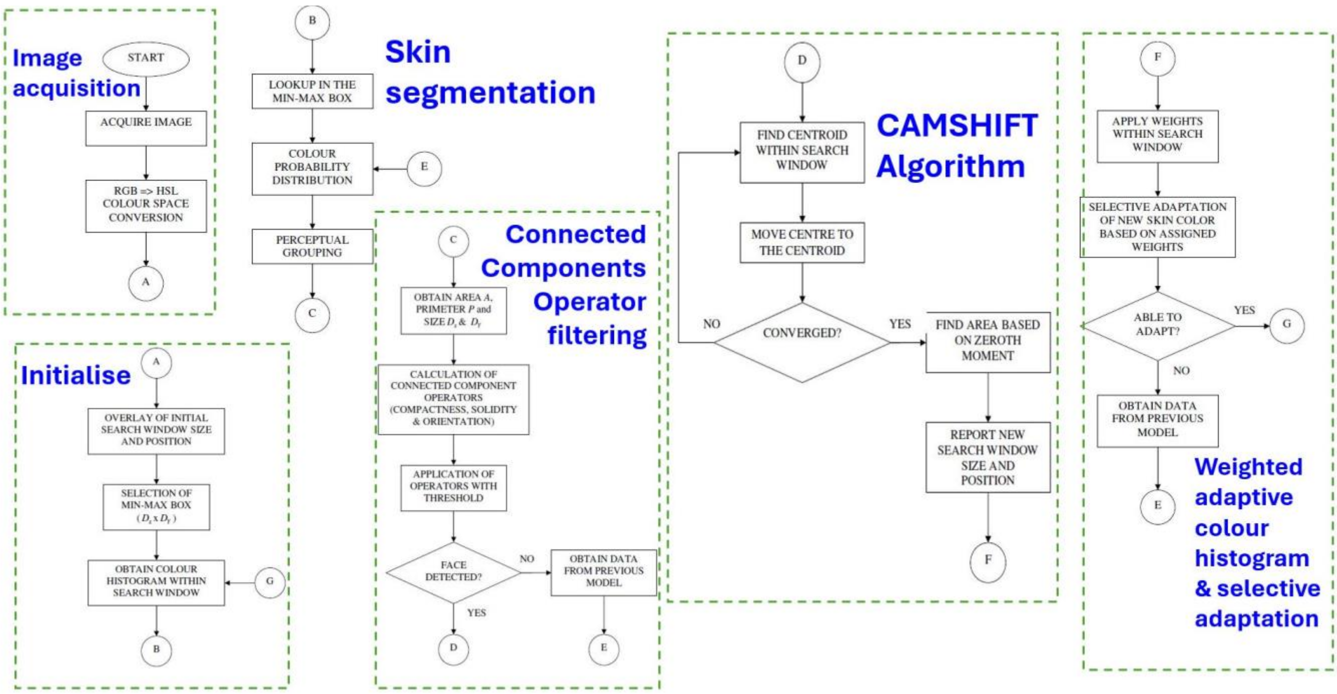

Fig. 8. Depicts the detailed flowcharts with the combination of all the six (6) stages of this proposed, developed and implemented robust and resilient CAMSHIFT model.

4. THEORETICAL ANALYSIS AND IMPLEMENTATION

The mathematical formulation and theoretical analysis of each individual method used in each stage are presented in this chapter. Firstly, the authors will discuss on how the colour Probability Distribution is formed. This is followed by the introduction of Perceptual Grouping, which is effective in skin segmentation. Next, the calculations of the three Connected Component Operators will be shown and described. The basic Mean-Shift algorithm will also be presented detailing the mathematical manipulation. With the basic algorithm explained, the steps to apply Continuously Adaptive Mean Shift Algorithm (CAMSHIFT) operation are derived. Lastly, the application of weighted adaptive colour histogram and selective adaptation are demonstrated clearly in the following.

When the program is executed, the user will be prompted three options are displayed on the screen. The first option indicates the real-time tracking using the enhanced algorithm. The user will then be given a choice whether to record the video for further analysis. With a countdown timer of 10 seconds before actual tracking begins, the skin sample will be obtained in the initial window. The search window is set at 40 x 40 pixels. The second and third option is to load and playback an AVI video for Enhanced or Original Algorithm analysis.

4.1 Colour Probability Distribution

Probability of an object of interest can be computed based on Baye's rule [26]. Considering each of the pixel and a single measurement vector m_k:

$$\mathrm{p}(\mathrm{o_n}|\mathrm{m_k}) = \frac{\mathrm{p}(\mathrm{m_k}|\mathrm{o_n})\mathrm{p}(\mathrm{o_n})}{\sum_\mathrm{i}\mathrm{p}(\mathrm{m_k}|\mathrm{o_i})\mathrm{p}(\mathrm{o_i})} \quad (1)$$

The above equation converts the frame of each incoming image into probability distribution image. The measurement vector m_k is considered as Hue vector of the image plane in this case. Conditional probability of the desired object occurs when the Hue translation has already occurred. In other words, colour plane of an image has already been converted into probability of Hue color plane, provided the Hue of that pixel has been detected first. Equation 1 is used to determine the probability of each pixel that considered as the desired object. This equation can be analogized by considering generated histogram of skin and the entire image.

The probability of a colour vector (Hue, Saturation), or simply (h, s), for a given skin is approximated by:

$$p(h,s|skin) \approx \frac{h_{skin}(h,s)}{N_{skin}} \quad (2)$$

where $h_{skin}(h,s)$ is the histogram of skin of Hue and Saturation channel, and $N_{skin}$is the total number of skin pixel.

The probability of a skin pixel in an image can be approximated by a fraction of observed pixels known to be skin, as shown below:

$$p(skin) \approx \frac{N_{skin}}{N_{total}} \quad (3)$$

where N_{total} is the total number of pixel of the image

The probability of a colour vector is approximated by:

$$p(h,s) \approx \frac{h_{total}(h,s)}{N_{total}} \quad (4)$$

Based on Baye's rule, the probability of skin given a colour vector can be found by:

$$p(skin|h,s) = \frac{p(h,s|skin)\cdot p(skin)}{p(h,s)} \quad (5)$$

Equation 5 is similar to 1, can be further simplified to the ratio of the two histograms:

$$p(skin|h,s) \approx \frac{h_{skin}(h,s)}{h_{total}(h,s)} \quad (6)$$

The result computed above forms equation 6, which is used as the lookup table to transform each pixel of every image frame into probability distribution (PD) image. It is suggested that more complicated measure sets for more object discrimination can be considered elsewhere as suggested by authors, Fukunaga and Hostetler earlier in the reference.

The probability of many local measurement vectors over a region of the image can be found by:

$$p(o_n|\Lambda_k m_k) = \frac{\prod_k p(m_k|o_n)p(o_n)}{\sum_i \prod_k p(m_k|o_i)p(o_i)} \quad (7)$$

Note that equation 7 is not applied in the proposed algorithm. The reason is because equation 1 or equation 5 is sufficient in producing the probability distribution image.

To enable face tracking using a flesh color model, users are instructed to center their face within an onscreen box, facilitating the sampling of flesh areas. The search window is set at 40 x 40 pixels, as depicted in Figure 9 and 10 for the RGB colour plane and transformed Hue image plane respectively. The hues obtained from the flesh pixels in the image are then sampled from the H channel and organized into a one-dimensional (1D) array histogram. This simplifies both computational and space complexities, facilitating the clustering of similar color values as postulated by Allen et al [27]. Histogram back-projection is a primitive operation that associates the pixel values in the image with the value of the corresponding histogram bin. The back-projection of the target histogram with any consecutive frame generates a probability distribution image where the value of each pixel characterizes probability that the input pixel belongs to the histogram that was used.

Given that m-bin histograms are used, we define the n image pixel locations $\{x_i\}_{i=1\ldots n}$ and the histogram $\{\hat{q}_u\}_{u=1\ldots m}$. The author also defines a function $c\colon \Re^2 \rightarrow \{1\ldots m\}$ that associates to the pixel at location x_i^* the histogram bin index $c(x_i^*)$. The unweighted histogram is computed based on Allen et al as:

$$\hat{q}_u = \sum_{i=1}^{n} \delta[c(x_i^*) - u] \quad (8)$$

In all cases the histogram bin values are scaled to be within the discrete pixel range of the two-dimensional array (2D) probability distribution image as follows:

$$\left\{\hat{p}_u = min\left(\frac{255}{max(\hat{q})}\hat{q}_u, 255\right)\right\}_{u=1\ldots m} \quad (9)$$

In other words, the histogram bin values are rescaled from $[0, max(\,q)]$ to the new range [0, 255], where pixels with the highest probability of being in the sample histogram will map as visible intensities in the 2D histogram back-projection image.

Fig. 9. Depicts the initial startup with small search window size of 40 x 40 pixels.

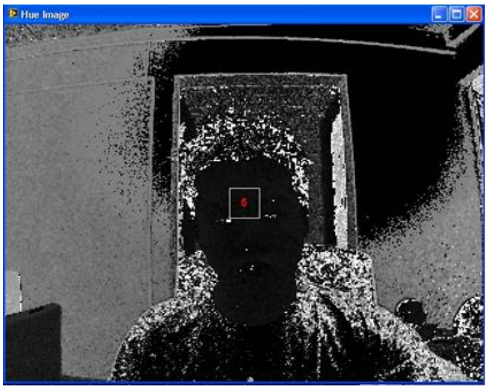

Fig. 10. Depicts the image transformation in the hue plane.

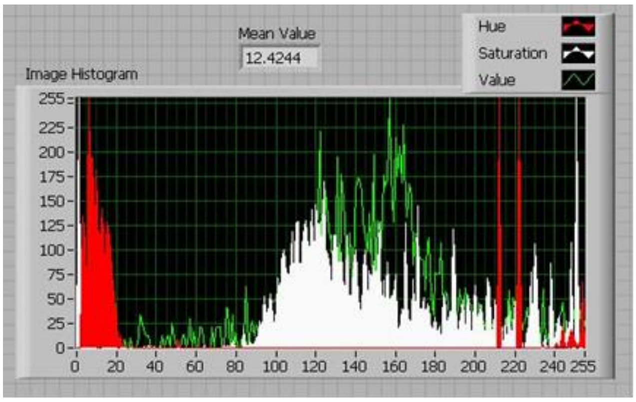

Fig. 11. depicts the full histogram in HSV planes.

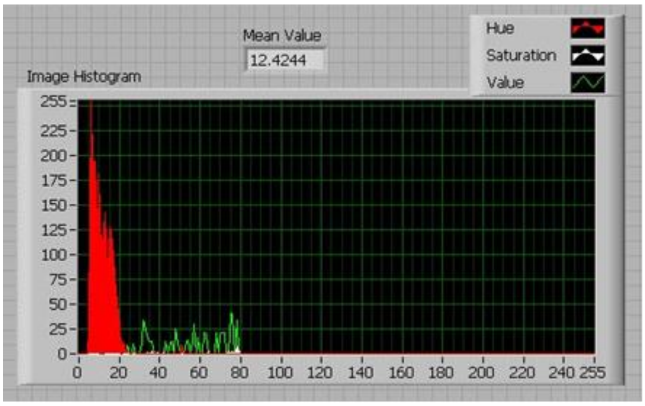

Fig. 12. Depicts the normalized histogram extracting only hue plane.

Since the histogram is one dimensional, the lookup table used can only transform a single plane of colour space instead of all the planes. Thus, it is necessary to extract only the Hue plane for further processing as shown in Figure 12. After numerous attempts, the threshold mean value of a human skin is found to be below 25.

4.2 Perceptual Grouping

In this study, the face is regarded as a point of selective attention on an image screen, as it is chosen for tracking. To fully rely and depend solely on a colour probability distribution image proves inadequate for face tracking. The probability map generated by a skin colour model may contain background noises, such as skin-like background pixels. Hence, perceptual grouping is introduced as a method to filter the image. Perceptual grouping involves the organization of image regions by a vision system into emergent boundary structures, with the goal of distinguishing objects from the scene background. The procedure is shown below:

1) *Compute log probabilities of the foreground in the image. This results in a probability distribution image,* I^o.
2) *Apply morphological erosion to* I^o*. This reduces noise and erroneous foreground and yields image* I^{er}.
3) *Let* $I^* = I^{er}$ *, then iterate the following operation to a desire number of times:*

$$I^* = \frac{1}{2}(I^* \otimes \text{low-pass filter} + I^o) \quad (10)$$

where $\otimes$denotes convolution

Such an operation effectively performs perceptual grouping in the resulting image I^* . The number of iterations can be adjusted by the user. As the number of iterations on this process increases, lesser noise will be generated in the probability image. As shown in Figure 13, the number of iterative cycles/loops (i.e. number of loop iterations range from 0 to 20 times) were used to prove this.

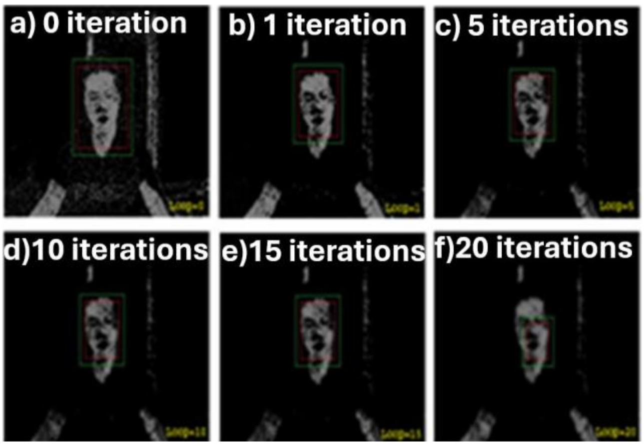

Fig. 13. Depicts the effects of increasing the iteration loops from zero to twenty times, evidently enhances the image quality, significantly eliminates noises.

4.3 Connected Component Operator

The method used to detect and track faces in colour image sequences using skin colour analysis is called Connected Component Operators. They are non-linear filters that eliminate parts of the image, while preserving the contours of the remaining parts. Compactness Solidity and Orientation are a set of shape based connected operators that make use of basic assumptions about shape of the face. These simple but effective decision criteria reply on the combinations of the area, A, the perimeter, P, and the size, D_x and D_y of the min-max box of the connected component as postulated by as postulated by Kuchi et al [28] et al.

Compactness of a connected component is defined as the ratio of its area to the square of its perimeter:

$$Compactness = \frac{A}{P^2} \tag{11}$$

The optimization is greatest for circular objects, and since faces have a nearly circular shape, the face components demonstrate a heightened value for this operator. Through observations and experimentation on diverse face components, a threshold range of 0.061 to 0.063 is established for this operator in order for face tracking to sustain. Any component with a compactness value within this range is selected for further analysis, while those outside the range are as deemed non-face and excluded.

Solidity of a connected component is defined as the ratio of its area to the area of the min-max box, as postulated by Kuchi et al as follows:

$$Solidity = \frac{A}{D_x D_y} \tag{12}$$

Solidity gives a measure of area occupancy of a connected component within its min-max box dimensions. The solidity also assumes a high value for face components. If the solidity of a component is lesser than a specified threshold value, it is eliminated, otherwise retained for further analysis. After observing various face components, the threshold value for this operator is found to be between 0.05 and 0.65.

Orientation is the aspect ratio of the min-max surrounding the component as indicated by Kuchi et al as follows:

$$Orientation = \frac{D_x}{D_y} \tag{13}$$

We presume that face components typically exhibit orientations within a specific range, determined to be between 1.01 and 1.16 through experimental observations in this research. If a component's orientation deviates beyond this range, it is excluded. To monitor a human face in a given video sequence, the aforementioned detection step is executed over a single frame as it is suggested by Kuchi et al. If any of the three operators fail to meet their threshold values, indicating the detection of a non-face, skin-like object, the search window will continue to search and track the image for possible human face.

Face tracking is illustrated using both the Red Green Blue (RGB) image and the probability distribution (PD) hue-plane image sequence, as depicted in Figures 14 and 15, respectively. The connected component operation/technique are deployed into the tracking algorithm with the bounding boxes within the human face as depicted in the mentioned Figures.

Fig. 14. Depicts the RGB video sequence where the illustration of Connected Components Operators at work.

Fig. 15. Depicts an illustration of Connected Components Operators in probability distribution (PD) hue-plane image sequence.

The three Connected Component Operators, namely, Compactness, Solidity and Orientation operator indices versus frame image sequence numbers are presented from Figures 16 to 18 respectively. To track face in any given video sequence, the detection step described above is performed combining the three mentioned operators and processing. Failing to meet the threshold values of any of the three operators will indicate a detecting of a non-face skin-like object, resulting in tracker searching over.

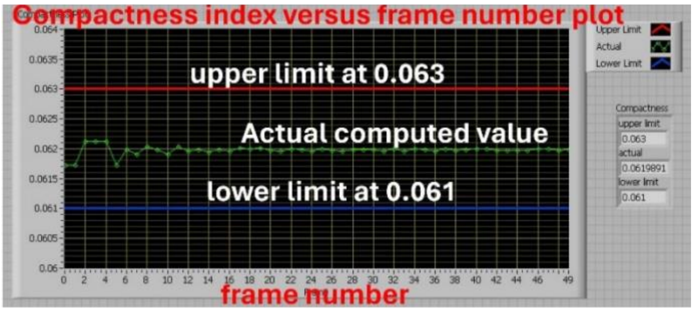

Fig. 16. Depicts the compactness operator index versus frame number sequence of the processed PD images.

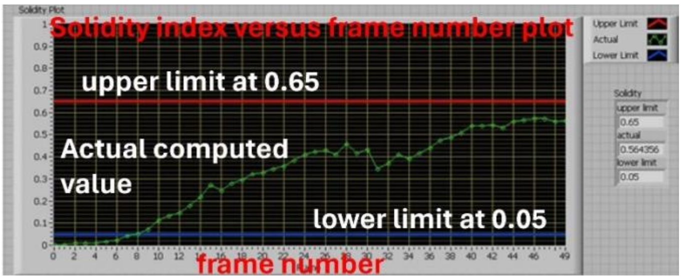

Fig. 17. Depicts the Solidity operator index versus frame number sequence of the processed PD images.

Fig. 18. depicts the Orientation operator index versus frame number sequence of the processed PD images.

4.4 Mean-Shift Algorithm

Mean shift is a nonparametric, iterative procedure first introduced by Fukunaga and Hostetler for seeking the mode of a density function represented by a set S of samples. It is based on the concept that the value of a density function at a continuity point can be estimated using the same observations that fall within a small region around that point. In earlier reference's citation, researcher, Cheng revisited Mean shift, and he had developed a more general formulation and demonstrated its potential uses in clustering and global optimization. It is also defined as a simple iterative procedure that shifts each data point to the average of data points in its neighbourhood. The following will describe the generalized mean shift procedure mathematically:

Let X be an n-dimensional real Euclidean space and S a set of sample vectors in X. Let w be a weight function from a vector in X to a nonnegative real. Let the sample mean m with kernel K at $x \in X$ be defined such as the following:

$$m(x) = \frac{\sum_{s \in S} K(\|s-x\|^2) w(s) s}{\sum_{s \in S} K(\|s-x\|^2) w(s)} \tag{14}$$

Let $M(T) = \{m(t): t \in T\}$, Let $T \subset X$ be a finite set (the "cluster centers"). One iteration of Mean-Shift is given by $T \leftarrow M(T)$. The full Mean-Shift procedure iterates until it finds a fixed point $T = M(T)$. The difference $m(x) - x$ is called mean shift. The repeated movement of data points to the sample means is called Mean-Shift algorithm. In each iteration of the algorithm, $s \leftarrow m(s)$ is performed for all $s \in S$ simultaneously. Further detailed discussion of the procedure, its definitions, and constraints, reader is invited review articles cited earlier from Fukunaga and Hostetler or Cheng.

The concept of a kernel is fundamental to the Mean-Shift procedure and, indeed, Mean Shift is conventionally defined in terms of a kernel, postulated by Fashing and Tomasi [29]. Kernel helps to provide density estimate of a set of data points. Different kernel used can get different results. If a flat kernel is used as search window, all the data points used within the kernel share the same weight and the data points outside the kernel are rejected. In the case of Gaussian kernel, all the data points within the kernel are considered. However, each data point has its own density distribution different from each other. Information provided from the data points that are further away from the center may not be available much in the computation of sample mean. It is reported that Mean-Shift algorithm using a special type of kernel will be in the gradient direction of the density estimate, as postulated by Cheng. Such a special kernel is also known as "shadow" kernel.

Shifting of sample mean in the gradient direction means that the sample mean is calculated in each iteration will climb the slope of gradient along the distribution until it reaches the mode. Epanechnikov kernel, as described by Fashing and Tomasi is selected in the tracking algorithm. Since it is a type of "shadow" kernel, the search window will be shifted along the gradient direction until it reaches the mode of the distribution. When mean shift reaches the mode (or peak), the algorithm converges. The proven mathematical procedure can be found in research paper by Fashing and Tomasi as already cited in earlier reference.

Mean-Shift algorithm is designed for static distributions. It cannot be used to handle face tracking issue/problem which involve dynamically changing distribution, such as objects in real time video sequences. Furthermore, object that moves so that the size and location of the distribution changes in time cannot be supported by Mean-Shift. In the case of dynamic distribution, the mode of the distribution keeps on changing the location and causes the kernel travels aimlessly without reaching the final goal, leading to poor localization, as postulated by Collins [30]. It also discussed that a kernel that has too big in size will include background clutter as well as the foreground object pixels. Moreover, large, big kernel can also fail by encompassing multiple modes. To compensate the limitation of mean shift algorithm, it has been modified to become CAMSHIFT (Continuously Adaptive Mean Shift) in which the size of the kernel is changing adaptively to deal with both static and dynamically distribution as postulated by

Bradski, cited in earlier reference. The details on CAMSHIFT are explained in the next section.

4.5 CAMSHIFT Algorithm

The Continuously Adaptive Mean Shift algorithm (CAMSHIFT) is an adaptation of the Mean-Shift algorithm for object tracking that is intended as a step towards head and face tracking for a perceptual user interface reported by Allen et al earlier. The CAMSHIFT algorithm proposed by Bradski can be summarized in the following steps:

1) Set the region of interest (ROI) of the probability distribution image to the entire image.
2) Select an initial location of the Mean-Shift search window. The selected location is the target distribution to be tracked.
3) Calculate a colour probability distribution of the region centered at the Mean-Shift search window.
4) Iterate Mean-Shift algorithm to find the centroid of the probability image. Store the Zeroth moment (distribution area) and centroid location.
5) For the following frame, center the search window at the mean location found in Step 4 and set the window size to a function of the Zeroth moment. Go to Step 3.

As discussed earlier in section 4.4, the Mean-Shift (centroid) within the search window of the discrete probability image computed in Step 3 can be calculated using the Zeroth, First and Second moments. Given that is the intensity of the discrete probability image at within the search window as follows:

Zeroth moment,

$$M_{00} = \sum_x \sum_y I(x, y) \quad (15)$$

First moment,

$$M_{10} = \sum_x \sum_y xI(x, y) \quad (16)$$

Second moment,

$$M_{01} = \sum_x \sum_y yI(x, y) \quad (17)$$

Further, the mean search window location (Centroid) is:

$$x_c = \frac{M_{10}}{M_{00}}; y_c = \frac{M_{01}}{M_{00}} \quad (18)$$

The Mean-Shift component of the algorithm is implemented by continually recomputing new values of (x_c, y_c) for the window position computed in the previous frame until there is no significant shift in position reported by Allen et al, cited in the reference. The algorithm must terminate in the case where M_{00} is zero, which corresponds to a window consisting entirely of zero intensity. For 2D colour probability distributions where the maximum pixel value is 255, the window size s should be set to the following:

$$s = 2 \times \sqrt{\frac{M_{00}}{256}} \quad (19)$$

An example of the CAMSHIFT in operation can be seen in the next two Figures, one of them in RGB images while the other is in PD images in a video sequence.

Fig. 19. Depicts an illustration of the standard CAMSHIFT tracking algorithm in RGB images

Fig. 20. Depicts an illustration of the standard CAMSHIFT tracking algorithm in probability distribution images.

In our experimental work and observation, the authors of this paper has asumed that for tracking faces, a typica window width is set to s and window length to 1.2s since faces are statistically reported to be elliptical in shape as in agreement with what researcher, Bradski presented in his seminal paper on CAMSHIFT.

4.6 Weighted Histogram

As mentioned earlier in this paper, the tracker would fail if Mean-Shift algorithm was implemented alone. It may consist of useless information such as background pixels within the initial search window or during the tracking time which was presented by the current authors, Alex See and Liaw in another earlier cited work in our reference. The generated 2D probability distribution image will be influenced by the untrusted lookup information. Hence, an isotropic kernel, with a convex and monotonic decreasing kernel profile $k(x)$, is chosen to assign smaller weights to pixels farther from the center, as postulated by Comaniciu et al's work cited earlier in the reference. The profile of a kernel K is defined as a function $k: [0, \infty] \rightarrow \Re$ such that $K(x) = k(\|x\|^2)$. The following equation indicates one of such kernel, called Epanechnikov Kernel, which was mentioned earlier in section 4.4, from author Y. Cheng's work.

$$K(x)\begin{cases}(1-\|x\|^2) & if\|x\|\le 1\\ 0 & if\|x\|>1\end{cases} \tag{20}$$

The weight distribution within a search window of a captured image was illustrated clearly in Figure 7 earlier. Using these weights increases the robustness of the density estimation since the peripheral pixels are the least reliable, being often affected by occlusions (clutter) or interference from the background in consistent with other workers, Comaniciu et al cited earlier. The function $b\colon \Re^2 \rightarrow \{1\ldots m\}$ associates to the pixel at location x_i^* the index $b(x_i^*)$ of its bin in the quantized feature space. The probability if the feature u = 1…m in the target model is further computed as follows:

$$\hat{q}_u = C\sum_{i=1}^{n} k(\|x_i^*\|^2)\,\delta[b(x_i^*)-u] \tag{21}$$

where δ is the Kronecker delta function and C is the normalization constant derived by imposing the condition

$$\sum_{u=1}^{m}\hat{q}_u = 1; C = \frac{1}{\sum_{i=1}^{n} k\left(\|x_i^*\|^2\right)} \tag{22}$$

since the summation of delta functions for u = 1…m is equal to one.

4.7 Selective Adaptation

As discussed in section 3.6, in practical face tracking video sequence, where unfortunately, probability distribution images are signficantly affected by background noise such as pixel form by skin-like objects, such as hand occlusion which cause erroneous frames and tracking loss. In this work, the authors proposed the selective adaptation algorithm imbued into the existing CAMSHIFT algorithm. There is a need to decide whether the new skin model is allowed to adapt or not. The adaptive mixture model seeks to maximize the log-likelihood of the color data over time. The normalized log-likelihood, $L^{(t)}$, of the data, $X^{(t)}$, observed from the object O at time t is given by McKenna et al cited earlier as in the reference.

$$L^{(t)} = \frac{1}{N^{(t)}}\sum_{x\in X^{(t)}} log\, p(x\,|\,O) \tag{23}$$

At each time frame, $L^{(t)}$, is evaluated. A sudden large drop in its value will be observed if the tracker loses tracking the object. Adaptation will be suspended until the object is again tracked with sufficiently high likelihood in consistent with McKenna et al's work.

A temporal filter was used to compute a threshold, T_t. Adaptation was only performed when $L^{(t)} > T_t$. The median, v, and standard deviation, σ , of L were computed for the $n = 2f$ most recent above-threshold frames, where $n \le L$. The threshold was set to the following:

$$T = v - k\sigma \tag{24}$$

where k was a constant found to be 1.5 and   denotes the frame rate in Hz.

The entire Selective Adaptation algorithm/process was applied accordingly to track a human subject's face, depicting a video sequence containing probability distribution (PD) images as shown in Figure 21. In Figure 22, lost of tracking was analysed by examining the selective adaption plot versus image frame number, where the the normalized log-likelihood was lesser than the threshold value, $L^{(t)} < T_t$ occurring between frame no. 18 and 21.

Fig. 21. Depicts an illustration of the implementation of the Selective Adaptation algorithm implemented in the CAMSHIFT tracking algorithm revealing in probability distribution images.

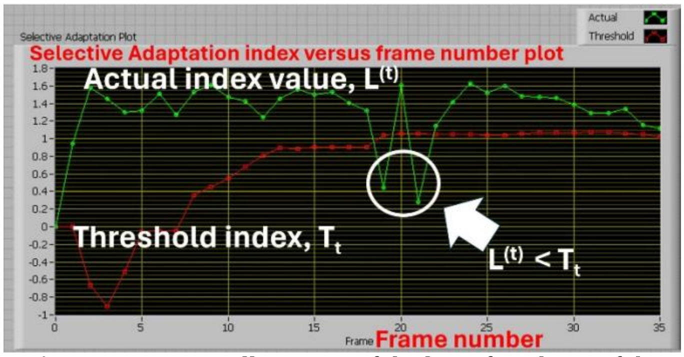

Fig. 22. Depicts an illustration of the lost of tracking of the Selective Adaptation algorithm implemented in the CAMSHIFT tracking algorithm

Figure 20 and Figure 21 show that when the tracking lost its target between frame no. 18 and 21, there is a sudden drop in the log-likelihood value. Since adaptation is performed only when $L^{(t)} > T_t$, therefore adaptation is suspended until the target is tracked again with sufficiently high likelihood, which in this case is shown in frame no. 22. In order to resume adaptation, the Hue value will be increased for effective tracking.

5. Results and Discussions

In this section, the experimental results and discussions will focus on addressing these key challenges. They are as follows:

- Occlusions
- Tracking under varying illumination environment
- Tracking of a near versus far face
- Tracking with skin-like background or object
- Disturbance from multiple faces

To benchmark and performed a fair performance comparison, the robust and resilient developed CAMSHIFT by the authors will be compared with the original CAMSHIFT developed by Bradski. For optimal results, the same video file was used for each individual test. The number of iterations for

Perceptual Grouping was set at 3 for the enhanced algorithm, unless stated otherwise. Before proceeding to the first test, a simple face tracking was performed using Bradski's Original CAMSHIFT algorithm. Referring to the probability distribution (PD) tracked images shown in Figure 24, noises appeared to be obvious. Furthermore, tracking became inaccurate due to the skin-like doors in the background as shown in frame no. 45.

Fig. 23. Depicts RGB images of Bradski's original CAMSHIFT algorithm.

Fig. 24. Depicts PD images of Bradski's original CAMSHIFT algorithm.

5.1 Occlusion

Occlusion is one of the major problems in human face detection and tracking. In this work, the author resolved by implementing the newly robust and resilient enhanced CAMSHIFT algorithm. It can be due to a hand that may trick and deceive the tracking algorithm. First, the author demonstrated using Bradski's original algorithm for a hand occlusion and secondly, follow by another unwanted interference human face entering into the tracking scene. Performance observations are reported in this section.

Fig. 25. Depicts RGB images of Bradski's original CAMSHIFT algorithm with hand occlusion. Tracker was unable to differentiate between face and hand occlusion at frame no. 28 based on the enlarge size of the white colour tracking window.

Fig. 26. Depicts PD images of Bradski's original CAMSHIFT algorithm with hand occlusions. Notice the PD images are much nosier as can be seen in the images. Tracker was unable to differentiate between face and hand occlusion at frame no. 28 based on the enlarge size of the white colour tracking window.

From Figure 25 and 26, it becomes obvious that when the hand came near to the face, the original CAMSHIFT was deceived into falsely detection and tracking the object. It can be observed that the tracking bounding window has increased its computation size to include both the face and hand as depicted in frame no. 22 and 28. As the hand leaves the face, the tracking followed the hand instead. The same video sequences were processed using the new enhanced robust and resilient CAMSHIFT, as depicted in Figure 26. The combined application of perceptual grouping, connected components operations, weighted histogram and selective adaption has enabled the tracking algorithm to be very stable and locking-in to the face target and remain within even though there was hand occlusion movement, observing frames no. from 19th to 28th.

Fig 27. Depicts PD images of enhanced robust and resilient CAMSHIFT algorithm with hand occlusions. Notice the PD images are less noisy as can be seen in the images. The red colour tracking window at frame no. 22 and 28 remains intact with the face and continues to remain stable when the hand slowly moves away from the target face towards the right side.

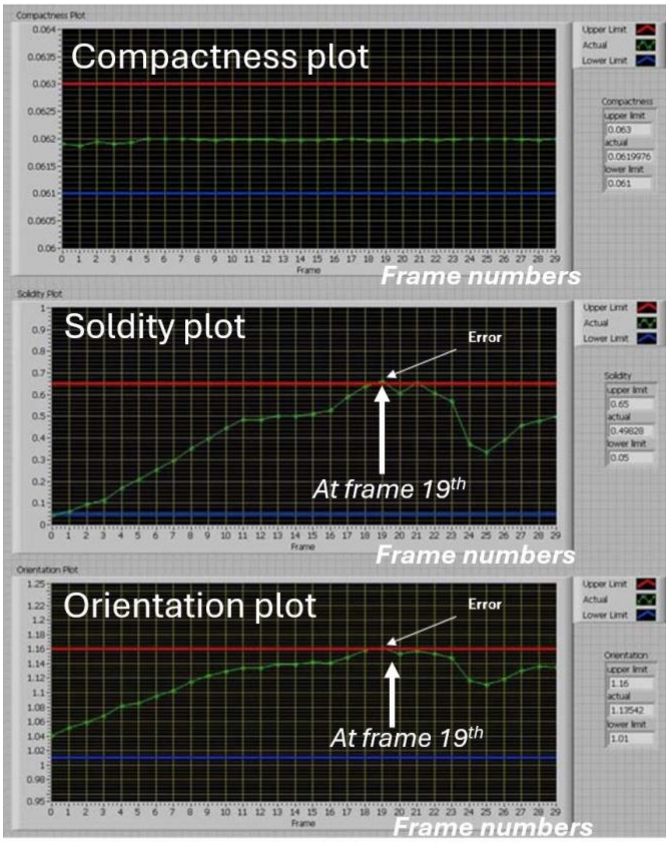

Fig. 28. Depicts analysis of Connected Component Operations, frames no. from 18 to 21 of corresponding PD images in Fig. 27.

With reference to both Figures 27 & 28, for the connected component operations, with the interference of the hand occlusion from frame no. s 19 to 21, observations are clear that this enhanced CAMSHIFT is resilient to such interference, and this tracker is able to successfully detect the error, filter it and this tracking has managed to remain in focus on its intended face target. In a separate test where another interference human face will be entering the detection and tracking scene and subject will be passing the target face in the background from left to right as depicted in Figure 29 and 30.

Fig. 29. Depicts occlusion by human subject in the background, moving from left to right direction, tracking is via original CAMSHIFT.

Fig. 30. Depicts occlusion by human subject in the background, moving from left to right direction, tracking is via original CAMSHIFT, showing PD images.

The author observed that the algorithm failed to track the target due to the extensive noisy images. The passing face also contributed to these noisy images. Unlike the Original algorithm, tracking using the enhanced robust and resilient algorithm is more accurate. The search window successfully track the target face and it did not leave it at all as depicted in Figure 30, frame no. 37, when the interference human subject's face move passed the target.

Fig. 31. Depicts occlusion by human subject in the background, using the enhanced robust and resilient CAMSHIFT

5.2 Tracking under varying illumination environment

Face tracking is often reported to be unreliable under changing illumination conditions. The recommended method is to convert the RGB colour space to a Hue, Saturation, Value (HSV) model. This is done in the two types of algorithms in concern. The following two tests are tracking under bright illumination environment and the dim condition with lux readings approximated about 290 lux and 60 lux respectively. The author began the bright illumination test, using the Original CAMSHIFT algorithm, in an enclosed room where there's no external light source from any windows. The only bright illumination came from the lights in the ceiling.

Fig. 32. Depicts RGB Images under bright illumination using the Original implemented CAMSHIFT algorithm.

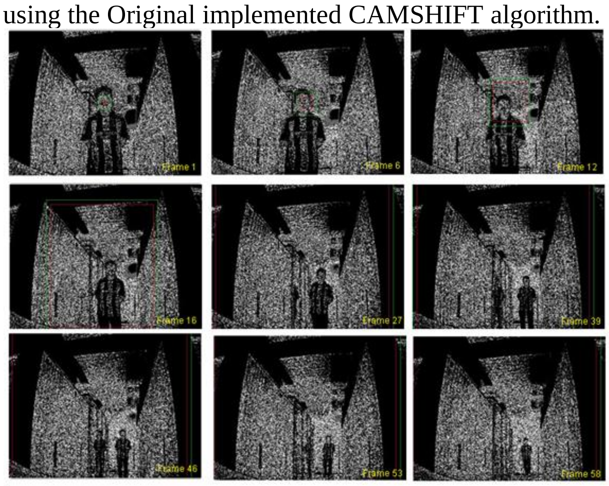

Fig. 33. Depicts PD Images under bright illumination using the implemented Original CAMSHIFT algorithm.

It can be seen from Figure 32 and 33 that the bright lights reflected onto the cabinets caused the probability distribution images to be noisy. Therefore, it created false images for the algorithm to track. The reason for the author to choose such an environment is to prove that Bradski's original CAMSHIFT algorithm may only work on certain level of bright illumination, for example a light source from a corner.

To solve this problem, the noise found in the probability distribution images must be removed, or at least be reduced. As reviewed in section 4.2, Perceptual Grouping is a process whereby a vision system organizes image regions into emergent boundary structures that aim to separate objects and scene background. It has also been discussed in section 4.2 that as the number of iterations on this process increases, lesser noise will be generated in the probability image. The author analysed the same video, which is used for the Original CAMSHIFT algorithm, chose to apply 5 iterations of the Perceptual grouping process. The tracking can be seen in RGB images and PD images in Figure 32 and 33 respectively. Comparing the probability distribution images captured in Figure 34 and 35 for the original versus enhanced CAMSHIFT respectively, it is obvious that the noise has been almost totally removed. With Perceptual Grouping technique implemented in the enhanced robust and resilient CAMSHIFT algorithm, it can successfully track the targeted face even in the most extreme bright illumination environment. This reveals the superiority of the robust and resilient enhanced CAMSHIFT developed by the authors over the original CAMSHIFT developed by Bradski.

Fig. 34. Depicts PD Images under bright illumination using the implemented enhanced robust and resilient CAMSHIFT algorithm.

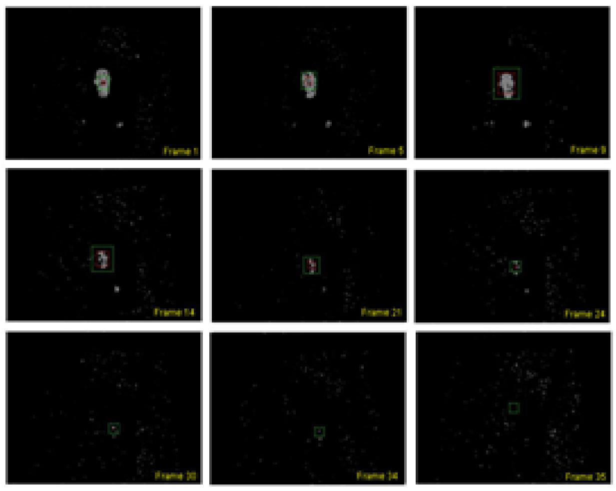

Fig. 35. Depicts PD Images under dim illumination using the implemented robust, resilient enhanced CAMSHIFT algorithm.

5.3 Tracking of a Near/Far Face

To minimise errors due to illumination, lights are maintained at a reasonable level with some of the lights switched off during the recording. In this experimental setup, the approximated total distance between the target face and the camera acquisition system was set at approximately 6 metres apart as depicted in the Figure 36. The In this experiment, the author first applied using the Original CAMSHIFIT algorithm and repeated it, following by the robust & resilient enhanced CAMSHIFT. The target face was placed initially at about 30 cm away from the camera, and target face is moved away from the camera with the distance recorded, while the algorithms are in execution. In both runs, the author recorded the video acquisition against the distance away from the camera. Initially, the original CAMSHIFT tracker managed to follow the target, but as the distance increases, due to the background noise generated, the tracker failed badly. However, the author further applied the robust and resilient enhanced CAMSHIFT algorithm with a Perceptual Grouping iteration of 1 on the same video used. The reason for this is because as the face move further away, the image became less clear. Therefore, filtering must be at selected at minimum.

Fig. 36. Depicts RGB images for Near/Far Face using the Original Algorithm.

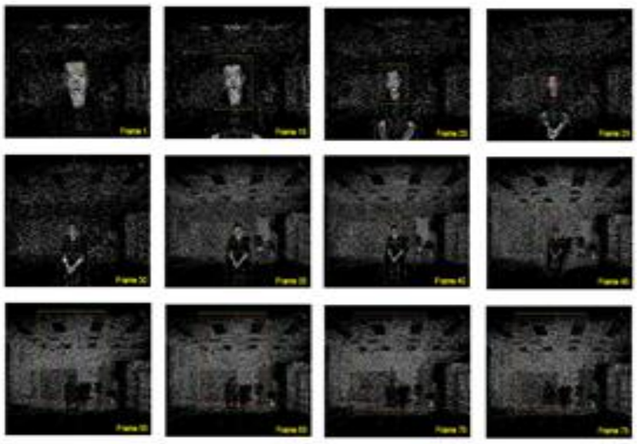

Fig. 37. Depicts PD Images for Near/Far Face using the Original CAMSHIFT algorithm.

Fig. 38. Depicts PD images for Near/Far Face using the enhanced robust and resilient CAMSHIFT algorithm.

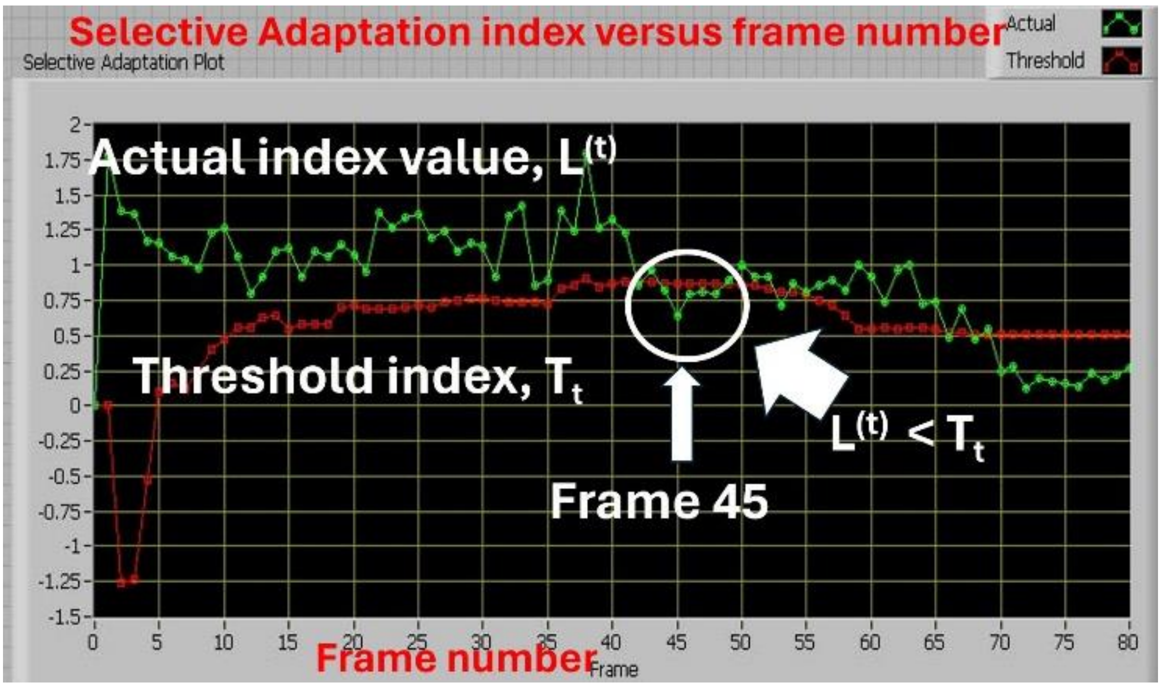

Fig. 39. Depicts the analysis of the selective adaptation for the robust and resilient enhanced CAMSHIFT.

As shown in Figure 39, one of the key experimental results is that the selective adaptation was suspended at the frame no. 44, as the tracker lost the target. This is because the normalised log-likelihood $L^{(t)} < T_t$, as illustrated at frame no. 45th. The enhanced algorithm successfully managed to resume tracking at frame no. 50. As the face moved further away, the log-likelihood began to fall below the threshold. By frame no. 75, the tracker has completely lost tracking or discontinued because the target face was too far away from the camera acquisition system, exceeded the tracking limit.

The author noted the distance of the target human face away from the camera acquisition system and plotted the result in Figure 40. The face tracking ratio that is based on the total face area covered in the search window to total area of the search window is expressed as percentages. This is to cater for optimal and fair comparison with Bradski's Original algorithm, which does not have Connected Components Operators or Selective Adaptation features built in. It is evident that examining frame no. 46, the Original CAMSHIFT algorithm can only potentially track faces up to about 350 cm away from the camera, while the robust & resilient enhanced CAMSHIFT tracker target face can potentially track as far as 500 cm, which is an improvement of 42.9%. The tracking ability suffer due to significant noise in the PD images was experienced by the Original CAMSHIFT from the 200 cm distance from the camera. However, this is not the case for the robust & resilient enhanced CAMSHIFT algorithm developed by the authors.

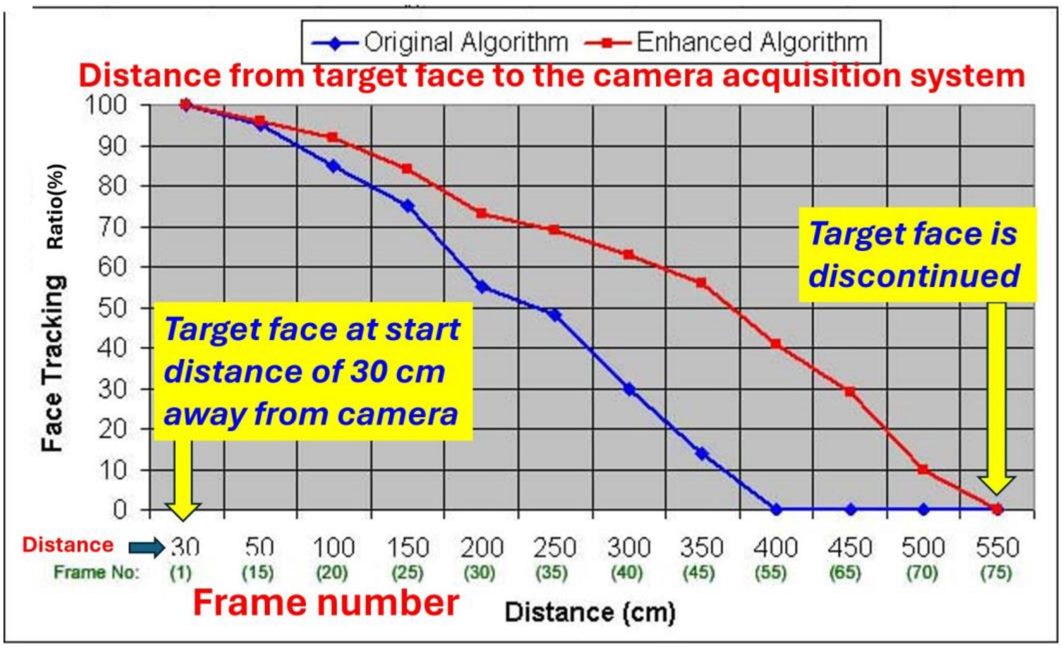

Fig. 40. Depicts the analysis of the face tracking ratio (%) versus target face's distance from the camera system.

5.4 Tracking with skin-like background or object

Typically, CAMSHIFT algorithms rely on a skin color model for human face tracking. Consequently, difficulties may arise when tracking against a background or object with a skin-like appearance. In the upcoming test video, the author intentionally selected a door and a carton box with colours resembling skin to assess and compare the performance of the Original CAMSHIFT versus the developed robust & resilient CAMSHIFT model. This is depicted in the RGB video sequence in Figure 41. It can be observed that the Original CAMSHIFT algorithm started the face detection and tracking reasonably well but began to become confused and mistakenly regarded the door as part of the face target as depicted in Figure 41, frame no. 16. In fact, at frame no. 22, both door and box are included during the computation and original tracker failed to recognize its erroneous detection and tracking.

Fig. 41. Depicts RGB images in a video sequence for skin-alike background experimentation using Original CAMSHIFT.

Fig. 42. Depicts PD images in a video sequence for skin-alike background experimentation using Original CAMSHIFT.

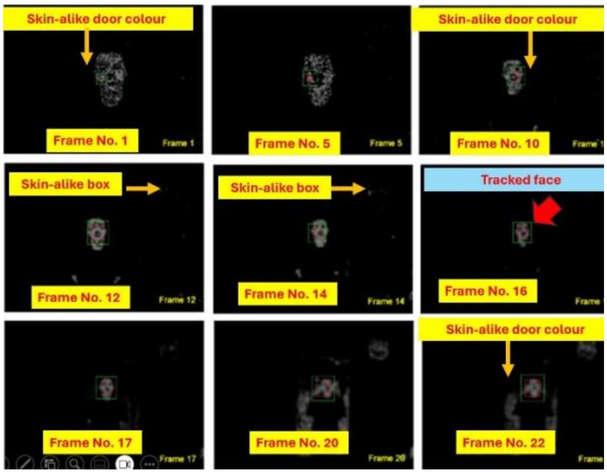

Fig. 43. Depicts PD images in a video sequence for skin-alike background experimentation using robust & resilient enhanced CAMSHIFT model.

In the probability distribution images generated by the Original algorithm depicted in Figure 42, the author observed that as early as frame no. 10, the tracker mistook the skin-like door as the target. The centroid, within the search window computed in the previous frame, followed the skin-like door as its target. As new values were recomputed at every new frame, the search window eventually covered the entire door instead of the face.

With the application Perceptual Grouping, the probability distribution image of the skin-like door in the background may be minimised as depicted in Figure 43, which is the robust & resilient enhanced CAMSHIFT. The author began the analysis with 3 iterations. But as the face became closer to the skin-like door, the noise in the probability images grew larger. At frame no. 10, the number of iterations was increased to 10 to reduce the noise. At frame no. 15, the number of iterations used increased again to 15 iterations. Finally, when the face reached the skin-like door, the noise was at the maximum.

The author has applied a total of 20 iterations of the Perceptual Grouping technique to reduce this level of noise. Therefore, the time taken to process each frame became longer. The results are shown in the Figure 43.

It is evidently that after application of 20 iterations using the Perceptual Grouping technique, the probability distribution images of the skin-like door are still slightly visible. With the addition of applying both the Connected Components Operation and Perceptual Grouping, the tracker stayed with the targeted face.

5.5 Disturbance from multiple faces

In the final experimental test on the robustness of the enhanced algorithms from external disturbance due to the presence of multiple faces in the foreground. Four other human faces were introduced and deliberately seen moving randomly in front of the targeted face. This can be challenging since most human faces have the same hue value and the tracker rely heavily on this to track. The author began the testing using the Original CAMSHIFT algorithm. The RGB video frames and PD images were depicted in Figure 44 and 45 respectively.

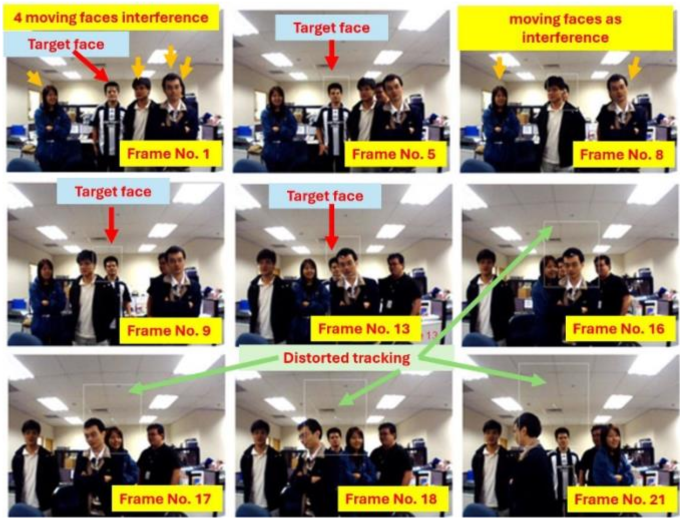

Fig. 44. Depicts RGB images in a video sequence using Original CAMSHIFT model for disturbance from multiple faces.

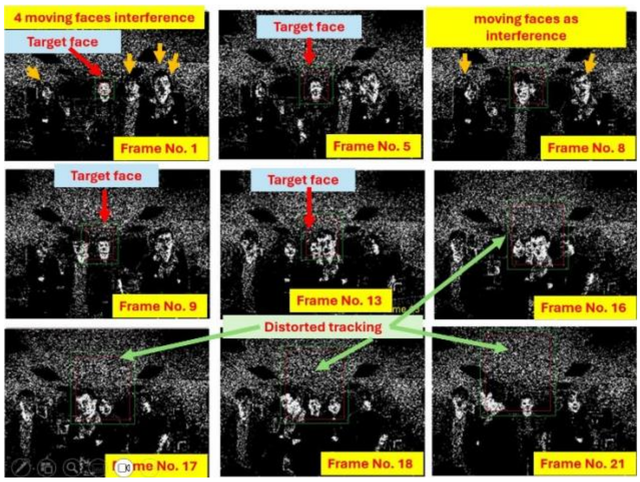

Fig. 45. Depicts PD images in a video sequence for disturbance from multiple faces experimentation using Original CAMSHIFT.

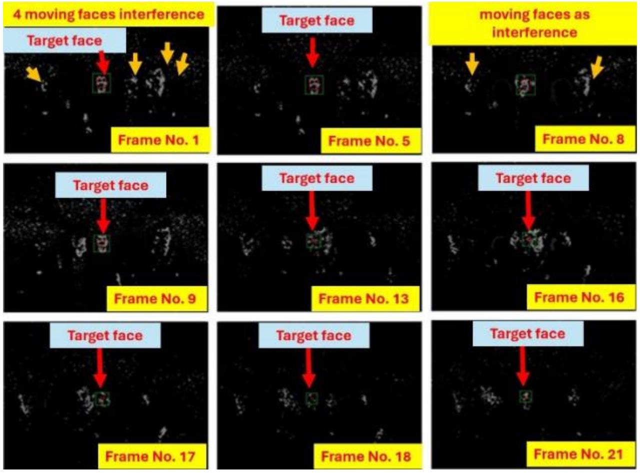

Fig. 46. Depicts PD images in a video sequence for disturbance from multiple faces experimentation using robust & resilient enhanced CAMSHIFT model. Stable and consistent tracking is observed and reported.

As shown in the probability distribution images in Figure 45, the tracking managed to stay focus on its target until at frame no. 9 where the onset of distracted tracking took place. Although the ceiling's illumination has contributed to much of the noise generated, the noise remains constant, but the moving faces were the ones that distracted the tracker. From frame no. 13 onwards, Original CAMSHIFT has become unstable, this is also observed and consistent when the time the random faces started to move about in the video sequence.

The authors experimented with the new robust & resilient enhanced CAMSHIFT model in the following test using the same video sequence. From the probability distribution images in Figure 46, it can be seen that the tracker has successfully detected and tracked the desired target face and it disallowed other interference faces as they moved across the targeted face. But the author noticed that every time a face passed by the target, the image became less clear. Overall, face tracking has been reported to be very successful even with disturbance and interferences from multiple faces.

5.6 Limitations of this study

In this study, several limitations are acknowledged and may require further research. One of the few limitations that this robust & resilient enhanced algorithm faced is the limited distance that the tracker can performed its tracking at. As demonstrated in 5.3 section, tracking of a near/far face, this algorithm/tracker is only able to track its target up to a distance of 500 cm or approximately 5 metres away from the camera acquisition system. This will not be practical for commercial usage, unless a more advance camera system is proposed with the algorithm reconFigure to suit its intended purpose. Another limitation will be the time taken to process a noisy and lengthy video. Due to the larger number of iterations used for Perceptual Grouping to reduce the noise generated, this algorithm will need more time to process each frame of the video. This may be overcome by increasing the processor's computation power of the machine under consideration. This enhanced algorithm is also unable to successfully track face with fast motion, which is purely due to the limitation of the existing hardware setup in this work. Lastly, the algorithm is unable to perform any tracking of its targeted face when it leaves and re-enters the camera's field of view. The authors believe that the above-mentioned limitations can be further addressed by either upgrading the hardware setup and re-visit the algorithm with further development.

6. CONCLUSION AND FUTURE WORK

This paper has presented a detailed discussion on the theoretical face detection and tracking on this developed robust & resilient enhanced CAMSHIFT model. Firstly, this developed model leverage the use of combined Perceptual Grouping technique, three Connected Component Operators, application of Weighted adaptive colour histogram and Selective adaptation technique. Secondly, through the experimental findings and analysis, it is evident that this proposed system can successfully outperform the Original CAMSHIFT model in the key five different scenarios as follows:

- Occlusions
- Tracking under varying illumination environment
- Tracking of a near versus far face
- Tracking with skin-like background or object
- Disturbance from multiple faces

Further, from the experimental data analysis for selective adaptation, it can be said that the normalised log-likelihood index, $L^{(t)}$ is a powerful, authentic marker/indicator to analyse as a sudden decrease in this value falling below the threshold target will imply that the face tracking has been suspended or discontinued. The Connected Component operations, namely compactness, solidity and orientation are also strong indicators/markers when used to detect tracking errors in the video sequence.

For occlusions, this enhanced CAMSHIFT algorithm has proven its robustness defined as its ability to maintain stability and functionality in its continuous tracking, adaptation of the target face, despite external interferences such as hand and face which cause occlusion. This model has also proven its worthiness as been resilient, which is defined as system's ability to quickly adapt and recover from loss of target tracking and outperform the traditional CAMSHIFT algorithm developed by Bradski. From the experimental data and analysis of the tracking of a near/far face, the robustness of the enhanced CAMSHIFT model to the Original CAMSHIFT is justifiable with data. Original CAMSHIFT algorithm can only track faces up to 350 cm away from the existing camera setup, while the robust & resilient enhanced CAMSHIFT tracker target face can track as far as 500 cm, which is an improvement of 42.9 %. The tracking ability suffer due to significant noise in the PD images is experienced by the Original CAMSHIFT from the 200 cm distance from the camera position.

Tracking with skin-like background or object and disturbances from multiple faces were presented and results analysed. The experimental results prove the superiority of the robust & resilient enhanced CAMSHIFT model versus the Original CAMSHIFT algorithm by a large margin.

In conclusion, this experimental research work has been deemed to have contributed to the body of literature in the detection and tracking of face in particular to CAMSHIFT algorithm. For the current research, the usage of webcam does provide convenience in relation to the software development of the algorithm per se. However, it is envisaged that the use of high-performance camera systems or setup will further benefit this research work in the future.

ACKNOWLEDGEMENT

The authors wish to express their gratitude to the institutions where they are currently employed.

REFERENCES

[1] S. Du. CAMShift-Based Moving Object Tracking System. IEEE 3rd International Conference on Electronic Technology, Communication and Information (ICETCI) 2023.

[2] Z. Jiang, R. Li and C.Z. Zhu. Remote Sensing Image Target Recognition System of Tennis Sports based on CAMSHIFT Algorithm. International Conference on Information System, Computing and Educational Technology (ICISCET) 2022.

[3] H. Sun, H.H. Chen, X. Cui and J. X. Wang. Vehicle Flow Statistics System in Video Surveillance based on CAMSHIFT and Kalman Filter. International Conference on the Software Process 2021;362–6.

[4] H. Wang, Q. Zhang, L. Yu and Z. Wang. Research on CAMSHIFT Algorithm Based on Feature Matching and Prediction Mechanism. IEEE International Conference on Mechatronics and Automation (ICMA) 2021.

[5] V.T. Nguyen, A.T. Nguyen, H.A. Bui and X.T. Nguyen. Real-time Target Human Tracking using CAMSHIFT and LucasKanade Optical Flow Algorithm. Advances in Science, Technology and Engineering Systems Journal 2021;6:907–14.

[6] S. Guo, C. Handong, J. Guo and J. Xu J. A Novel Target Tracking System for the Amphibious Robot based on Improved Camshift Algorithm. IEEE International Conference on Mechatronics and Automation (ICMA) 2021; pg 1419-1424.

[7] X. Hu and B. Huang. Face Detection based on SSD and CamShift. IEEE Joint International Information Technology and Artificial Intelligence Conference 2020, pg 2324-2328.

[8] Y. Zhang. Detection and Tracking of Human Motion Targets in Video Images Based on Camshift Algorithms. IEEE Sensors Journal 2020;20:11887–93.

[9] P.P. Roy, P. Kumar and B.G. Kim. An Efficient Sign Language Recognition (SLR) System Using Camshift Tracker and Hidden Markov Model (HMM). SN COMPUT. SCI. 2, 79 (2021). https://doi.org/10.1007/s42979-021-00485-z

[10] N. Zhang, J. Zhang, Optimization of Face Tracking Based on KCF and Camshift, Procedia Computer Science,Vol. 131, 2018, pg 158-166, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2018.04.199.

[11] G.R. Bradski. Real time face and object tracking as a component of a perceptual user interface. In: Applications of Computer Vision. 1998. p. 214–219.

[12] M.S. Khalid, M.U. Ilyas, M.S. Sarfaraz and M. A. Ajaz. Bhattacharyya Coefficient in Correlation of Gray-Scale Objects. Journal of Multimedia 1(1); (2006): pg57-61.

[13] L. Soni and A. Waoo. A Review of Recent Advances Methodologies for Face Detection, International Journal of Current Engineering and Technology (2023): 13. 86-92. 10.14741/ijcet/v.13.2.6.

[14] K. Hasan, S. Ahsan, A. Mamun, A.-Mamun, S. Newaz and G. Lee. Human Face Detection Techniques: A Comprehensive Review and Future Research Directions. Electronics (2021);10:2354.

[15] A. Kumar, A. Kaur and M. Kumar. Face detection techniques: a review, Artificial Intelligence Review (2019): 52, 927-948. https://doi.org/10.1007/s10462-018-9650-2

[16] F. Alqahtani, J. Banks, V. Chandran, J. Zhang. 3D Face Tracking Using Stereo Cameras: A Review. IEEE Access 2020;8:94373–93.

[17] K. Fukunaga, L.D. Hostetler, "The Estimation of the Gradient of a Density Function, with Applications in Pattern Recognition", in IEEE Transactions on Information Theory (1975), Vol. 21, Issue 1, January 1975, 32-40.

[18] S.J. McKenna, Y. Raja, S. Gong, "Tracking colour objects using adaptive mixture models", in Image and Vision Computing (1999), vol. 17, pp. 225-231.

[19] S. Gong, S.J. McKenna, A. Psarrou, "Review of Dynamic Vision: From Images to Face Recognition", in Imperial College Press. 2000.

[20] K. Schwerdt, J.L. Crowley, "Robust Face Tracking using Color", in Automatic Face and Gesture Recognition (2000), Fourth IEEE International Conference on 2000, page(s): 90-95, ISBN: 0-7695-0580-5

[21] S. Spors, R. Rabenstein, "A Real-Time Face Tracker For Color Video", in Acoustics, Speech, and Signal Processing (2001), Proceedings of IEEE International Conference (ICASSP '01), Volume 3, 7-11 May 2001, Page(s): 1493-1496, ISBN: 0-7803-7041-4

[22] T. Wang, Q. Diao, Y. Zhang, G. Song, C. Lai, G. Bradski, "A Dynamic Bayesian Network Approach to Multi-cue based Visual Tracking", in Pattern Recognition (2004), ICPR 2004, Proceedings of the 17th International Conference, Vol. 2, 23-26 Aug 2004, Page(s)167-170, ISSN: 1051-4651, ISBN: 0-7695-2128-2.

[23] Y. Cheng, "Mean Shift, Mode Seeking, and Clustering", in IEEE Transactions on Pattern Analysis and Machine Intelligence (1995), Vol. 17, Issue 8, August 1995, Page(s) 790-799

[24] Alex See Kok Bin and Yee Kang Liaw, Face Detection and Tracking Utilizing Enhanced CAMSHIFT Model, International Journal of Innovative Computing(IJICIC), Vol. 3, Issue 3, pp 597-608, ISSN 1349-4198.

[25] D. Comaniciu, V. Ramesh, P. Meer, "Kernel-Based Object Tracking", in IEEE Transactions on Pattern Analysis and Machine Intelligence (2003), Volume 25 Issue 5, May 2003, Page(s): 564-577, ISSN: 0162-8828.

[26] B.W. Silverman, "Density Estimation for Statistics and Data Analysis", in Monographs on Statistics and Applied Probability, London: Chapman & Hall (1986)

[27] J.G. Allen, R.Y.D. Xu, J.S. Jin, "Object Tracking Using CamShift Algorithm and Multiple Quantized Feature Spaces", in Proceedings of the Pan-Sydney area workshop on Visual information processing (2004), Pages: 3 – 7, ISBN ~ ISSN:1445-1336, 1-920682-18-X

[28] P. Kuchi, P. Gabbur, P S. Bhat and S.S, David. Human Face Detection and Tracking using Skin Color Modeling and Connected Component Operators. Iete Journal of Research 2002;48:289–93.

[29] M. Fashing, C. Tomasi, "Mean Shift is a Bound Optimization", in IEEE Transactions on Pattern Analysis and Machine Intelligence (2005), Volume 27, Issue 3, March 2005, Page(s) 417-474

[30] R.T. Collins, "Mean-shift Blob Tracking Through Scale Space", in Computer Vision and Pattern Recognition (2003), IEEE Computer Society Conference, Volume 2, 18-20 June 2003 Page(s): II - 234-240