IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received April 27, 2020, accepted May 3, 2020, date of publication May 14, 2020, date of current version June 2, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2994283

3D Face Tracking Using Stereo Cameras: A Review

FALEH ALQAHTANI, (Member, IEEE), JASMINE BANKS, VINOD CHANDRAN, (Senior Member, IEEE), AND JINGLAN ZHANG
School of Electrical Engineering and Computer Science, Queensland University of Technology, Brisbane, QLD 4000, Australia
Corresponding author: Faleh Alqahtani (dr.faleh@outlook.com)

ABSTRACT This review considers previous research, regarding the background and applications of 3D face-tracking systems with a focus on stereo camera-based systems. Stereo cameras are less expensive than laser ranging systems, and they are widely available on devices such as smart phones. This review aims to spur further development and applications of face tracking in this domain. Many studies on face tracking have used concepts such as the Kanade-Lucas-Tomasi method, particle filters, tracking-learning-detecting, probability hypothesis density, mean shift/cam shift, and others. As imaging constraints are relaxed, facial tracking becomes more challenging. This review presents an exposition of the most common challenges in face tracking, such as occlusion and clutter, pose variations, changes in facial resolution, illumination variations, and facial deformation. Five forms of pose estimation are discussed: appearance template methods, detector arrays, flexible models, geometric methods, and tracking methods. Applications of the listed 3D face tracking systems are also discussed, including face modelling, film editing, access control, security, and surveillance.

INDEX TERMS 3D face tracking, face detection, real-time tracking, facial landmark tracking, pose estimation, object detection, stereo camera, Kanade-Lucas-Tomasi (KLT) method, tracing learning detection (TLD), mean shift, stereo vision, illumination, object tracking, occlusion, clutter, point detector, face modelling.

I. INTRODUCTION

The use of 3D face tracking involves the application of both physiological and, to a higher degree, behavioural traits of an individual, which has attracted the attention of research organisations [1]. In the meantime, there has been a high demand for the algorithms used in this technique, which can meet real-world challenges. However, numerous obstacles have arisen from this new development because of the vulnerability of technological advancements [2]. Some challenges that the technique faces include image acquisition and imaging conditions. 3D face tracking is particularly associated with the circumstances of lighting, variations associated with large poses, and ageing occlusions. The rigidity of the head is incomplete, and this also presents difficulty. Most of these challenges are connected to illumination variations, where the images that are used in tracking patterns are affected by the presence of lighting, including the spectra and camera characteristics; among the camera characteristics are sensor response and lenses.

The associate editor coordinating the review of this manuscript and approving it for publication was Gianluigi Ciocca.

The problem of lighting has proven so prominent that it has hindered the perfect use and building of a robust system of camera control. This is the main challenge system designers have faced; however, for the generation of simple images of the same person, ignoring facial dissemination due to variations in light capacity requires the application of a pose-robust albedo estimation. However, even albedo estimation has a serious limitation. Another problem associated with the use of automated 3D face tracking is the pose or viewpoint, whereby the images under scrutiny may vary as the result of the camera's position of inclination. This factor may force some parts of the face to be partially or wholly occluded from the image representation. Another aspect associated with 3D face tracking is the working of stereo vision cameras. Stereo vision cameras work in a 3D format and can be used to establish the 3D impression of an object that is under view or being analysed by a given person. The use of 3D face tracking in reconstruction and object tracking requires a geometrical relationship between the left and right cameras, which already have fixed configurations in a majority of stereo systems. The major challenge of stereo vision cameras is their inability to take in in-field calibration. The cameras

are often pre-calibrated within the factories in which they have been bound and cannot be matched with the needs of the individual who wants to take the photo. Additionally, the images created by the stereo camera, particularly those stored within the camera, can be easily distorted and do not reflect the capability of a given individual to match the existent changes. The rectification of an image to any size could lead to significant attraction, which may alter the quality of the original image that was created. Figure 1 [3] depicts a block diagram for face tracking.

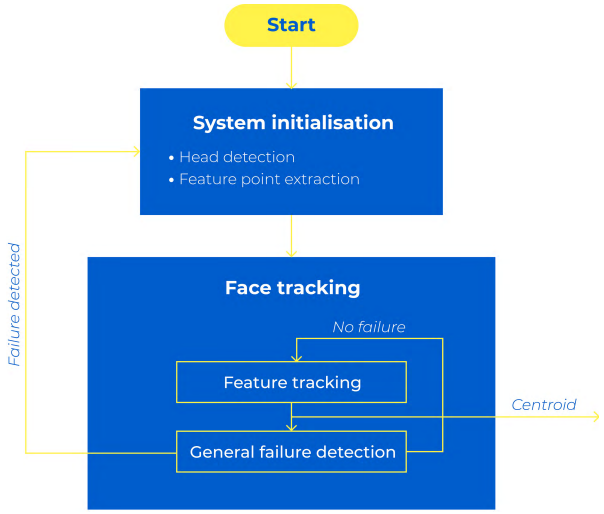

FIGURE 1. Block diagram of face tracking system [3].

II. FACE TRACKING METHODS

Face tracking methods can be grouped on the basis of the nature of the utilised technique; the methods discussed include KLT feature trackers [4], particle filter/Kalman filter method [5], PHD filters for multiple target tracking [6], TLD [7] method, and Shift/CamShift [8] methods. Tracking methods are important in 3D model-based recognition systems, in which estimation of the 3D model is achieved through the utilisation of the input video. The description of the methods shows that they fall into four major categories, as seen in their respective discussions: these include feature tracking, head tracking, image-based tracking, and model-based tracking. Face tracking algorithms are also categorised based on how they perform face tracking, which may be the whole of the face or a single feature as perceived by facial elements [9]. It is vital to track the faces through the video sequence, which helps in identifying the information that the recognition system will process. Also, face tracking is categorised based on whether it is in 3D pose space or 2D image space. Table 1 gives an overview of advantage and disadvantages of face tracking methods.

A. KLT METHOD

The KLT [4] method of face tracking is a method in which faces are tracked or detected via the use of feature points. The method is very effective because it can track a human face irrespective of whether the subject moves his head closer or away from the camera [10]. Under this method, the process is divided into three: facial detection, identifying the facial feature that can be tracked, and tracking the face [11]. In detecting a face, a detector is utilised, referred to as Cascade Object Detector System object, within a video frame [12]. Usually, the cascade object detector makes use of a detection algorithm known as the Viola and Jones [13] together with an effective trained classification model. This is because there is a default configuration for the face detection. The result of the algorithm of detection is illustrated in Figure 2 [14].

FIGURE 2. Detected face from the algorithm.

Secondly, the algorithm determines whether the facial features are reliable for tracking within the video frame. This process will identify some features on the face that can be tracked. An example of a featured detected face is shown Figure 3 [14].

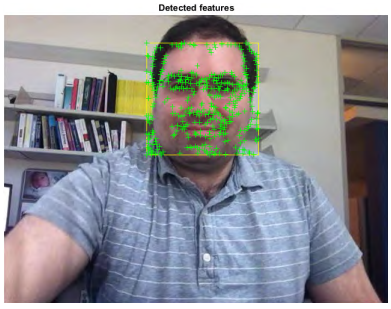

FIGURE 3. Tracked featured points from the algorithm.

The tracker is then initialised to track the points and then a video clip is initialised in order to display the results obtained from tracking the facial features [15]. In the last stage, the face of the individual is tracked. From a given frame to another, the points are tracked and a function, estimateGeometricTransform, is used to estimate the face motion. This results in the illustration shown in Figure 4 [14].

The Lucas-Kanade approach considers the flow to remain a constant within the surroundings of individual pixels, in order to ensure reducing the issues associated with the aperture in relation to well-structured regions of the pixel [10].

vector $u = [ux, uy]T$ is representative of the flow associated with $p = [px, py]T$ in relation to multiple images,

TABLE 1. Overview of advantage and disadvantages of face tracking methods.

Face Tracking Method	Advantages	Disadvantages
KLT method	(a) Most well-known real-time face detection algorithms (b) It has high detection accuracy. (c) False positive rate very low (d) Detection accuracy can be improved while computation time is reduced through construction of cascade classifiers.	(a) Very lengthy training time (b) The head poses are limited (c) Does not detect dark faces
Particle filtering method	(a) Effective in description of image texture (b) Detect moving objects through background subtraction (c) The approach is simple (d) It is tolerant to changes in monotonic illumination and computational simplicity	(a) Insensitivity to small changes in localization of the face (b) inaccurate (c) Can only be used in grey and binary images (d) Insufficient for changes in non-monotonic illumination
Trace learning detection	(a) Does not require initial knowledge on face structure (b) Easy to implement (c) Easy and simple to program (d) Does not require knowledge about weak learner (e) Does feature selection resulting in comparatively simple fast classifier	(a) Results are dependent on data with weak classifier (b) Relatively slow training (c) Overfitting resulting from weak classifier too complex (d) Low margins due to weak classifiers (e) Sensitivity to outlier and noisy data
The PHD filter	(a) Can deal with variations in illumination and sensor during object detection (b) Quickens standard Snow classifier (c) Efficient computation	(a) Little color variation compared to brightness (b) Most of the misses comprises regions that are very similar to regions of grey values present in image which may be detected as face

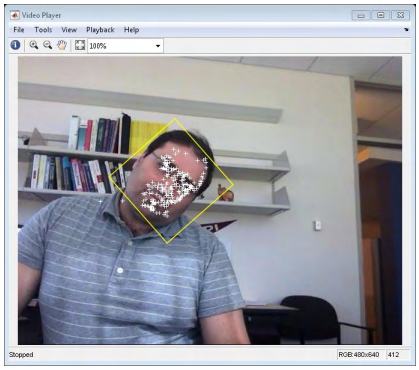

FIGURE 4. Result of tracking.

taken in consideration of the application of the minimum value in consideration of the error principle derived within a window. This is considered in relation to dual consecutive images of I and J respectively, also referred to 2D greyscale images.

Considering the window sizes of WX and WY, there is an inverse relationship in consideration of robustness and accuracy. To calculate the placement of the pixel and ensuring its accuracy, it is more preferable to have a smaller window. Inadvertently, a larger window would result in greater coarseness in coordinating the motion within the window [10].

In implementing the Lucas-Kanade algorithm in a pyramidal or iterative process, it enables the detection within the window by applying simplified motions. On a corresponding note, the impact of the window size on to the accuracy of tracking is reduced.

Such pyramidal representation enables greater efficiency in how consecutive images are viewed, which in turn concludes with representation associated with image I. Could be referred to in relation to precisions in calculating the anti-aliasing filters within the kernel.

In relation to an image of (640 $\times$ 480), images $I1$, $I2$, $I3$, and $I4$ are respectively represented in sizes of (320 $\times$ 240), (160 $\times$ 120), (80 $\times$ 60), and (40 $\times$ 30) while undertaking a scaling factor of 0.5. I0 is considered the original of the image I. The index within the pyramidal images is considered in relation to the levels within the pyramid, with the actual number of levels being a function of the image size captured relative to what is being tracked [10].

The advantages of the Kanade-Lucas-Tomasi (KLT) feature tracker are its greater convenience and affordability compared to other face detection systems. A greater spatial density of information ensures that a high resolution can be achieved, and it is also faster, especially when comparing only a few images [16]. However, KLT methods also have disadvantages due to limited registration algorithms and as a result it will often be difficult to synchronize tracking in video scenes to real situations.

B. PARTICLE FILTERING METHOD

Similarly, another method of face tracking is the particle filtering method [5]. In order to detect a face in this method, a face detector from Open CV 2.1 is used, known as cascaded Haar-feature [13]. It usually recognises regions rectangular in shape in an image that contain faces. The results are then classified in pixels as either a skin image or not-skin,

as shown in Figure 5 [17]. If the rectangular display exhibits less than 30% skin, the elimination of the face is done as false positive [18].

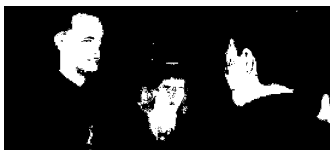

FIGURE 5. Result as seen under pixels [17].

In order to track faces in the particle filtering method, a number of samples according to a model of motion are drawn [12]. The samples will display different particles observed from different locations on the rectangular box used. The image in Figure 6 [17] below shows at which points of the rectangle the particles were tracked:

FIGURE 6. Image showing matching of the rectangles with the points [17].

After the detection of the first face, an RGB histogram is used to represent the detected face. The tracking process also requires the extraction of an image region that bears a similar height and width of the face in all the respective sampling points [12], [18]. A RGB histogram is then calculated for each sample is the same way as the previous technique. Therefore, particle filtering is purposely applied in the process of finding optimal values in the dimensional space within the rectangles. The process of getting the sample print can be done either in the X, Y axes or in the W, H axes [18]. Thus, this can only be improved by the use of better similarity scores than those obtained in the histogram correlation. However, this process could cause delay in computing and can also lead to a reduction of the frame rate of the tracking mechanism [18].

The primary advantages of particle filtering method revolve around the inherent provisions for the elements of nonlinearity. It is possible to achieve accuracy even in situations where the physical system manifests a great deal of dynamism [19]. On the other hand, the degeneracy problem is the primary disadvantage, which can occur in cases where there are a large number of iterations.

C. TRACING LEARNING DETECTION (TLD)

Tracing learning detection (TLD) is a framework that is intended for tracking unidentified faces in the video stream in the long term. In its original formulation, the whole detector was learned online from a single frame. A randomised forest was then used for representing the boundary in decision between the background and the object. Unlike in general tracking, in face tracking, there is no crucial need for building the entire detector because there is a range of faces available in the system [18], [20]. In this system, validating a face is considered less significant than detection. Also, face validation takes only a few iterations compared to face detection, because only a few faces need to be validated [20]. During learning, the validator is built; its main function is to compare the face patch so as to verify if it corresponds to that of the target. TLD also has a tracker, which provides an estimate of the motion that exists between executive frames, based on the assumption that there is a limited frame-to-frame motion, such that there is achievement of visibility for the object [20]. In case there is movement of the object from the camera, the track may fail and never be able to recover.

The other part of the TLD is the detector, which treats every frame independently and ensures a complete scanning of the face, to help in localising the faces that have been studied over the past period. This detector makes two kinds of errors, as is common with many detectors – the false positive and the false negative [18]. This implies that precision during face validation should be very high, so as not to confuse between two faces. Learning is used for observation of the performance of both the tracker and the detector. It assumes that there is a chance of failure in both devices. The objective of the TLD learning framework is to enhance the performance of the face detector through processing of the video stream online. Through learning, the detector can generalise the appearance on the faces and discriminate against the background.

Tracking learning detection (TLD) is useful in situations where multiple aspects of the objects have to be described using a single algorithm. The primary requirements include the location and extent of the image, which allows a highly accurate detection [21]. The tracker operates by following the target from frame to frame until a consistent pattern can be achieved. However, the method has disadvantages centered on the high-quality measures that have to be maintained. These include the reliability and positivity of the positive labels, where the number of positive examples cannot be ascertained and errors tied to numerous numbers of negative examples.

D. THE PHD FILTER

The probability hypothesis density (PHD) filter was first proposed and introduced by Mahler [6] as the first moment for multi-target posterior [12], [18] [20]. PHD addresses

the issues of computational intractability that attend Bayer's filters. PHD contains the dimensionality of one target state; to sample efficiently there has to be a given sum of particles that corresponds to the studied targets, which results in linear complexity. PHD does not give any information that pertains to the identity of the targets, but it can recover the posterior density function when there is an increase in the objects.

PHD can be implemented using the sequential Monte Carlo (SMC) methods. PHD-SMC is made possible by following three steps [22]. The particle step represents the PHD filter corresponding the multi-target state, which is formulated as transition density $fk|k(x)$. where x is the stated space given at time $k-1$.

The particle representation is used in the six stages of implementation. This intensity is an estimation of the expected target numbers. which is not always equal to one. The particle set is represented by the equation $(xi, wi)Nki = 1$ [22], in which Xi is the resultant weight and NK is the number of particles that has been estimated at the time of $tk - i$.

The first stage predicts the target density. Here, the particle set that has been gained from the previous steps is resampled. The particle represents the intensity that is over the state space. It can also be interpreted that each of the particles provides a representation of the target state (microstates in thermodynamics language). This allows for prediction of the entire set through application of a transition model to each particle, thereby allowing for addition of noise to it [20]. The weight remains constant. The second step is computing the correction term. Third, the target state is estimated based primarily on those targets that are detected in the time step tk. In the fourth state, the covariance is estimated. Step five is mainly on update, and then from the update, in the sixth step a standard technique for resampling is used [18].

The implementation of the PHD has also been made possible through Gaussian mixed probability hypothesis density (GMPHD). The GMPHD filter is still PHD but in a closed format, and assumes the linear Gaussian system [14]. The first assumption is that every object (face) follows a straight Gaussian model, such that:

$$fk|_{k-1}(X|\zeta) = N(x; f_k - 1_\zeta) \quad (1)$$

$$gk(_{z|x}) = N(z; H_{kx}, R_k) \quad (2)$$

In the equations, the Gaussian density is denoted by $N(m, P)$; m is the mean while P is the covariance of Gaussian density. F_{k-1} refers to the state transition matrix; the process associated with noise covariance denoted by Qk_1; Hk is equivalent to the observation matrix, and Rk is equated to observational noise covariance. The probability for survival and detection are PD,K, and ps,k, respectively. From the equation, it is possible to derive the intensity of the spontaneous birth random finite sets (RFS).

A primary benefit of The PHD filter is that random finite sets can be used as substitutes to the multi-target sets. As such accurate and dynamic face detection can be achieved in limited conditions. The technology and algorithm use data fusion to deal with multiple aspects that inhibit face detection [23]. On the other hand, a primary drawback is a requirement for a methodical approach. The PHD filter is a complex face detection technique that operates mainly using a one-dimensional scenario and the target object must be moving along a predefined line segment to achieve the desired result.

E. THE MEAN Shift CamShift METHOD

The mean shift is used in image processing and cluster analysis, in a computer vision that can be utilised in the maxima of a specific destiny function. Fukunaga and Hostetler [8] are the minds behind the introduction of the technique, which is considered to be dominant in the field of computer vision. The mean shift method, as utilised in the clustering problem, works on the assumption that every point provided is a representation of samples associated with a given function of probability density; where the areas with high density function of the sample are regarded to have correspondence to the local maxima of the said distribution. The algorithm's mechanism for locating the local maxima is that it permits the attraction of the points to one another, through a supposed low level of gravitational force. The gravitation of the points towards the greater density regions makes them merge at various points; the local maxima are found close to the region where the points converge. By locating local maxima, we are able to achieve the best resolution, because we are able to have the maximum distribution of pixels in our 3D face tracking. In summary, the mean shift algorithm functions by displacing the window to an area that has the highest density. Mean shift is regarded as "a hill climbing algorithm" [17] that is constituted of a shifting process of given kennels iterative to what is considered as a higher destiny region until it converges; this description summarises the use of mean shift in clustering. The clustering action of mean shift is as shown in Figure 7 [25].

FIGURE 7. Sample clusters found by mean shift algorithm [25].

As can be seen from Figure 7 the original segmentation of the toy car can be achieved through clustering. With this, as shown in the bottom image, the key shapes of the toy are able to be pinpointed, unlike when we just had the image of the toy itself.

Mean shift is utilised in various fields, including image processing and cluster analysis in a computer vision. Various studies shows that it is normally utilised in the detection of the different modes of the given destiny [12], [18]. The algorithm can also be applied to visual tracking; a very simple algorithm would result in the creation of a confidence map within the target image. The creation depends on the histogram's color for the given image, that was initially within the initial image. Mean shift is also applicable when identifying the peak of a specific confidence map that is regarded as near the object's initial location. Research has shown that the confidence maps refer to the probability density function on new images [17], and the result is achieved by associating a given probability to each pixel of the target image.

A primary advantage of the mean shift is that tracking can be achieved for multiple scenarios including those with interchangeable resolutions. It is also possible to concentrate on the color alone to enhance multiple properties of the object from a distance. It also has greater capacity to track in a video sequence [26]. On the other hand, a fundamental problem with the system is that it relies on there being an easily detectable single color of the object. If a color is not resolved, the results cannot be relied upon. The same case applies if illumination does not produce enough contrast.

The CamShift is utilised in the identification of the peak within a given distribution. The CamShift can also be used in the filtering of particles that are used in the imaging field and process [17]. The method is an advancement in the mean shift technique, since it provides more details regarding the face. The studies reiterate that while mean shift uses a window of a given size, the CamShift permits the subject to modify the window size so that it changes depending on the target's magnitude and rotation [17], [24]. CamShift, among other methods such as object tracking and ensemble tracking, tend to expand on the idea of mean shift [24]. However, a gap still exists in how to merge the useful ideologies about the two methods, CamShift and mean shift, in order to come up with an advanced face tracking method.

One of the key advantages of the CamShift method is its seamless flexibility. The color average of the different regions targeted for recognition can be manipulated to suit the desired results. Simplicity and robustness can also be enhanced using essential algorithmic features tied to variation. However, it is can be troublesome when two colorimetric channels cannot be formulated automatically. In addition, back projection does not often align with movements as it is traditionally tied to fixed images. The accuracy of the final product also rests on the soundness of the interest region pertaining to the target.

F. DISCUSSION

This study investigates the application of different face-tracking methods and how they lend themselves to stereo camera scenarios. The article evaluates several face-tracking methods, including the particle filters method, the KLT method, PHD filters, and the mean shift/cam shift method, discussing the challenges and short comings of each. Through this evaluation, it is apparent that the mean shift/cam shift model lends itself well to face tracking in the event of occlusions and lighting distortions. Moreover, the model allows for smoothing and segmentation when detecting faces.

The other models discussed in the article face considerable challenges in recreating 3D artefacts from the image being examined, primarily because they are unable to account for changes in lighting and posture. In addition, an algorithm like KLT that uses frames may fail to recognize certain sequences over a number of frames, although all the algorithms considered by the authors were capable of tracking and recognizing multiple objects under the right conditions.

From the information presented in the article, it is clear that face tracking and recognition algorithms have matured to the level that they can be used reliably in real-world scenarios, but there remains a need to develop them further to account for less-than-ideal scenarios in which optimal conditions cannot be ensured or when the images and videos are of poor quality. In essence, the algorithms, each suitable for numerous cases, must be further developed to aid in the realization of reliable and accurate computer vision. The improvements needed are in the areas of occlusions, multiple object tracking, and dealing with lighting conditions.

III. CHALLENGES IN STEREO VISION CAMERAS

A. BACKGROUND

Stereo vision cameras are premised to function as the human eye, albeit in a 3D manner, with the assistance of hardware and software. These cameras in most instances have at least two lenses. The stereo vision cameras operate by having two sets of different pictures of the same object taken simultaneously. These cameras have high resolution and pixels. However, they have been introduced only recently into the market, and there are challenges in the process of manufacturing them. Having still and moving images that are truly 3D has not yet been optimised through existing knowledge. These challenges in the production of stereo cameras are discussed extensively in this chapter.

B. DISCUSSION

The use of 3D face tracking in reconstruction and object tracking requires a geometrical relationship between the left and right camera, which already have fixed configurations in a majority of the stereo systems. Inferring this geometrical relationship is an achievable task, however far from the ideal canonical system. There are challenges, as proposed in [27] that are observed in stereo vision.

The main challenge is the reconstruction of 3D information of a scenario from two images that have been taken from distinct viewpoints. The essential tasks in stereo vision are calibration, correspondence, and 3D reconstruction. It is assumed that two cameras are able to distinctly meet the required setup structures of the canonical stereo system. Obtaining the corresponding pixels in both the images enables depth information to be computed and hence allow 3D reconstruction [27], [28].

In regard to a canonical system, the main objectives of stereo matching algorithms are to compute the disparities of horizontal placement that correspond with the pixels. The disparities that are observed in pixels are actually represented in a disparity map that has the same size as the stereo images, which in turn contains the identity of each pixel in an intensity value. Ideally, the disparities of two corresponding points should be determined separately, because the individual scene points of the images are projected into two image points that have features which are identifiable with each other [27]. Factors such as noise also affect the image created by cameras. It is challenging that the resultant pictures created by the left part of the camera do not completely match with the images that are created by the right part of the camera [2]. This means that there is a loss of points of the common features and this gives rise to dissimilarities. Henceforth, the main challenges in stereo visual cameras that need to be overcome are the matching of algorithms, especially stereo matching algorithms which face occluded regions, textureless regions, reflections in images, and/or the periodically textured areas that have objects that are very thin.

As seen in [27], the challenge met in stereo matching, apart from obtaining the corresponding points of the right and left image, is photo consistency, whereby the colors and intensity vary in relation to the viewpoint. This is usually brought about by having different camera sensor characteristics, as well as the electronics of the camera while acquiring the image produce noise, which in turn affects the quality of the image. It is, however, noteworthy that the differences created by these disturbances are insignificant, since they do not distort the consistency of the images [27]. The other challenge these authors note is one of uniquely matching two points, because there is luminance that is created as a result of having large regions, and this makes it difficult to find a unique match.

Stereo vision cameras face the burden of matching the pixels of the left and right image [29] with the corresponding pixels, when one side of the camera does not have the same view as the other. This occurrence is common when some of the scene parts are easily seen by one side of the camera, while the other side is occluded by various obstacles [29], [30]. Calculation of the 3D reconstruction becomes impossible if relying on pixels from one side only, and this is observed even in the calculation of the depth for this particular pixel. The stereo matching algorithms usually use constraints that are able to provide results that are acceptable to solve the stereo challenge of correspondence. The constraints are mostly in regard to the following features: ordering, uniqueness, smoothness, photo consistency, and epi-polar constraint. Consideration of the various constraints reveals other challenges in stereo vision cameras.

The smoothness constraint is the disparity of adjacent points, which vary smoothly, except at the depth of the boundaries [31]–[33]. The smoothness assumption is the consequence of observing objects and assuming that their connection with the surfaces are smooth. The algorithm of smoothness is effective in scenes which only have compact objects, while it is ineffective in computing scenes which are considered as having finely structured shapes such as hair or grass [34].

In regard to the uniqueness constraint, it is considered that the pixels of one view correspond to the pixels of the other view. If the pixels of one view fail to correspond to the pixels of the other view, it is considered that there is an occluded view [31]. Hence, the uniqueness constraint, while effective with opaque surfaces, cannot be proved while computing for surfaces that are transparent because of the depth which has been provided by both sides. The correspondence of uniqueness is also lost in scenarios where there is a slanted surface. This is because the projections from 3D lines are different from the length lines [35].

Most of the research in stereo vision has been geared towards tackling the correspondence challenge [34]. In this sense, there has been a development of various algorithms which have mainly focused on solving the challenge of stereo correspondence, and this has made it possible to evaluate the algorithms in regard to their performance. As a result of this, it is possible to analyse the boundaries, surface interiors, and scene objects [36]. The depth of object surface is, however, not smooth, though the object boundary depth is non-smooth, and this causes the appearance of depth discontinuity. Due to these discrepancies, suggestions have been made that there should be simultaneous estimations of the surface orientations, depth, creases, and occluding contours, while designing the algorithms. Such an approach is expected to enable the recovery of surface interiors through estimating the creases and the occluding contour [31], [32].

In regard to the uniqueness constraint, it is difficult to apply to the occluded and transparent regions. There is the need of defining the occlusion region so that the image region where there is no disparity should be easily identifiable [37]. The existing algorithms handle differently the regions that have been occluded. It is fundamental that there should be no assumption regarding the transparent image region uniqueness, since the uniqueness accompanies the depth discontinuities, which may be assumed as asymmetric one way or symmetric two ways [32]. Assuming the uniqueness in one way inevitably leads to every pixel that is in the reference image being assigned single disparity. However, each of the disparities can be easily recognised in multiple pixels, primarily from the second image. In instances where there is two-way uniqueness assumed, the outcome is that the individual disparity will be pointed out by both sets of pixels from both images of the stereo pair.

The computational demands that are encountered in running interactive stereo vision algorithms in order to accurately produce images of 3D from scenes are challenging [38]. This is because in developing real time systems, such as in the development of robots, it is challenging to have enough computational resources that will iteratively reproduce the data sets. Systems that are able to estimate fast 3D image depth have been developed [33]. Hence, there are two major categories that are employed in determining stereo

matching algorithms. They are global and local algorithms. The local algorithms mostly involve using constraints in a given count of pixels that are actually surrounding the anchor pixel. Likewise, the global algorithms are based on the global constraints, which involve the usage of scanning lines or even the entire image. The local algorithms have time and again proven to be efficient as well as simple, although they have the shortcoming of having low textured areas along the depth borders, as well as over-occluded regions, despite being able to achieve fast real time frame performance [39]. As for global algorithms, they are considered able to exploit the nonlocal constraints as well as additionally support the reduction of sensitivity towards the local regions of the image when they fail to match as a result of either occlusion or uniform texture. Using these constraints, however, does lead to the computational complexity observed in global matching being significantly greater than that observed in local matching [35].

Application of local matching methods at first seems to utilise the features of images such as the edges, corners and curves with the corresponding parts of the image, and this does lead to stereo matching [40]. The block areas in each pixel are made to attain the resulting matching pixel pairs in the stereo image pair; this is as a result of having feature-based algorithms, which ensure there is a robust uniform appearance and depth discontinuities in images that are limited by region-specific features of an image [1]. This does, however, affect the density of the points by limiting the regions of support into being specific in the features of the images. This in turn affects the estimation of the density points during the generation of the disparity map [36].

Application of window-based correlation, which entails matching corresponding pixels of the image stereo pair [36], can be implemented by having each pixel from the right image matched with the image of the left side through considering the colour in one of the images and applying it to the other side of the image. However, this type of corresponding method does not in any manner consider continuity or smoothness, and this does lead into having false matched pixels, due to the existence of a lot of candidate pixels satisfying the color similarity conditions [38].

The main challenges that are observed in this windowed correspondence is the selection of the particular window size. Windows that will be deemed as too small will make wrong matches on the pixel-pixel matching. Large windows, on the other hand, are observed as making less window matches, and this in turn affects the local smoothness by having constant disparities in the windows [41]. In order to have results that yield positive performances, there is need to consider using adaptive windows, whereby there are smaller and bigger windows. Smaller windows will be used in cases of near discontinuity, while larger windows will be used in cases of away discontinuities [37].

In the process, algorithms are derived that will be used for the virtual distance. Amini *et al.* [37] and Lopez-Quintero *et al.* [39] sought to realise the multimodal immersion of the human operator through the means of auditory, visual, tactile, and kinesthetic display and sensing capabilities. However, they realised that there are some drawbacks in their theory of stereo vision telepresence [37]. The drawbacks hindered the feasible implementation of a stereo vision system, hence the ideal properties, such as having communication with no time delay, or having curved camera chips, as well as displays which gave a distortion-less, extended field of view were unattainable. The studies affirm that implementing the stereo vision telepresence system will in effect lead to realising the ideal properties of stereo vision systems [42], [43]. The available components of the planar camera display and chips that are used for displaying and capturing were typically from the NTSC standard video [44]. As technology advances, the resolution and number pixels of cameras improve. However, deficiencies that will impact on a real stereo system, include latency caused by communication delays [39]. Additionally, there were delays in physical communication as well as in the data packet switching. Further challenges that are eminent in stereo camera vision arise from fixing camera set-ups. Having small apex angles, which are combined to the fixed cameras, actually leads to having only partial simultaneously visible parts of a scenario [45]. Consequently, this requires overlapping the images so that they can merge to form spatial impressions, thereby deteriorating the stereo vision quality by a significant degree [46]. This greatly hinders telemanipulation tasks, which differ with close object distances. There is a need to have natural vergence angles from human operator's eyes so that the right impressions of the depth of the remote environment are created [47].

Flórez *et al.* [41] discuss data fusion and locally consistent time of flight. They bring up the concept of depth estimation of dynamic scenes as one that presents an obstacle in the field of computer vision [41]. Solutions such as stereo vision systems, light coded cameras, and time of flight cameras, have been tested to to solve this challenge. However, stereo vision systems have tremendously improved this area of research by improving the quality of estimated geometry; hence the results that are observed have been relatively satisfactory, because texture information coding of the scenes has been limited [48]. 3D geometry has been estimated robustly by usage of time of flight and light coded cameras, which have served to estimate real time 3D geometry [47]. The downsides of these systems have been the limitations associated with low spatial resolutions.

Flight and stereo data in visual systems have been complementary to each other, with the major challenge these two sets face being the problem of them fusing together. The major goals of having stereo data and time of flight fusion is to bring forth the information of the stereo system and of time of flight camera systems, so as to obtain 3D geometry that is significantly improved [48], [49]. This can be achieved by combining the best features of both systems; for instance, high resolution as well as robustness in regard to the different scenes. It is interesting to note that stereo vision will require operators to carry the cameras to each place, and this means

that there will be consumption of more mass and energy from the vision system.

The algorithms for stereo vision, apart from being computationally intensive and having real time processing from hardware dedicated to the architecture, bring challenges that need to be overcome in coming up with the distances within various ''points in the environment'' [50]. Up to now, it has not been clear which particular algorithm will be able to provide an obstacle strong enough to inhibit the flapping wings. As for the DelFly, it has not been able to come up with a computationally efficient algorithm. Devising this type of algorithm will require consideration of system constraints to be taken into account. The bigger challenge however is that of overseeing the manufacturing of small scale stereo vision that has image lines that have been synchronised [1].

Consideration of the current stereo vision of an artificial compound eye does provide some interesting properties which are applied in the field of digital imaging. The current studies have investigated the aspects of fingerprint capturing, color imaging, and multispectral imaging. In regard to stereo vision, researchers are still trying to develop new advances that will be used towards obtaining 3D information, as well as being able to reconstruct images that are of high resolution [51]. The method that has been proposed to solve this challenge was to have a system that had its pixels modified and rearranged so as to improve estimation of the distance of the objects. This method was to utilise multiple images that are collected from various viewpoints, and thereafter captured by a thin observation module that is bound by optics. The other means was to use sum of squared difference, which would in turn reconstruct 3D images into ones of high resolution. It is disappointing that there has never been an approach developed to conclusively address the challenges of 3D visual cameras [1]. The 3D stereo vision camera has been advanced mostly by eCley, which has, over time, advanced its functions by improving on them gradually, hence producing outstanding as well as desirable functions [51].

Although stereo cameras provide useful depth information in the form of pixel displacements or disparities between the images, there are challenges associated with their use. However, this is not the case in all instances, as there are some challenges that are associated with such vision cameras. In instances where there are vision cameras fitted in cars, the distance that is given to an object is not always determined through just looking at the image, unless the object is nearer to the right judgment through pre-qualification of its correct distance [11], [52]. Another challenge is having sensors that work closely together to provide both up-to-date motion and sounds; for instance, sensors in vehicles. Stereo vision cameras have been used in remote sensing, such as automating cars or use of robots, but the ability to sense the surroundings or environment through identification of all aspects that have been programmed becomes challenging.

Ambiguous correlation from the same scene, resulting from different angles, is another potential challenge that the cameras may bring up, thereby resulting in misinterpretation. Differences in camera sensors are another challenge associated with stereo cameras [53]. Through creating various colors and resolutions, the camera gives or sends wrong signals that may hinder the overall view or perception of the overall image. In instances where the left camera is not congruent with the right camera in terms of obstruction, 3D reconstruction is hard to achieve. Detecting various objects in one go requires combinations of more than one image, which may be limiting. These combinations are supposed to help compile the relevant data for analysis and judgment. Efficient correlation is important to ensure that there are different viewpoints that may give different perspectives of the same image. Occlusion may pose another challenge in the use of vision cameras [54].

In summary, it is evident from this review of research that stereo vision in cameras has been a challenging area, one in which development of 3D face tracking systems will require substantive contributions to the field if we are to realise optimal results. 3D camera is a developing area, with a number of challenges that have been highlighted in this literature review. Nevertheless, the current advances indicate that realisation of stereo vision in 3D face tracking systems will soon become a reality, since the foundational work in this area is being advanced by the numerous research studies conducted across the globe on a daily basis [8].

Cameras are comparatively the most used appliance because of the need to capture information at a very accurate angle that is comparable to the images perceived by the human eye. The only difference between the cameras and the eye is the beautiful details that are captured by the camera, which give the image minutiae and a higher resolution. During camera selection for jobs, all three-dimensional camera factors such as camera sensors are analysed with monocular camera sensors. This system consists of two cameras mounted upon each other, and the resolution is tested at the same time. This brings the comparison recently to a mono camera and stereo camera with ADAA system [55].

The monocular video camera cannot do a lot of things the stereo camera can, but it can still identify objects clearly and accurately. The challenges faced by the stereo vision cameras can be related to image recognition and matching consistency between images. The stereo camera usually misses because colors tend to vary with intensity depending on the viewpoint [56]. The differences are so small and it could be assumed that they are constant, with the photo consistency of stereo matching algorithms, but this is still a serious problem. Secondly, the matching of two points is due to the existence of a wide position with luminance that is constant throughout the process. These regions are usually consistent with more than one correspondence point. Identification therefore depends on the number of unique points [57]. The most significant problem is now with the presence of pixels in the left image, that does not have corresponding images. This is because of the fact that during some camera captures, certain objects are occluded from the capture. When there are no

corresponding pixels, 3D construction and depth calculation is hard to analyse.

IV. CHALLENGES IN 3D FACE TRACKING

This section will discuss various challenges that are commonly experienced in face tracking and that have made the use of this technique difficult. The various challenges in 3D face tracking can be increased by scale changes in different ways. Multiple challenges have to be considered while designing the 3D face tracking system: illumination, facial expression, scaling and occlusion or clutter [58].

Firstly, the illumination and pose variations are expected to be robust. As such, the observer will often lose track of any essential changes. Even in instances where the final product can be easily determined, it is necessary to define individual alterations. This paves the way for possible improvements while defining the final image. Super-illumination will lead to loss of track given that the parameterized function would find it difficult to monitor the deviations of the essential image points. Scaling will also have a greater impact when the illumination constancy constraint of the optimal flow cannot be sustained. It also compromises the training images when a variety of data has to extracted from multiple illumination vectors [58].

It is also difficult to guarantee the quality and variety of data through the changes of expression. Lighting is the primary reference for alteration and may lead to a deviation from the expected performance as for the case of occlusion and clutter problems. Overall, changes in facial expression can be most challenging. First, facial resolution inhibits the 3D face tracking system by altering the tracking algorithm. Second, facial deformation is also an critical challenge to consider as changes in facial expressions do not necessarily complement the system.

Scaling presents multiple alternatives paving the way for a generalized or optimal face tracking. Scaling also increases problems by hampering the tracking algorithm and in most cases reducing performance. The low quality of images in the face tracking database may hamper the process of face tracking. 3D face tracking is a form of new technology that may not blend with older technologies of 1D and 2D, thus hampering the process of face recognition.

A. ROBUSTNESS TO POSE AND ILLUMINATION VARIATIONS

Pose and illumination variation has been a persistent challenge in the tracking of faces, despite having been widely studied. It is proposed in [59] that the two factors affect low level tasks such as tracking and registration. Consequently, the effectiveness and accuracy of the algorithms for tracking is reduced [42]. This implies that it is imperative that 3D face tracking methods be robust to variations in pose and illumination [60]. Illumination refers to variations in light; changes in illumination result in varying magnitude associated with the light intensity that arises from the reflection of an object; it also affects shading and shadows that are visible in an image [61]. When there are changes in lighting, the image of the same person usually appear differently. When there is a larger change induced by illumination, the systems are not able to identify the input image [54]. This is a challenge, because accurate estimation of the conditions for illumination is often difficult. This complicates the process of factoring the illumination into tracking strategies.

Researchers have proposed various methods that can be applied to handling the illumination problem [40], [62]. For instance, a person can reduce illumination variation by getting rid of the most basic eigenfaces. This method involves the use of principal component analysis (PAC) regarding the distribution of faces (eigenvectors) together with the pixel intensity features of a face [62]. Each of the face images is supposed to contribute to the eigenvector. Proof has also been provided for discarding the first few eigenfaces [40]. The only problem with this technique is that it causes degradation of the system performance for frontal input images that are taken under frontal illumination [45], [54].

Some researchers have also proposed that video-based systems can be applied such that motion is used to prove cues for face segmentation, and tracking and having more data can enhance tracking performance [63]. There are also a few challenges that are associated with video-based tracking and face tracking. Some of these challenges include: tracking and segmentation over time; low resolution of the region around the face; development of the integration measures; integration of the information of the entire sequence; and 3D modelling [56], [62].

The other widely applicable method for dealing with illumination difficulties is through modification of the brightness constancy constraint in the optical flow. Through active appearance, the model framework uses illumination 3D tracking [40], [45]. The main challenge with the technique relies on the fact that it requires an intense preparation of the images that are used to build the model; additionally, the resultant images are based on the variety and quality of such data. The main advantage of the 3D models is that they are robust in providing variation; however, they are not robust enough for illumination [64].

Another proposed solution uses image comparison. The problem with this technique is that the measures are not illumination-invariant, because when the illumination has changed, the measures for a pair image also changes. The other technique that has been applied in resolving the problem is the use of illumination subspace for a fixed viewpoint [65]. Under this fixed viewpoint, the results from tracking may be illumination-invariant. The main challenge is that the method requires many images for a person to enable construction of basic images for illumination subspace.

The other attempt at resolving the illumination challenge, as shown in [66], is the use of principal component analysis (PCA) for solving the problem of shape-from-shading by parametric statistics. This is done through reconstruction of the 3D face surface from one image on frontal illumination. This method is seen as being successful, although there is

also a challenge in reconstructing 3D surfaces from a single image [67].

With regard to pose variations, the difficulties come about due to changes in the observer's viewing angle and the rotation in the position of the head. 3D face tracking systems are usually able to detect angles of small rotation. However, when the rotation is higher, and the image that is available only has a frontal view, the result could be incorrect identification [10]. This is because the performance of systems drops significantly when the variations of the pose are presented in input images. Pose and illumination result in loss of track.

There are three types of solution that have been suggested for the challenge of variations in posture [68]. The most popular method of resolving the problem is through using multiple images of every person for the training stage and then selecting only one image in the database in the tracking stage. The other solution uses knowledge from single image methods. Alternatively, multiple images of people can be used both in the training stage and in the tracking stage [10]. Illumination techniques are a major challenge to face tracking. The response to camera sensor and intensity of lighting may have a direct influence on the quality of illumination and accurate face capture during the face tracking process [69]. Illumination variations that are related to face tracking would require taking of images under various illumination standards to ensure ease of access to the face that is to be subjected to analysis.

B. OCCLUSION AND CLUTTER

Occlusion occurs when the object of interest (a face, in this context) is partly hidden behind another object or objects in the image, for example, if the subject is wearing eyeglasses. Clutter refers to the presence of numerous objects in the background that interfere with the extraction or tracking of the object of interest, such as a face. [10], [68]. There are two primary methods that have been used to solve the challenge of clutter and occlusion when recognising faces.

The first approach is to detect some facial features. This can be done through exploitation of the relative geometric arrangement. The other technique is based on classifying the brightness pattern in an image window. This is done when the whole image is swept as non-face or face. The method for classifying brightness patterns have 90% success rate for faces that in cluttered background [70].

Other challenges in face detection are high sensitivity to partial occlusion of the face, such as a hair style that partially covers the face, or wearing glasses. Such partial occlusion or overlapping facial images distort the face, making it hard to detect. A more robust method of face tracking can be achieved through exploitation of depth information; the face can easily be separated from the background while using prior knowledge of the geometric structure of the face [52], [55]. The face position can be refined further by detecting face symmetry through the use of color for locating the point lying just above the nose and in between the eyes [40]. Also, particle filters can be used to solve the challenge.

C. POSE VARIATIONS

The pose and viewpoint of an individual is another challenge. The camera face pose may lead to a variance in the quality of the image that is taken or recorded and may alter the facial recognition features that may need to be recorded for analysis [71]. A change in the pose of the face may lead to occlusion, and bring about perceived deformation in the face image being recorded for analytical purposes [72]. Occlusion may result when the image of an individual is clustered with other images and may serve to hamper the process of identification of the image. Occlusion may lead to an image being partially identified and may not serve to bring about the full view of a face that is under analysis.

D. FACIAL RESOLUTION

The problem of low resolution occurs when the resolution of the face image that needs to be recognised is below 16×16. This is a common challenge in most of the surveillance applications, including commercially used cameras such those found in banks, supermarkets and CCTV in public streets [73]. The problem is that the images that are taken from the surveillance camera comprise a very small face area that is not sufficient for 3D face tracking. Given that the face of a person is situated further from the camera, the resultant image is less than 16×16 [10].

A low-resolution face image has minimal information, hence tracking of the face is drastically reduced. The challenge can be easily solved by lowering the camera so that it is much closer to the images [65], [74]. A super-resolution approach can also be used to overcome these challenges. However, using super-resolution also presents with some challenges, due to the detailed facial features that need accurate modelling. Facial super-resolution can also be achieved by using an active appearance model (AAM). This requires separate multiple image registration followed by interpolation. Alternatively, the super-resolved texture in the nth frame can be used to track the $(n + 1)$th frame to improve super resolution output through tracking [65].

E. TRACKING THROUGH FACIAL DEFORMATIONS

The wild face is challenging, given that its dominance is facilitated by the existence of robust visual data that is captured in an environment that has no constraints by commercial cameras. To deal with the complexity presented by the wild face, AFR methods are enhanced to sort out the complexity in real world backgrounds [63]. Such complexities include variations in skin color, gender variety, multiple face scenes, and unavoidable challenges such as resolution, image quality, facial pose correction, and illumination [63].

Face acquisition through videos intrinsically leads to facial dynamics as a result of a number of factors, including change in the point of view, camera, motion and head movements. This requires that AFR performs in real time. Inter-class similarity and intra-class variations can be dealt with through application of frames or image pre-processing, for example, face alignment [56]. Also, AFR engines can process multiple

or single camera views or synthesise the face model in 3D from one camera, which causes wider study of the computational face.

F. VARIATIONS IN EXPRESSIONS

The face is one of the imperative biometrics in humans, which, due to its unique features, has an integral role in conveying the emotions and identity of people. Particularly, people's expressions comprise macro-expressions which can express disgust, fear, anger, happiness, surprise, sadness, and other involuntary facial patterns [10], [73]. A range of emotions makes the moods of people vary, resulting in a variety of facial expressions. The hairstyle and makeup applied on the face also change the facial expressions. When the outward expression of the face is changed due to these factors, it becomes very difficult for the system utilised in 3D face tracking to provide a match to the face that has been stored in the database [10].

Several ways have been identified which can be used in tracking faces that have different expressions or have been deformed. Such methods include: tracking and reconstruction; and a data-driven technique used for tracking and recognising facial motions that are not rigid. The 3D morphable model has also been used to synthesise a variety of facial expressions, implying that this method is also suitable for tracking the coefficient set of functions that represent the morphable model. A dense optical flow field can be computed to accommodate the different facial dynamics [75].

G. AGING

The face of a human being changes over time due to ageing. The image of a person who was taken when they were ten years old will be different from one taken when they are twenty-three [62]. The facial changes due to ageing appear to be more marked when someone is below the age of eighteen [62]. People above the age of eighteen have minor changes which may involve texture and some small changes in shape, primarily due to weight changes. This ageing process poses some challenges when trying to recognise faces.

Overcoming the issues of facial ageing in 3D face tracking requires that the pattern for ageing is understood. During face ageing, shapes and lines are usually modified, and other aspects such as hairstyle may also be changed. One way of face tracking that takes ageing into account is the generative model, which applies a 3D ageing model and a pose correction method [66]. The 3D model captures ageing patterns well due to the 3D nature of the ageing process. The discriminative model has also been used successfully in facial tracking [67].

Also, some elements of the face, including moles, freckles, and scars, are important in matching images of the face. The overt utilisation of facial marks is significant due to the presence of high-resolution sensors, which are compatible with manual identification [65], [68]. The growing size of the databases available for faces also contributes to the role of facial marks in recognising faces.

Wrinkles associated with age can also be a challenge, owing to the inadequate database of images that would show the variance in the age of an individual or a given face under analysis. Wrinkles may not necessarily be used to estimate the age of an individual because of the various factors that may contribute to wrinkles. The use of makeup and progression in age may contribute to the realisation of the age of an individual.

V. POSE ESTIMATIONS

In robotics and computer vision, a distinctive task is identification of an object in an image so as to determine the orientation of the object and its position relative to a coordinate system. The data may be used by a robot to avoid moving into a specific object or to allow manipulation of a given object. A problem in pose estimation may be solved in numerous ways, depending on choice of methodology and image sensor configuration [71].

Three types of methodologies may be distinguished. Firstly, there are geometric or analytic methods. Secondly, there are genetic algorithm methods; these tend to apply the principle that if the posture of a specific object must not be calculated in actual time, then a geometric algorithm should be used [57]. Researchers claim that this method is robust precisely when the different descriptions have not been faultlessly calibrated [57], [76]. Such methods tend to use a simulated learning-based system that tends to learn the plotting from the different 2D image structures with the aim of posing transformations. Therefore, a huge collection of images containing varying poses should be offered during the learning stage. This means that when this stage has been finished, the system should represent a given approximation of the subject's pose given the particular object's image [22], [77].

A. APPEARANCE TEMPLATE METHODS

One of the critical techniques for the pose estimations that research has focused on is the appearance template mechanism/method. According to research findings, this fundamental approach is essential when the image use is based on comparison metrics to achieve the desired conditions and targets [8], [9]. From this perspective, researchers believe that the approach is valuable in comparing a new image with the set exemplars. It is essential to note that an appearance template provides a platform for the assessment or comparison of the resulting head's image to the set of exemplars with the intention of obtaining or acquiring the most similar view [8], [9]. From this perspective, the technique plays an intrinsic part in the adoption and development of the image-based comparison metrics, with the objective of matching the view of the subject's head to the given pieces of the exemplars, as per the potential labels for a pose.

Using an adaptive view based model (AVAM), it is easier to determine the pose. This template captures various levels of pose and is universal in its application to the pose model [65]. The model allows for creation of a user-specific model of tracking any slight movement or change. Using the 3D view

eigenspaces and model, the pose estimation can be calculated using $P = \{\bar{I}, VIZ, \bar{V}Z\}$ [65], where P is the eigenspaces, where $\bar{I}$ and $\bar{Z}$ are the mean intensity and depth for all the views [78]; VI andVZ act as the reference extent of eigenspaces attributed to our model. To achieve this, windows Pi found within the eigenvector matrices as depicted by every view is the template: $Pi = \{\bar{I}i, VIi, \bar{Z}i, VZi, \varepsilon i\}$ [78]. Using this variable, it is possible to get the right pose model. In the simplest implementation, the image that results from the queries obtains a similar pose which is attributed to similar templates [79]. Some of the representative examples utilise the normalised cross-correlation at multiple image resolutions as well as the mean squared error (MSE) over the sliding window [78], [79].

Evidently, appearance templates have certain advantages in comparison to the more complicated mechanisms or approaches. In the course of exploring this mechanism, researchers and practitioners have been able to examine the effectiveness and efficiency of the tool in addressing the pose estimation problem in modern/contemporary society [80]–[82]. For instance, appearance template methods provide the opportunity for the researchers to engage in easy or simple explanation, as well as the expansion of the data set for efficiency in addressing the pose estimation issue [8]. Moreover, in the utilization of this approach, researchers believe that appearance template mechanism does not need training samples, as well as facial feature points. Based on this, researchers believe that the fundamental pose estimation technique proves to be effective and appropriate for low and high-resolution imagery [80]–[82].

On the other hand, in spite of the features and effectiveness or benefits of this fundamental approach, researchers have pointed out its shortcomings [9]. For instance, in recent years, researchers have explored the limited nature of the appearance template technique, based on its approach to estimating discrete poses only [9]. They highlight its ineffectiveness and inefficiency in addressing the desired goals and targets in pose estimation in the digital world or society [71]. In exploring the limited nature of the technique, researchers have identified the critical factors affecting the implementation of the appearance template methods [71].

For example, the implementation or efficiency of the appearance template method is dependent upon the reliability of the head region detection in handling the pose estimations. Moreover, the technique proves to be computationally expensive, especially in the midst of the large datasets [60], [83]. The technique is associated with the concept of pair wise similarity, which does not necessarily translate to the essence of pose similarity [84]. Researchers have adopted and incorporated these attributes in the course of handling and exploring the effectiveness and efficiency of the appearance template method to address the pose estimation issue [85]. Researchers have highlighted that appearance templates are only capable of providing estimates for the location of discrete poses in the absence of the interpolation technique or method [86]. This approach is based on the assumption that the head region undergoes initial detection and location, which leads to a potential degradation of accuracy because the head pose estimate results from localisation errors in the implementation of the technique [78], [87].

The technique also suffers from efficiency concerns because of the tendency to adopt and incorporate more templates in the exemplar sets, contributing to an increasing demand for more computationally expensive image comparisons. In addressing this, researchers have proposed integrating the training of the set of support vector machines (SVMs) in the course of detecting and localising the face [78]. This provides the opportunity to use the support vectors as appearance templates in the estimation of the head pose [60], [86].

Despite the limitations, the major issue with the template method is that it follows faulty assumptions [61]; these assumptions are evident in the observation of pairwise similarity in the space taken by the image, that is usually equated to the pose [88]. This means that the effect of identity might contribute to issues of more dissimilarity in the image when changes in the pose are compared, and thus to an improper association of the image with the incorrect pose [88]. To lessen the impact of pairwise similarity, researchers have documented numerous approaches to various distance measurements and image transformations which serve to reduce the errors associated with pose estimation [61], [88]. For instance, it is possible to incorporate the Laplacian-of-Gaussian filter to emphasise common facial contours, as well as to remove certain identity-specific texture differences that exist among various individuals [79]. Moreover, it is possible to convolve the images with the complex Gabor wavelet in order to focus on the directed features such as the vertical lines of the nose, as well as the horizontal orientation of the mouth [81].

B. GEOMETRIC METHODS

In addition to the appearance template model or mechanism, researchers have explored the adoption and implementation of geometric methods [83]. This relates to the utilisation of the precise configuration of the local features. There is a big distinction between the approaches of computer-vision and the results of psychophysical experiments [82]. The latter focuses on the human perception of the head pose in the course of relying on the cues, as exemplified by the deviation seen in the nose angle, as well as the ''deviation of the head from the bilateral symmetry'' [78] 3D cameras have been used to track geometrical shapes of the heads of people, thus giving accurate data. The geometry-based technique is important to help steer the calculation of distances and angles to give a clear view of the object for interpretation and judgment. Through having different angles and points in setting the cameras, it is possible to get all the variables that may exist.

Effects such as the face location with regard to the head's contour have strong implications or influences on the human view of the head pose; this gives a salient cue to the head's orientation [89]. Evidently, geometric tactics for pose estimation focus on the utilisation of the head shape, as well

as the precise configuration of the local features in the estimation of the pose. These methods are interesting because of they exploit the properties that influence human head pose estimation [80]. Previous approaches sought to focus on estimating the pose through the set of feature locations. Geometric approaches use diverse mechanisms to estimate the pose through the configuration of the features [60].

The facial symmetry axis occurs through connection of the line that is found in the eyes' midpoint and extends to the mouth. It is possible to use these five points in the determination of the head pose, with the measuring point originating right from the normal towards the plane from the planar skew-symmetry [59]. In this methodology, researchers have also explored the concept of the pitch that is arrived at through comparing the distances measured from the nose, right at the tip, (and the line of the eye) to an anthropometric model [8]. In contrast to the earlier models, the technique does not offer a solution for improving the pose estimation for the near frontal views. In recent years, researchers have proposed integrating the corners of the eyes and of the mouth to facilitate automatic detection in the image [59], [88].

One of the major advantages of this approach, according to the findings of the research, is its simplicity of implementation [59]. Geometric methods prove to be fast and simple in addressing the pose estimation issue in the digital world. On the other hand, researchers have highlighted the ineffectiveness of the geometric methods to achieve high precision location of the features in estimating the pose [59]. It is important to note that the approach fails to work with low-resolution images, thus, does not adequately address the problem of pose estimation to meet the demand [60]. Researchers have highlighted the issue of occlusion in the implementation of geometric methods to achieve the desired goals and targets [60].

C. TRACKING METHODS

Researchers have adopted tracking methods or mechanisms to meet the pose estimation problem in the digital world [42]. It is important to note that tracking methods implement temporal continuity in estimating the head pose through tracking the head. To this end, practitioners tend to start with the initial position of the head [78].

Studies have noted that tracking methods function through the head's relative movements occurring between the uninterrupted frames of the video sequence [78]. This achieves temporal steadiness and leveled motion restrictions to provide appealing visual estimates of the pose across time [83], [85]. These demonstrate significant accuracy levels when the trigger arising from the subject's head position is required. The subject must maintain a frontal pose between the systems, thus, the platform for reinitialisation of the pose whenever the track is lost. The approach reduces the problem of rotational ambiguity in order to provide the desired head direction for tracking purposes [61], [88].

From this perspective, the techniques always rely on manual initialisation and the camera view for the head pose to be forward-looking. Researchers have highlighted the tendency of the method to operate in a ''bottom-up manner following the low-level facial landmarks from the frame-to-frame'' [52]. This is the basis for the adoption and implementation of an intensive approach, which assumes that the human face acts in the form of a planar surface when viewed within the orthographic space. In this approach, there are two DOF recovered through weighted least squares to determine the best transformation that can be achieved between a number of frames [9]. The approach reduces the problem of rotational ambiguity in order to provide the desired head direction for tracking purposes.

Researchers have been able to explore the benefits and positive attributes of the tracking methods [87], [89]. For instance, the technique is essential for demonstration of possible accuracy. Additionally, the method effectively handles the challenges through the incorporation of the tracking algorithm [89]. However, there are some negative aspects of the tracking methods. Pose estimation by tracking methods demands a very accurate initialisation process. Furthermore, the technique proves to be semi-automatic in the achievement of the desired goals and targets [87].

D. DETECTOR ARRAYS

Detector arrays can be described as the linear planning of separate detectors on an integrated circuit chip [40], [54]. They place the image on a spectrometer plane, which allows a variety of wavelengths to be measured simultaneously. In this case, light is normally dispersed through a fixed grating. Because sensors have no mobile parts, a complete spectrum can be attained in less than a millisecond. Additionally, the compact and rugged design of a spectral, with monolithic sensors, permits reliable, precise data acquisition with no need for recalibration. Both the CCD and the NIR devices, which have a significant dark current, are normally stabilised for drift free operations or are thermoelectrically cooled [45]. Therefore, the diode-array bases of spectrometer detectors are considered more useful in an industrialised environment [70]. In such an environment where models rapidly pass, an application of a spectrometer detector may be a quality checkup of an LED in production. In order to perfectly understand the calibration requirements of a specific instrument, it is important for an individual to understand the nature of the detector array [75]. One needs to understand its major performance parameters as well as the way each of them is measured and defined.

The 3D face tracking system has taken centre stage in the modern world, where the technology is applied in many different dimensions. The use of an Okao software pack in Amazon's camera phone, due to be launched, has a way of identifying the faces of people through certain features that are attributed to determine one's gender, race and even age [22].

In some instances, the Okao system uses close coordinates, such as linking software features like X, Y and Z coordinates from stereo aligned cameras [90], [91]. These are some of the

gains achieved through the advancement of technology, thus helping reduce the use of 3D glasses, as the cameras have advanced to incorporate the features.

Detector arrays help in the evaluation of pose estimations by ignoring the number of occlusions made by the camera [56], [62]. They consist of a photophide array PDA [62], with discrete detectors depicting an arrangement that is linear. This works by only allowing a simultaneous form of measurement for a given number, which is followed by a fixed grating that is dispersed [65]. The sensors in the arrays have no moving parts, and therefore can get a full spectrum within a millisecond [63]. Also, the rugged and compact design of the special sensors allow only accurate and reliable information acquisition because of the array's monolithic nature [62], [74]. However, the arrays have no alternation for recalibration and use the availability of NIR and CCD devices to cool the significant dark currents flowing in it. This makes the arrays very useful in the industrial environment [67].

Initialisation and subsequent tracking of the three-dimensional monocular view of motion is aimed at solving a unified work purpose [63]. This is because a pose is compared to the estimated tracking and then matched into smooth trajectories. Many times, the matching of the templates have formed the primary foundation of the proposed estimation methods [62]. A template, in contrast, refers to a combination of sets of projected model structures and the alternate pose parameters of the hand model in reference to a given pose. The creation of a large number of templates depends on the projection of computer graphics in a process that provides pose space approximation [45], [70], [75].

E. FLEXIBLE MODELS

Flexible models, also known as deformable templates, are widely used with the intention of aiding image interpretation [92]. These models are best illustrated by the case of individuals who tend to utilise models that are made by hand in the form of transaxial slices and face through vertebrae; while others use models that have contours based on the expansion of trigonometric functions [66]. Others have applied statistical techniques with the intention of learning relationships between specific shapes and other kinds of variables for the morphometric analysis. Studies show that the successful achievement of a given structured deformable template matching is dependent upon the accuracy achieved in describing the shape class [66], [92]. This includes the expected example of shapes as well as the evident variations [62]. The eventual aim of computer vision is to simulate the insight of the end user (human) besides interpreting the world that surrounds them. The main difficulty experienced in the activity of tracking of an object is to achieve ways of integrating as well as interpreting the diverse local image cues, including texture, gradient and intensity [56].

In other findings, it emerges that bottom-up methods tend to fail as a result of poor contrast, adverse viewing conditions, noise, and occlusion [63], [74]. Model-based shape matching refers to a very popular issue that is normally found in the case of a computer vision. The earliest research mainly focused on rigid shape matching in instances where the given shapes were modeled through the application of transformations; among the notable transformations were affine transformation, scaling, rotation, and scaling to the model template [63]. This can be recovered by correlational-based matching or by Hough transformation of the shape model through its flexibility, as well as its ability to trigger geometrical constraints upon the shape and integrate the evidence of the local image. The exploration of deformable models can be classified into two categories: free-form models and the parametric deformable models. The marker model looks at two variables: marker-less and a marker-based method [93]. The marker model tends to look at the endoscopic image to reach out to various aspects that are critical in making informed decisions. Through the flexible models, physicians can have an internal look at ways of improving surgery.

The free-form models may represent any form of arbitrary shape provided the constraints of general regulations are satisfied. The constraints considered in this case include smoothness and continuity. They are commonly known as active contours. These active contours were introduced through snake models [65]. In such an approach, a contour that follows nature minimises the type of energy known as a snake, which is usually controlled through the action of three different energies or forces. Firstly, there is an internal contour energy that tends to enforce smoothness. Secondly, there is the image force that tends to attract the contour to the different desired features. Thirdly, there is anexternal constraint force. A snake is normally modeled as having the ability to deform through elasticity. However, any kind of deformation tends to increase its internal energy which results in restitution force. This in turn tries to retain its original shape, but the snake remains to be immersed within a probable energy field [64]. Table 2 gives an overview of advantage and disadvantages pose estimation methods.

VI. APPLICATIONS OF FACE TRACKING

Object tracking and detection are significant in many of the vision applications within computers [8]. Some of these vision applications include automotive safety, activity tracking and surveillance. For example, to develop a system sample that can be used in face tracking, one should divide the process into three main parts: detecting a face, identifying different facial features that are of interest, and tracking the face [8]. The process automatically detects as well as tracks a face by using different feature points. This approach tends to keep track of a person's face even when the given individual tilts his or her head. It also keeps track of the individual's face should the person move away from or towards the camera. Markerless face tracking, also commonly referred to as motion capture (mo-cap) is a computer vision technology that obtains data from video sequences and still images by tracking facial landmarks in real time [8], [82]. This technology then analyses the input with the aim of perceiving

TABLE 2. Overview of advantages and disadvantages of pose estimation methods.

Pose Estimation Methods	Advantages	Disadvantages
Appearance template method	(a) Can accommodate both/all levels of resolution imagery, low and high (b) Easily adjustable to varying conditions (c) Accurate in finding the most similar view	(a) Can only estimate discrete poses (b) Time consuming
Geometric methods	(a) Save on time because they are faster (b) Simple to undertake	(a) Vulnerable to severed illumination and lower levels of resolution (b) Require exact locations of the face's target features
Tracking methods	(a) Good accuracy (b) Not easily affected by external linked errors (c) Fast to undertake	(a) Computationally expensive
Detector array methods	(a) Can take note of the previous pose (b) Do not need the step for head localization (c) Easily adapt to new conditions	(a) Quite expensive because of the sums required for training detectors used in discrete pose
Flexible models	(a) Present a platform for good invariance associated with error for head localization (b) Flexibility observed in predicting complicated poses (c) Usable for both cases of poses, discrete and continuous	(a) Computationally expensive when determining the models' optimal shapes (b) Reactive to initialization

facial expressions and head poses, and eventually rendering the information to a specific application.

Markerless facial tracking tends to capture as well as convert a particular subject's movements and expressions by pinpointing different facial landmarks with high-definition cameras and wires. When detecting a face, an algorithm must initially follow the given face by using vision [69], [94]. The next step is to initialise a tracker with the aim of tracking the points. After identifying the feature points, one can then use the system of vision, PointTrackerSystem, to locate the object with the intention of tracking them [95]. For every point that exists within the original frame, normally the point tracker attempts to find the matching points within the current frame. After this, the estimateGeometricTransform function is normally used to estimate features such as "rotation, scale and translation between the new points and old points" [95]. The witnessed transformation usually pertains to the particular bounding box found around the face. The results may be stored or played back on a video player, this is normally achieved by creating video player objects to display the video frames [94]. When tracking the face, it is important to track the points from one frame to another as well as using the estimateGeometricTransform function to estimate the motion of the face [96]. Thus, the more facial points tracked, the more accurate the illustration of facial features will be. There are different applications of face tracking: surveillance, security, access control, face modelling, and film editing [97].

A. FACE MODELLING

Face modelling refers in the beauty industry to the beauty shots that are usually taken solely of faces [98] to promote cosmetics. Cosmetic brands will hire models to appear in product shots and campaigns with a focus on different facial features. Companies like Maybelline, Max Factor, and Rimmel, need face models to promote their new lipstick, mascara or foundation to a larger audience [98]. Featuring these products on a specific model's face drives sales because consumers can see the effect and color of the product. Feldman [77] claims that face modelling is a phrase that numerous aspiring models will use to discover whether they have the correct features to succeed in specific industries. Brands will seek persons who possess a look that is attractive, with even and symmetrical features. An unbalanced look may not be found appropriate for the beauty industry, which has a huge focus on the person's facial appearance [91]. A model's commercial look is desired to meld with the company's specific vision by using their beauty products. Reference [99] explain that in such cases, the features of the individual will be assessed in terms of pleasing shapes, contours of jawline and cheekbones, and colouring of nose, lips and eyes. Reference [100] propose that digital enhancement is a reality in the modelling industry and mostly applies to the beauty industry.

For instance, mascara that supposedly makes one's eyelashes longer and fuller will probably be enhanced by Photoshop, and false lashes are used to create the desired effect [95], [97], [101]. Enhancement of facial images to make them more aesthetically pleasing and to remove blemishes has long been done using tools such as Photoshop. Extending this to the face modelling domain has further applications in the beauty industry.

As in the case of Bellus3D Face Camera, various techniques have been in put into place to help model the face to a certain degree of trueness [91]. The Bellus 3E model

with the 3D face tracking provides state of the art technology, combining an accurate face model within seconds, with high resolution in the camera's capture option, analysis, and even when need be, animation. This camera provides accurate features that have been lacking previously [22]. The camera uses the 3D shape (DepthShape) to structured-light depth to give accuracy with a high resolution mobile front camera to capture detailed aspects such as skin pores and wrinkles in real time [77]. These are just some of the features that were previously missing. The 3D vision application requires high accuracy which is provided by the 3D modelling technique that is used in face tracking.

The face modelling technology is utilised to determine motion, shape and appearance of the human face in activities such as forensic audits and identity detection in a given case. Face tracking is utilised in a number of applications, including communication through the internet, control of access, law enforcement and surveillance purposes.

Face modelling is an intricate process in the establishment of a face database that is then utilised for forensic, database, security and computerised purposes. The process begins by uploading an image or video source that is then uploaded in the face detection application. The next stage is the process of face localisation, which then leads to the face normalisation stage. To start off the process of face determination and realisation, the face localisation procedures involve the establishment of the pose and size of the image that is under analysis,. Face normalisation then leads to the extraction of features to establish the most common features that pinpoint the real identity of the image, particularly in cases of human identity. Feature matching is then determined through the use of a feature vector to bring out the 3D image program.

B. SECURITY

In this study by [86] it is argued that too many individuals regard their home as their castle; it is viewed as one place where individuals feel safe [60]. Even the slightest possibility of having their inner sanctum invaded by unwanted visitors is regarded as one of their greatest nightmares. The home security industry has grown drastically over the past few years [91]. There are different devices that tend to use face tracking as one of the ways of limiting entry, or as a way of informing householders about any intruders. For instance, there is an IntelliVision's 3D face tracking product that often detects, recognises and records individual faces appearing in a specific camera's field of view [102]; one or more individuals in a particular scene are verified or recognised through checking the image against a stored database of faces. The 3D features of cameras have enabled security to be enhanced. Security is looked at from the perspective of access, and tracking illegal entry and activities of other people who may be harmful to us. For instance, in airports, 3D cameras are used to check those who access the premises by in-depth 3D scanning for illegal items. This can be used to track down terrorists who have been put on terror alert, and give real time information on an individual's whereabouts. Through integrated features, the cameras are able to send real time signals on platforms that are created for quick and immediate response. The aim is to protect those who have been exposed to risks to their security [86]. 3D face tracking is used for public and urban area monitoring [22] and is applied in law enforcement, ports and airports.

There are different distinct features that help a face tracker improve security [22], [57]. Firstly, a face tracker enables the security personnel to shortlist, target and identify loiterers, intruders and other potential miscreants. Secondly, it is realisable as a practical system, such as Linux, Windows, Cloud solution, Inside camera and other embedded systems. Thirdly, it has high accuracy; 3D face tracking accuracy is 95 % on the public standard data set [100]. Another essential quality is image and contrast quality [94] to avoid blurred images. Real time is another important quality of face tracking in security enhancement, and tracking offline [91]. In addition, a face tracker is available with SDK and API for external systems and partners. This has several benefits, such as creating an automated log of individuals in the scene or camera, which may be used by security professionals for forensic investigations. It enables automated matching against a specific watch list with real time alerting. It is also easy to deploy a huge variety of cameras and achieve high accuracy [102].

C. SURVEILLANCE

Surveillance is defined as the act of carefully watching an individual or something particularly with the aim of detecting or preventing a crime [96]. It entails monitoring of different activities, behaviours and other kinds of changing information with the aim of directing, managing, influencing and protecting people [95], [101]. Surveillance cameras are video cameras that are used with the aim of observing a specific area; in most cases, they are connected to a specific IP network or recording device, and may be controlled by a law enforcement or security guard [22].

In a tracking learning detection (TLD) method, a face model is usually stood in for, with the collection of ''non-target and target patches'' observed [91]. The main framework usually consists of three major components. Firstly, there is a tracking component that uses a median-flow tracker with the aim of finding a face correspondence in frames. Secondly, there is a detection component that employs three layers of a cascaded classifier with the intention of selecting a patch that is very similar to the specific target face model [91]. Thirdly, there is a learning component that employs an n-expert and p-expert to select non-target and target patches with the aim of updating the face model. This method has different strengths. Firstly, it tracks constantly as long as the various appearances do not change very much from the observations [100]. Secondly, TLD learns the appearance of a particular target with respect to samples considered to be non-target and target, and hence automatically retrieves track after reappearance [96], [103]. However, it does have different weaknesses. Firstly, the face model tends to be less adaptive. In addition, the vulnerability of the TLD to drift is

usually high in a scene that is clustered. Further, it performs an exhaustive search of faces, which increases the processing time. Thirdly, tracking failure sometimes occurs if an object with the same appearance as the specific target appears in the particular scene.

Incremental Visual Tracking (IVT) is another method used in surveillance. Research shows that facial models are usually presented in a low kind of sub-space [76]. For one to find a correspondence during 3D face tracking, particle filter-based affine motion parameters are necessary, with the help of Mahalanobis and Euclidean distances for association of data. IVT has different strengths. Firstly, there is face depiction that lies on eigenspace; hence, it is likely to pose different clutter. Another strength is that on-line learning tends to incrementally update facial models based on the changes in the scene [99], [102]. 2014]. However, it does have several weaknesses. Firstly, it is susceptible to drifting, because it can slowly and effectively adapt to the non-target points in case of an update. Secondly, it lacks mechanisms to detect as well as correct drift, because it does not have any global constraints [104]. Another method is Discriminative Sparse Coding (DSC) tracking. Sparse Code is normally used for face modelling representation. The candidate regions are normally compared with the static and observation models. The main strength is that it tracks well when the appearance of a given face does not change greatly from the first frame. However, it does have several weaknesses. Firstly, a track tends to fail if a future state fluctuates drastically in comparison to the observation in the first frame. Secondly, the adaptive model in most cases fails.

Surveillance is closely linked to the security features of the camera. Using the 3D model camera, it is easy to identify and determine the changes that may be made in the process of looking at the persons who visit certain places or premises [77]. This gives a detailed look at the individual and even has a time recorder to track the time taken either within the premise or when something happens. Combining features such as vision camera as photometric stereo together with sensor and illumination technology gives the stereo cameras a positive green light in enhancing surveillance thus enhancing security [105]. Through analysing multiple sources that combine both night and day or brighter places, surveillance is improved.

D. FILM EDITING

Film editing is regarded as a technique, art and practice of assembling different shots into a sequence that is coherent. Face tracking is an important feature of film editing [83]. If one wants to change the face of an individual in a video, one can use a program known as Wondershare Filmora. This program often entails a face-off feature that automatically tracks the rotation and position of the original head in a specific picture. All that one needs to do is to choose a portrait and apply it with a particular click; the portrait will be put over the specific face, and adjusted for the right size [91]. There are different ways in which one can replace faces in a video. Firstly, the editor needs to import the videos. To import the source videos to this particular face changer, there are two available options. This includes clicking ''import'' with the intention of browsing the file folder on a computer and then loading them together. The next step is to drag and drop the wanted crops into the primary window. The face replacement software tends to support almost all the video formats [103]. Hence, one does not need to worry about an incompatibility issue. The next step is to apply a face off in a click. To replace a face, you highlight a video clip that you want to apply to a face on a particular timeline. In the specific pop-up window, then click the 'Face Off' submenu [69]. There are numerous mosaics that are available. The next step is to save the video that has the changed face [10]. This is done by hitting the button 'Create' with the intention of exporting one's video with the replaced faces. In the computer's pop-up output window, there are a variety of options to save it [102].

The use of 3D cameras is essential to ensure that the relevant data or information that is needed is captured and filtered as per the specifications [106]. Filtering of such data is through the use of editing features that are necessary for the overall identification of needed data. In films, for example, scenes can be portrayed in 3D, to give real time information on the gaps and fill them through 3D editing [99].

E. ACCESS CONTROL

In the fields of information security and physical security, access control refers to selective restriction to a specific place or any other resource. The permission to access a specific resource is known as authorisation. One of the renowned face tracking systems in access control is FaceSentinel. This is an entirely new concept of biometric access control [100]. It integrates with existing access control systems through industry standard protocols [100]. Most of the studies show that this facial tracking system is designed for applications that exhibit higher security in situations where there is need for reliability [94]. Additionally, the sensor element is totally non-contact, simple to use, and highly tolerant [94].

Abiola explains that ''face tracking for high security access control verification'' has different features [92]. It is completely non-contact, light immune, simple and fast to use, tends to store images of all the users, and has a Wiegand interface. Being completely non-contact, FaceSentinel is hygienic as well as easy to use. It is designed to assimilate with the existing access control systems through RS485 or Weigand. In addition, FaceSentinel stores images of all the transactions [92]. Therefore, the images of all individuals who fail to gain or gain access to the system are usually recovered. FaceSentinel normally comprises of any IR sensor that is within reach with an LED ceramic that is associated with one of the ultra-small factor PCs.

VII. CONCLUSION

In conclusion, the use of 3D face tracking involves the application of both physiological and, to a higher degree, behavioral traits of an individual [35], [36]. The reviewed studies

describe how 3D face tracking is employed as part computer technology that is used in the identification of human faces in the form of digital images. This technology has attracted the attention of research organizations. There has been a high demand for the algorithms used in this technique, which has the ability to face real world challenges; this area has not been well addressed by the studies. The studies have explored a number of techniques used for tracking; among these are appearance templates, tracking methods, and geometric methods. However, numerous challenges have arisen from this new development because of the vulnerability of technological advancements. Some of the challenges that the technique presents are image acquisition and imaging conditions.

Stereo cameras have provided the necessary data and images that are useful. However, there are still some challenges are associated with these vision cameras. In instances where vision cameras are fitted in cars, the distance that is given to an object is not always accurate, unless the object is nearer, allowing right judgment through pre-qualification of its correct distance. Another challenge is having sensors that work closely together to provide both up-to-date motion and sounds, for instance, sensors in vehicles. Stereo vision cameras have been used in remote sensing such as automating cars or use of robots, but the ability to sense the surrounding environment through identification of all aspects that have been programmed presents challenges.

REFERENCES

[1] I. Labutov, C. Jaramillo, and J. Xiao, "Generating near-spherical range panoramas by fusing optical flow and stereo from a single-camera folded catadioptric rig," *Mach. Vis. Appl.*, vol. 24, no. 1, pp. 133–144, Jan. 2013.

[2] G. de Croon, M. Perçin, B. D. Remes, R. Ruijsink, and C. De Wagter, *The DelFly Design, Aerodynamics, and Artificial Intelligence of a Flapping Wing Robot*. Cham, Switzerland: Springer, 2015.

[3] V. Atienza and J. M. Valiente, "Face tracking algorithm using grey-level images," in *Proc. Int. Conf. Signal Process. Commun.*, 2000, pp. 507–512.

[4] C. Tomasi and T. K. Detection, "Tracking of point features," Carnegie Mellon Univ., Pittsburgh, PA, USA, Tech. Rep. CMU-CS-91-132, 1991.

[5] P. M. Djuric, J. H. Kotecha, J. Zhang, Y. Huang, T. Ghirmai, M. F. Bugallo, and J. Miguez, "Particle filtering," *IEEE Signal Process. Mag.*, vol. 20, no. 5, pp. 19–38, Sep. 2003.

[6] R. Mahler, "PHD filters of higher order in target number," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 43, no. 99, pp. 1523–1543, Oct. 2007.

[7] Z. Kalal, K. Mikolajczyk, and J. Matas, "Face-TLD: Tracking-learning-detection applied to faces," in *Proc. IEEE Int. Conf. Image Process.*, Sep. 2010, pp. 3789–3792.

[8] H. Boughrara, M. Chtourou, C. B. Amar, and L. Chen, "Facial expression recognition based on a mlp neural network using constructive training algorithm," *Multimedia Tools Appl.*, vol. 75, no. 2, pp. 709–731, Jan. 2016.

[9] A. Danelakis, T. Theoharis, and I. Pratikakis, "A survey on facial expression recognition in 3D video sequences," *Multimedia Tools Appl.*, vol. 74, no. 15, pp. 5577–5615, Aug. 2015.

[10] C. Nock, O. Taugourdeau, S. Delagrange, and C. Messier, "Assessing the potential of low-cost 3D cameras for the rapid measurement of plant woody structure," *Sensors*, vol. 13, no. 12, pp. 16216–16233, 2013.

[11] L. Wang, X. Zhao, and Y. Liu, "Adaptive appearance learning for human pose estimation," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2015, pp. 1125–1129.

[12] R. A. Newcombe, A. Fitzgibbon, S. Izadi, O. Hilliges, D. Molyneaux, D. Kim, A. J. Davison, P. Kohi, J. Shotton, and S. Hodges, "KinectFusion: Real-time dense surface mapping and tracking," in *Proc. 10th IEEE Int. Symp. Mixed Augmented Reality*, Oct. 2011, pp. 127–136.

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

[14] MathWorks. (Sep. 2014). *Face Detection and Tracking Using the KLT Algorithm*. Accessed: Jun. 10, 2019. [Online]. Available: https://www.mathworks.com/help/vision/examples/face-detection-and-tracking-using-the-klt-algorithm.html

[15] M. Karpushin, G. Valenzise, and F. Dufaux, "Good features to track for RGBD images," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 1832–1836.

[16] N. Al-Najdawi, S. Tedmori, E. A. Edirisinghe, and H. E. Bez, "An automated real-time people tracking system based on KLT features detection," *Int. Arab J. Inf. Technol.*, vol. 9, no. 1, pp. 100–107, 2012.

[17] K. Brocklehurst. (2016). Particle filtering for face tracking. Pennsylvania State University. Accessed: Jun.10, 2019. [Online]. Available: http://vision.cse.psu.edu/ people/kyleB/faceTracker/index.shtml

[18] M. Ye, Q. Zhang, L. Wang, J. Zhu, R. Yang, and J. Gall, "A survey on human motion analysis from depth data," in *Proc. Time-of-Flight Depth Imag. Sensors, Algorithms, Appl.* Cham, Switzerland: Springer, 2013, pp. 149–187.

[19] M. S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp, "A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking," *IEEE Trans. Signal Process.*, vol. 50, no. 2, pp. 174–188, 2002.

[20] N. Smolyanskiy, C. Huitema, L. Liang, and S. E. Anderson, "Real-time 3D face tracking based on active appearance model constrained by depth data," *Image Vis. Comput.*, vol. 32, no. 11, pp. 860–869, Nov. 2014.

[21] Z. Kalal, K. Mikolajczyk, and J. Matas, "Tracking-learning-detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 7, pp. 1409–1422, Jul. 2012.

[22] S. Kajalo and A. Lindblom, "The role of formal and informal surveillance in creating a safe and entertaining retail environment," *Facilities*, vol. 34, nos. 3–4, pp. 219–232, Mar. 2016.

[23] B.-N. Vo, S. Singh, and A. Doucet, "Sequential Monte Carlo implementation of the PHD filter for multi-target tracking," in *Proc. 6th Int. Conf. Inf. Fusion*, 2003, pp. 792–799.

[24] D. Ramunno-Johnson. (2016). *The Mean Shift Clustering Algorithm. Everybody's Favorite Data Blog: Adventures in Machine Learning and Data Science*. Accessed: Jun. 10, 2019. [Online]. Available: http://efavdb.com/mean-shift

[25] A. Pooransingh, C.-A. Radix, and A. Kokaram, "The path assigned mean shift algorithm: A new fast mean shift implementation for colour image segmentation," in *Proc. 15th IEEE Int. Conf. Image Process.*, Oct. 2008, pp. 597–600.

[26] N. M. Artner, "A comparison of mean shift tracking methods," in *Proc. 12th Central Eur. Seminar Comput. Graph.*, vol. 23, 2008, pp. 1–8.

[27] M. Gosta and M. Grgic, "Accomplishments and challenges of computer stereo vision," in *Proc. ELMAR*, Sep. 2010, pp. 57–64.

[28] T. Jiang, *Stereo Vision for Facet Type Cameras*. Berlin, Germany: Logos Verlag Berlin GmbH, 2016.

[29] M. S. Kristoffersen, J. V. Dueholm, R. Gade, and T. B. Moeslund, "Pedestrian counting with occlusion handling using stereo thermal cameras," *Sensors*, vol. 16, no. 1, p. 62, 2016.

[30] C. Bouvier and Y. Ni, "Logarithmic image sensor for wide dynamic range stereo vision system," *Procedia Comput. Sci.*, vol. 39, pp. 146–149, Jan. 2014.

[31] N. Sun, S. Murakami, H. Nagaoka, and T. Shigemoto, "A correction algorithm for stereo matching with general digital cameras and Web cameras," *Int. J. Space-Based Situated Comput.*, vol. 3, no. 3, p. 169, 2013.

[32] F. Hahn, S. Jensen, and S. Tanev, "Disruptive innovation vs disruptive technology: The disruptive potential of the value propositions of 3D printing technology startups," *Technol. Innov. Manage. Rev.*, vol. 4, no. 12, pp. 27–36, 2014.

[33] R. Srivastava and S. Roy, "Utilizing 3D flow of points for facial expression recognition," *Multimedia Tools Appl.*, vol. 71, no. 3, pp. 1953–1974, Aug. 2014.

[34] S. Asteriadis, K. Karpouzis, and S. Kollias, "Visual focus of attention in non-calibrated environments using gaze estimation," *Int. J. Comput. Vis.*, vol. 107, no. 3, pp. 293–316, May 2014.

[35] H. Liu, K. R. Hao, and Y. S. Ding, "New anti-blur and illumination-robust combined invariant for stereo vision in human belly reconstruction," *Imag. Sci. J.*, vol. 62, no. 5, pp. 251–264, 2014.
[36] T. Al Smadi, "Pedestrian crossings detection by using driving assistance systems," *J. Commun. Technol., Electron. Comput. Sci.*, vol. 9, p. 17, Jan. 2017.
[37] A. S. Amini, M. Varshosaz, and M. Saadatseresht, "Development of a new stereo-panorama system based on off-the-shelf stereo cameras," *Photogramm. Rec.*, vol. 29, no. 146, pp. 206–223, Jun. 2014.
[38] S. Morita and Y. Ozaki, "Moving-window two-dimensional correlation spectroscopy and perturbation-correlation moving-window two-dimensional correlation spectroscopy," *Chemometric Intell. Lab. Syst.*, vol. 168, pp. 114–120, Sep. 2017.
[39] M. I. López-Quintero, M. J. Marín-Jiménez, R. Muñoz-Salinas, F. J. Madrid-Cuevas, and R. Medina-Carnicer, "Stereo pictorial structure for 2D articulated human pose estimation," *Mach. Vis. Appl.*, vol. 27, no. 2, pp. 157–174, Feb. 2016.
[40] M. F. E. M. Senan, S. N. H. S. Abdullah, W. M. Kharudin, and N. A. M. Saupi, "CCTV quality assessment for forensics facial recognition analysis," in *Proc. 7th Int. Conf. Cloud Comput., Data Sci. Eng. Confluence*, Jan. 2017, pp. 649–655.
[41] S. A. Rodríguez Flórez, V. Frémont, P. Bonnifait, and V. Cherfaoui, "Multi-modal object detection and localization for high integrity driving assistance," *Mach. Vis. Appl.*, vol. 25, no. 3, pp. 583–598, Apr. 2014.
[42] H. Patil, A. Kothari, and K. Bhurchandi, "3-D face recognition: Features, databases, algorithms and challenges," *Artif. Intell. Rev.*, vol. 44, no. 3, pp. 393–441, Oct. 2015.
[43] J. Lee, M.-H. Jeong, J. Lee, K. Kim, and B.-J. You, "3D pose tracking using particle filter with back projection-based sampling," *Int. J. Control, Autom. Syst.*, vol. 10, no. 6, pp. 1232–1239, Dec. 2012.
[44] G. Fanelli, M. Dantone, J. Gall, A. Fossati, and L. Van Gool, "Random forests for real time 3D face analysis," *Int. J. Comput. Vis.*, vol. 101, no. 3, pp. 437–458, Feb. 2013.
[45] G. S. Walia and R. Kapoor, "Recent advances on multicue object tracking: A survey," *Artif. Intell. Rev.*, vol. 46, no. 1, pp. 1–39, Jun. 2016.
[46] D. Fidaleo and G. Medioni, "Model-assisted 3D face reconstruction from video," in *Proc. Int. Workshop Anal. Modeling Faces Gestures*. Berlin, Germany: Springer, 2007, pp. 124–138.
[47] H. Yu, O. Garrod, R. Jack, and P. Schyns, "A framework for automatic and perceptually valid facial expression generation," *Multimedia Tools Appl.*, vol. 74, no. 21, pp. 9427–9447, Nov. 2015.
[48] J. Behmann, A.-K. Mahlein, S. Paulus, J. Dupuis, H. Kuhlmann, E.-C. Oerke, and L. Plümer, "Generation and application of hyperspectral 3D plant models: Methods and challenges," *Mach. Vis. Appl.*, vol. 27, no. 5, pp. 611–624, Jul. 2016.
[49] J. Li, A. M. Kaneko, G. Endo, and E. F. Fukushima, "In-field self-calibration of robotic manipulator using stereo camera: Application to humanitarian demining robot," *Adv. Robot.*, vol. 29, no. 16, pp. 1045–1059, Aug. 2015.
[50] S. S. Rautaray and A. Agrawal, "Vision based hand gesture recognition for human computer interaction: A survey," *Artif. Intell. Rev.*, vol. 43, no. 1, pp. 1–54, Jan. 2015.
[51] M. T. Hussein, "A review on vision-based control of flexible manipulators," *Adv. Robot.*, vol. 29, no. 24, pp. 1575–1585, Dec. 2015.
[52] M. Parisa Beham, S. M. M. Roomi, and V. Kapileshwaran, "Robust face recognition using automatic pose clustering and pose estimation," in *Proc. 5th Int. Conf. Adv. Comput. (ICoAC)*, Dec. 2013, pp. 51–55.
[53] Y.-C. Chen, V. M. Patel, P. J. Phillips, and R. Chellappa, "Dictionary-based face recognition from video," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2012, pp. 766–779.
[54] P. Tome, J. Fierrez, R. Vera-Rodriguez, and D. Ramos, "Identification using face regions: Application and assessment in forensic scenarios," *Forensic Sci. Int.*, vol. 233, nos. 1–3, pp. 75–83, Dec. 2013.
[55] X. Hong, G. Zhao, and M. Pietikainen, "Pose estimation via complex-frequency domain analysis of image gradient orientations," in *Proc. 22nd Int. Conf. Pattern Recognit.*, Aug. 2014, pp. 1740–1745.
[56] P. Khandelwal, P. Swarnalatha, N. Bisht, and S. Prabu, "Detection of features to track objects and segmentation using GrabCut for application in marker-less augmented reality," *Procedia Comput. Sci.*, vol. 58, pp. 698–705, Jan. 2015.
[57] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "DeepFace: Closing the gap to human-level performance in face verification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1701–1708.
[58] M. C. D. F. Macedo, A. L. Apolinário, Jr, and A. C. D. S. Souza, "Kinectfusion for faces: Real-time 3d face tracking and modeling using a kinect camera for a markerless ar system," *SBC J. Interact. Syst.*, vol. 4, no. 2, pp. 2–7, 2013.
[59] Z. A. Othman and A. N. Faisal, "Face recognition model for age invariant," *Basrah J. Agricult. Sci.*, vol. 41, no. 1, 2015.
[60] H. Li, D. Huang, J.-M. Morvan, Y. Wang, and L. Chen, "Towards 3D face recognition in the real: A registration-free approach using fine-grained matching of 3D keypoint descriptors," *Int. J. Comput. Vis.*, vol. 113, no. 2, pp. 128–142, Jun. 2015.
[61] L. A. Jeni, J. F. Cohn, and T. Kanade, "Dense 3D face alignment from 2D video for real-time use," *Image Vis. Comput.*, vol. 58, pp. 13–24, Feb. 2017.
[62] D. Cosker, P. Eisert, O. Grau, P. J. B. Hancock, J. McKinnell, and E.-J. Ong, "Applications of face analysis and modeling in media production," *IEEE MultimediaMag.*, vol. 20, no. 4, pp. 18–27, Oct. 2013.
[63] M.-P. Schapranow, "Security in EPCglobal networks," in *Real-time Security Extensions for EPCglobal Networks*. Cham, Switzerland: Springer, 2014, pp. 31–59.
[64] C. Fuchs, F. Neuhaus, and D. Paulus, "3D pose estimation for articulated vehicles using Kalman-filter based tracking," *Pattern Recognit. Image Anal.*, vol. 26, no. 1, pp. 109–113, Jan. 2016.
[65] G. T. Marx and G. W. Muschert, "Personal information, borders, and the new surveillance studies," *Annu. Rev. Law Social Sci.*, vol. 3, no. 1, pp. 375–395, Dec. 2007.
[66] W. Li, C. Han, X. Yan, and J. Liu, "Adaptive sequential Monte Carlo implementation of the PHD filter for multi-target tracking," in *Proc. 16th Int. Conf. Inf. Fusion (FUSION)*, Jul. 2013, pp. 23–29.
[67] M. J. Marin-Jimenez, A. Zisserman, M. Eichner, and V. Ferrari, "Detecting people looking at each other in videos," *Int. J. Comput. Vis.*, vol. 106, no. 3, pp. 282–296, Feb. 2014.
[68] T. Zhang and H. M. Gomes, "Technology survey on video face tracking," in *Proc. Imag. Multimedia Anal. Web Mobile World*, Mar. 2014, p. 90.
[69] M. S. M. Asaari, B. A. Rosdi, and S. A. Suandi, "Intelligent biometric group hand tracking (IBGHT) database for visual hand tracking research and development," *Multimedia Tools Appl.*, vol. 70, no. 3, pp. 1869–1898, Jun. 2014.
[70] W. Wójcik, K. Gromaszek, and M. Junisbekov, "Face recognition: Issues, methods and alternative applications," in *Face Recognition: Semisupervised Classification, Subspace Projection and Evaluation Methods*. Rijeka, Croatia: InTech, 2016.
[71] T. Baltrusaitis, P. Robinson, and L. Morency, "3D constrained local model for rigid and non-rigid facial tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 2610–2617.
[72] M. Kim, S. Kumar, V. Pavlovic, and H. Rowley, "Face tracking and recognition with visual constraints in real-world videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–8.
[73] M. Sun, Y. Liu, Z. Liu, and M. Zhang, *Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data*. Cham, Switzerland: Springer, 2015.
[74] S. Degli Esposti, "When big data meets dataveillance: The hidden side of analytics," *Surveill. Soc.*, vol. 12, no. 2, pp. 209–225, 2014.
[75] L. Zhang, D. Tjondronegoro, V. Chandran, and J. Eggink, "Towards robust automatic affective classification of images using facial expressions for practical applications," *Multimedia Tools Appl.*, vol. 75, no. 8, pp. 4669–4695, Apr. 2016.
[76] T. K. M. Lee, M. Belkhatir, and S. Sanei, "A comprehensive review of past and present vision-based techniques for gait recognition," *Multimedia Tools Appl.*, vol. 72, no. 3, pp. 2833–2869, Oct. 2014.
[77] K. Gates, *Our Biometric Future: Facial Recognition Technology and the Culture of Surveillance*. New York, NY, USA: NYU Press, 2011.
[78] Y. Yang and D. Ramanan, "Articulated human detection with flexible mixtures of parts," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 12, pp. 2878–2890, Dec. 2013.
[79] Y. Yun, M. H. Changrampadi, and I. Y. H. Gu, "Head pose classification by multi-class AdaBoost with fusion of RGB and depth images," in *Proc. Int. Conf. Signal Process. Integr. Netw. (SPIN)*, Feb. 2014, pp. 174–177.
[80] M.-H. Yang, J. Ho, and K.-C. Lee, "Video-based face recognition using probabilistic appearance manifolds," U.S. Patent 7 499 574, Mar. 3, 2009.
[81] S. Cheng, S. Zafeiriou, A. Asthana, and M. Pantic, "3D facial geometric features for constrained local model," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2014, pp. 1425–1429.

[82] A. H. Al-Hamami and N. H. Batty, "Face recognition technique based on artificial neural network and principal component analysis," *Int. J. Adv. Stud. Comput., Sci. Eng.*, vol. 5, no. 11, p. 85, 2016.
[83] M. Humphreys, "Developmental trajectory of familiar and unfamiliar face recognition in children: Evidence in support of experience," *Plymouth Student Scientist*, vol. 10, no. 1, pp. 281–291, 2017.
[84] S. Alghowinem, R. Goecke, M. Wagner, G. Parkerx, and M. Breakspear, "Head pose and movement analysis as an indicator of depression," in *Proc. Humaine Assoc. Conf. Affect. Comput. Intell. Interact.*, Sep. 2013, pp. 283–288.
[85] P. Tome, L. Blazquez, R. Vera-Rodriguez, J. Fierrez, J. Ortega-Garcia, N. Exposito, and P. Leston, "Understanding the discrimination power of facial regions in forensic casework," in *Proc. Int. Workshop Biometrics Forensics (IWBF)*, Apr. 2013, pp. 1–4.
[86] R. Verschae, J. Ruiz-del-Solar, and M. Correa, "Face recognition in unconstrained environments: A comparative study," in *Proc. Workshop Faces Real-Life Images, Detect., Alignment, Recognit. (ECCV)*, Marseille, France, Oct. 2008, pp. 1–12.
[87] Y. Yan, R. Subramanian, E. Ricci, O. Lanz, and N. Sebe, "Evaluating multi-task learning for multi-view head-pose classification in interactive environments," in *Proc. 22nd Int. Conf. Pattern Recognit.*, Aug. 2014, pp. 4182–4187.
[88] C. Mayer, M. Eggers, and B. Radig, "Cross-database evaluation for facial expression recognition," *Pattern Recognit. Image Anal.*, vol. 24, no. 1, pp. 124–132, Mar. 2014.
[89] M. Hassaballah and S. Aly, "Face recognition: Challenges, achievements and future directions," *IET Comput. Vis.*, vol. 9, no. 4, pp. 614–626, Aug. 2015.
[90] A. Yeredor, I. Dvir, G. Koren-Blumstein, and B. Lachover, "Method and apparatus for video frame sequence-based object tracking," U.S. Patent 7 436 887, Oct. 14, 2008.
[91] G. Huang and L.-S. Sun, "An access control framework for reflective middleware," *J. Comput. Sci. Technol.*, vol. 23, no. 6, pp. 895–904, Nov. 2008.
[92] X. Cao, Y. Wei, F. Wen, and J. Sun, "Face alignment by explicit shape regression," *Int. J. Comput. Vis.*, vol. 107, no. 2, pp. 177–190, Apr. 2014.
[93] P. J. Bailey, *The Reluctant Film Art of Woody Allen*. Lexington, KY, USA: University Press of Kentucky, 2016.
[94] X.-J. Yu, "A study on the editing frequencies trends for films emotion clips," *Int. J. Organizational Innov.*, vol. 9, no. 3, p. 40A, 2017.
[95] A. Jordt and R. Koch, "Direct model-based tracking of 3D object deformations in depth and color video," *Int. J. Comput. Vis.*, vol. 102, nos. 1–3, pp. 239–255, Mar. 2013.
[96] J. Foytik and V. K. Asari, "A two-layer framework for piecewise linear manifold-based head pose estimation," *Int. J. Comput. Vis.*, vol. 101, no. 2, pp. 270–287, Jan. 2013.
[97] R. Rosales and S. Sclaroff, "Combining generative and discriminative models in a framework for articulated pose estimation," *Int. J. Comput. Vis.*, vol. 67, no. 3, pp. 251–276, May 2006.
[98] T. Danisman and I. M. Bilasco, "In-plane face orientation estimation in still images," *Multimedia Tools Appl.*, vol. 75, no. 13, pp. 7799–7829, 2016.
[99] F. Wang and M. Lu, "Robust particle tracker via Markov chain Monte Carlo posterior sampling," *Multimedia Tools Appl.*, vol. 72, no. 1, pp. 573–589, Sep. 2014.
[100] Z. Hu, T. Matsuyama, and S. Nobuhara, "Cell-based visual surveillance with active cameras for 3D human gaze computation," *Multimedia Tools Appl.*, vol. 74, no. 11, pp. 4161–4185, Jun. 2015.
[101] J. Hexner and R. R. Hagege, "2D-3D pose estimation of heterogeneous objects using a region based approach," *Int. J. Comput. Vis.*, vol. 118, no. 1, pp. 95–112, May 2016.
[102] V. R. Karimi, P. S. C. Alencar, and D. D. Cowan, "A uniform approach for access control and business models with explicit rule realization," *Int. J. Inf. Secur.*, vol. 15, no. 2, pp. 145–171, Apr. 2016.
[103] V. Q. Nhat and G. Lee, "Illumination invariant object tracking with adaptive sparse representation," *Int. J. Control, Autom. Syst.*, vol. 12, no. 1, pp. 195–201, Feb. 2014.
[104] Z. Kalal, K. Mikolajczyk, and J. Matas, "Forward-backward error: Automatic detection of tracking failures," in *Proc. 20th Int. Conf. Pattern Recognit. (ICPR)*, Aug. 2010, pp. 2756–2759.
[105] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 1, Dec. 2001, p. 1.
[106] R. Afrouzian, H. Seyedarabi, and S. Kasaei, "Pose estimation of soccer players using multiple uncalibrated cameras," *Multimedia Tools Appl.*, vol. 75, no. 12, pp. 6809–6827, Jun. 2016.

FALEH ALQAHTANI (Member, IEEE) received the B.Eng. and M.Eng. degrees from RMIT University, in 2011 and 2013, respectively, and the Ph.D. degree from the Queensland University of Technology, in 2019. His current research interests include artificial intelligence, object detection, face tracking, image processing, and computer vision.

JASMINE BANKS received the B.Eng. degree in electronics and information technology and the Ph.D. degree from the Queensland University of Technology, in 1993 and 2000, respectively. She is currently a Lecturer with the School of Engineering Systems, Queensland University of Technology. Her research interests include artificial intelligence, image processing, computer hardware, and electrical and electronic engineering.

VINOD CHANDRAN (Senior Member, IEEE) received the Ph.D. degree in electrical and computer engineering from Washington State University, in 1990. He retired as a Professor with the Queensland University of Technology (QUT), Brisbane, QLD, Australia, in 2016. He has authored or coauthored over 200 international journal articles and conference papers. His research interests include span signal processing, image processing, and machine learning—applied to biometrics and biomedical systems.

JINGLAN ZHANG received the Ph.D. degree in information technology from the Queensland University of Technology, in 2003. She is currently a Senior Lecturer with the Queensland University of Technology. Her research interests include visual and acoustic information (graphics, images, and sound) processing and retrieval, big data analysis and visualization, computer–human interaction, escience, software engineering, mobile and web applications, artificial intelligence, and information systems.