applied optics

Three-dimensional facial digitization using advanced digital image correlation

Hieu Nguyen,[1] Hien Kieu,[2] Zhaoyang Wang,[1,*] and Hanh N. D. Le[3]

[1]Department of Mechanical Engineering, The Catholic University of America, Washington, DC 20064, USA
[2]Department of Computer Science, Georgia Institute of Technology, Atlanta, Georgia 30332, USA
[3]Department of Electrical and Computer Engineering, Johns Hopkins University, Baltimore, Maryland 21218, USA
*Corresponding author: wangz@cua.edu

Received 8 November 2017; revised 23 January 2018; accepted 21 February 2018; posted 22 February 2018 (Doc. ID 312944); published 15 March 2018

Presented in this paper is an effective technique to acquire the three-dimensional (3D) digital images of the human face without the use of active lighting and artificial patterns. The technique is based on binocular stereo imaging and digital image correlation, and it includes two key steps: camera calibration and image matching. The camera calibration involves a pinhole model and a bundle-adjustment approach, and the governing equations of the 3D digitization process are described. For reliable pixel-to-pixel image matching, the skin pores and freckles or lentigines on the human face serve as the required pattern features to facilitate the process. It employs feature-matching-based initial guess, multiple subsets, iterative optimization algorithm, and reliability-guided computation path to achieve fast and accurate image matching. Experiments have been conducted to demonstrate the validity of the proposed technique. The simplicity of the approach and the affordable cost of the implementation show its practicability in scientific and engineering applications.

OCIS codes: (110.3010) Image reconstruction techniques; (110.6880) Three-dimensional image acquisition; (120.6650) Surface measurements, figure; (150.6910) Three-dimensional sensing; (170.3010) Image reconstruction techniques.

https://doi.org/10.1364/AO.57.002188

1. INTRODUCTION

Three-dimensional (3D) facial digitization is a process of generating a high-quality 3D model of the human face, and it has become a considerably attractive research area in a variety of fields, such as computer vision, optics, medical imaging, and experimental mechanics. Notable applications include, but are not limited to, face recognition, realistic animation for movies and video games, and cosmetic surgery. The existing popular measurement techniques suitable for the 3D facial digitization include the time-of-flight method, laser scanning method, moiré method, interferometry method, photogrammetry method, digital image correlation method, fringe projection or structured illumination method, etc. [1–9]. These methods typically involve using complicated measurement systems and procedures. Consequently, researchers and engineers have a great interest in exploring direct and easy schemes such as the human-eye-based stereo vision technique for accurate 3D facial digitization.

Recently, there are reported research activities that use face models and a classical reconstruction method, shape-from-shading, to reconstruct the 3D human face from a single image or from multiple images [10,11]. These techniques are convenient to use because they generally do not require camera calibration and instead rely on various face models to reconstruct many others. For the same reason, however, they cannot yield high accuracy for facial digitization, and the results are used only for limited tasks of face recognition. Another type of method that also requires 3D face templates in the reconstruction process is a deformation-based method [12–14]. The basic idea of the methods in this category is to generate the 3D face images by using a deformable face model; this involves deforming the 3D face model with 2D face images and a set of feature-point coordinates. Even though the deformation-based method can achieve 3D face imaging directly, the algorithms are complex in practice because the face alignment process must be very accurate yet is also very challenging. Furthermore, the lack of sufficient feature points in the regions of interest brings weaknesses to this method.

Using artificial markers or patterns either painted or projected on the human face is a common way to facilitate the stereo-vision-based 3D face digitization [15,16]. The techniques in this category represent an industry-standard scheme, but such an approach irritates the target person and can therefore notably affect the digitization performance. To cope with this issue, infrared projectors and infrared cameras are often adopted [17], and the drawback is the restricted resolution.

The scheme of using two or more images acquired from different views of an object to reconstruct its 3D model is highly demanded in practice. Such a technique may use one, two, or multiple cameras to capture the required two or more images. In the case of using a single camera, the corresponding technique employs a shape-from-motion algorithm where the video sequence captured by the camera during the motion of either the camera or the object provides the required images [18,19]. A major problem with this single-camera approach is the low 3D imaging accuracy. In regard to the techniques using multiple images, a notable work reported by Bradley *et al.* [20] shows that they can produce accurate 3D face images using neither a facial template nor painted or projected patterns. However, their technique requires a complicated setup with 14 high-definition cameras and nine large LED light fixtures. Another noteworthy attempt aims at using a low-cost stereo vision system to construct a 3D face model without calibrating the system, but the technique is subject to low accuracy [21].

In this paper, an advanced, accurate, easy-to-implement, and stereo-vision-based digital image correlation (DIC) technique for the 3D digitization of a human face is presented. The technique uses only two regular digital cameras, and it neither relies on active light illumination such as laser or structured light nor requires artificial patterns to be projected or painted on the human face. An illustration of the proposed 3D facial digitization system is shown in Fig. 1, where it can be seen that the basic principle of the technique is binocular stereo imaging. Unlike the existing stereo vision techniques, the proposed technique employs an advanced novel DIC scheme to achieve accurate 3D facial digitization from two images. It includes two primary crucial steps: (1) calibrate the cameras in advance to obtain the geometry information of the system; (2) perform matching of pixel points to link the same physical points between the two facial images. The technical details are elaborated in the following two sections.

2. PRINCIPLE OF THE 3D COORDINATE DETERMINATION AND CAMERA CALIBRATION

The proposed technique determines the 3D coordinates of points from system triangulation, so it requires obtaining the geometry information of the system. Figure 2 is a schematic

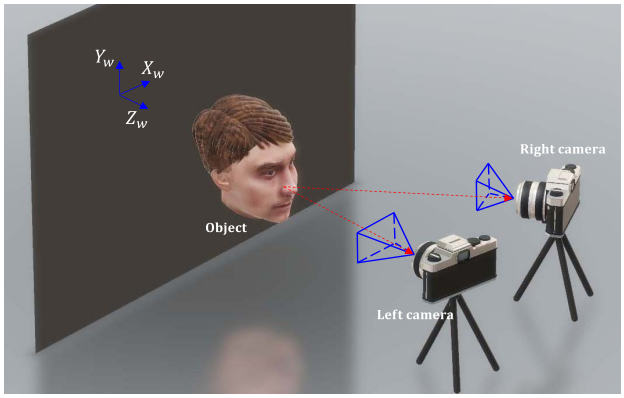

Fig. 1. Illustration of the 3D facial digitization system.

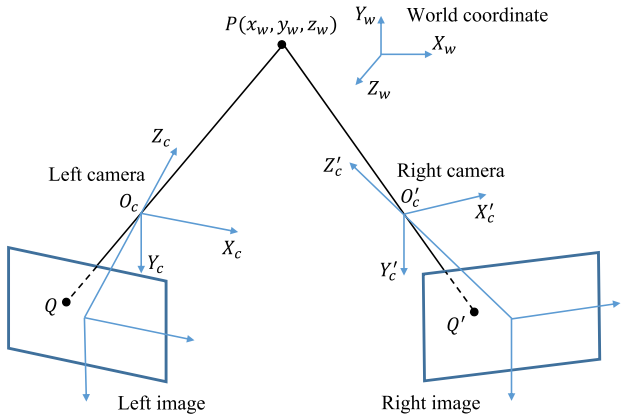

Fig. 2. Schematic of the binocular stereo imaging.

illustration of the technique. In the figure, O_c and O_c' are the optical centers of the left and right cameras, and an arbitrary physical point P is imaged as point Q and point Q' in the image planes of the left and right cameras, respectively. The goal here is to accurately determine the 3D coordinates of point P with respect to a global or world coordinate system from points Q and Q'. In theory, the triangulation-based calculation of the 3D coordinates demands that the physical distances among P, Q, and Q' be known. Consequently, it requires a rigorous description of how a 3D physical point is imaged as a pixel point in a captured digital image.

Following Fig. 2, an arbitrary point (x_w, y_w, z_w) in the global or world coordinate system can be expressed as (x_c, y_c, z_c) in a camera coordinate system by using the following equation:

$$\begin{Bmatrix} x_c \\ y_c \\ z_c \end{Bmatrix} = \begin{bmatrix} R_{11} & R_{12} & R_{13} & T_1 \\ R_{21} & R_{22} & R_{23} & T_2 \\ R_{31} & R_{32} & R_{33} & T_3 \end{bmatrix} \begin{Bmatrix} x_w \\ y_w \\ z_w \\ 1 \end{Bmatrix}. \tag{1}$$

In the equation, R and T components, named camera extrinsic parameters, indicate the rotation and translation parameters that transform the world coordinate system to the camera coordinate system. In the imaging plane of the camera, the pixel location (u, v) of the point can be described by a pinhole model as

$$\begin{Bmatrix} u \\ v \\ 1 \end{Bmatrix} = \frac{1}{z_c} \begin{bmatrix} \alpha & \gamma & u_0 \\ 0 & \beta & v_0 \\ 0 & 0 & 1 \end{bmatrix} \begin{Bmatrix} x_c \\ y_c \\ z_c \end{Bmatrix} = \begin{bmatrix} \alpha & \gamma & u_0 \\ 0 & \beta & v_0 \\ 0 & 0 & 1 \end{bmatrix} \begin{Bmatrix} x_{cn} \\ y_{cn} \\ 1 \end{Bmatrix}, \tag{2}$$

where $x_{cn} = x_c/z_c$, $y_{cn} = y_c/z_c$, α and β are the horizontal and vertical distances from the lens to the imaging plane in pixel unit, γ is a skew factor, and (u_0, v_0) are the coordinates of the principal point. These five parameters are often called camera intrinsic parameters.

By using a camera calibration target where the 3D world coordinates of the control points (such as the corner points of a checkerboard) are known, the camera intrinsic and extrinsic parameters, as well as the lens distortion parameters that must be considered in practice, can be determined from a

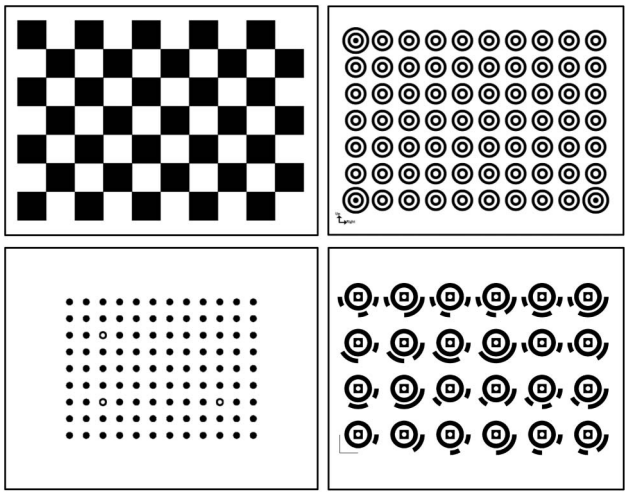

Fig. 3. Example of four commonly used camera calibration targets.

bundle-adjustment-based camera calibration process. Figure 3 demonstrates an example of four commonly used planar calibration targets. Camera calibration has been a very active research topic in the last three decades. The adopted algorithms can be found in Refs. [22,23].

With a 3D imaging system that contains two separate cameras, Eq. (1) gives the following equation for a typical point (x_w, y_w, z_w) in the global or world coordinate,

$$\begin{Bmatrix} x_{cn}z_c \\ y_{cn}z_c \\ z_c \end{Bmatrix} = \begin{bmatrix} R_{11} & R_{12} & R_{13} & T_1 \\ R_{21} & R_{22} & R_{23} & T_2 \\ R_{31} & R_{32} & R_{33} & T_3 \end{bmatrix} \begin{Bmatrix} x_w \\ y_w \\ z_w \\ 1 \end{Bmatrix}$$

$$\begin{Bmatrix} x'_{cn}z'_c \\ y'_{cn}z'_c \\ z'_c \end{Bmatrix} = \begin{bmatrix} R'_{11} & R'_{12} & R'_{13} & T'_1 \\ R'_{21} & R'_{22} & R'_{23} & T'_2 \\ R'_{31} & R'_{32} & R'_{33} & T'_3 \end{bmatrix} \begin{Bmatrix} x_w \\ y_w \\ z_w \\ 1 \end{Bmatrix}, \quad (3)$$

where the terms with a $'$ symbol are associated with the second camera. In Eq. (3), the extrinsic parameters R, T, R', and T' are acquired from camera calibrations in advance; and (x_{cn}, y_{cn}) and (x'_{cn}, y'_{cn}) are obtained from the two captured images using Eq. (2) and the lens distortion model. As a result, there are in total six equations as shown and five unknowns: x_w, y_w, z_w, z_c, and z'_c. Because Eq. (3) is an overdetermined linear system, the desired (x_w, y_w, z_w) can be directly obtained from its linear least-squares solutions.

3. IMAGE MATCHING

In the aforementioned governing equation, Eq. (3), for the 3D facial digitization, (x_{cn}, y_{cn}) and (x'_{cn}, y'_{cn}) must correspond to the same physical point (x_w, y_w, z_w). This requires matching the points in one image (the reference) with their corresponding positions in another image (the target) with subpixel accuracies. To comply with such a matching request, the human face must have sufficient texture patterns to ensure a correct detection of the pixel correspondences. Without painting or projecting artificial patterns on the face, the proposed technique relies on utilizing the natural textures, such as skin pores, freckles, and lentigines.

Solely depending on natural textures for the 3D facial digitization is challenging in practice. To tackle the difficulties, the proposed matching scheme comprises two steps: the first is feature-based matching, and the second is area-based matching.

A. Feature-Based Matching

The most well-known feature-based matching method is the scale invariant feature transform (SIFT) algorithm [24]. It is known to be very robust in handling variances in lighting, image rotation, translation, and scaling. Therefore, the SIFT matching scheme, including the SIFT algorithm and the relevant random sample consensus (RANSAC) algorithm [25,26], is employed in the proposed technique to carry out the first step of the image matching task.

The SIFT matching scheme can extract many features of interest from each face image, such as the skin pores and freckles, and build a unique descriptor for each feature point. The feature points between the reference and target images can then be compared by using their descriptors to find the best matching pairs. The feature-based matching can detect matching over the entire face image, and therefore can easily cope with the issue of geometric discontinuities, including occlusions and shadows.

The SIFT matching process is very fast; however, it detects matching only at tens or hundreds of discrete points on the facial images and cannot produce dense and full-field matching results. Figure 4 gives an example of using the SIFT method to extract feature points and detect matching between two facial images.

Following the feature-based matching process, an area-based matching process is conducted to expand the image matching to many more points with substantially enhanced accuracies.

B. Area-Based Image Matching

Based on the results provided by the feature-based SIFT matching process, the area-based matching process aims to yield matching results at every possible point with high accuracy. To speed up the analysis in practice, the area-based image matching usually does not run on every pixel but on pixels at a step size of N, and the full-field matching results are subsequently obtained by interpolation from those grid pixels. The step size is usually set to two or three and can be set to larger for a faster processing speed with reduced accuracy.

Unlike the feature-based algorithm, the area-based matching algorithm detects the same point in two different images

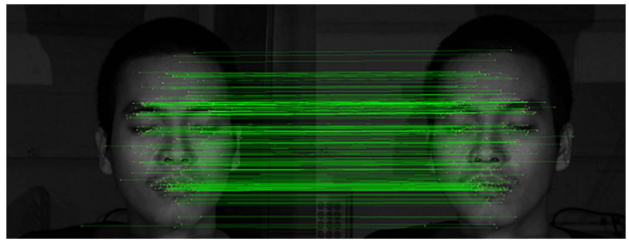

Fig. 4. SIFT matching on captured facial images.

through searching for the same group of pixels surrounding the interrogated point. The algorithm is described below.

For a pixel $(\tilde{u}_0, \tilde{v}_0)$ in the reference image to be matched, a square subset region of $(2M+1)\times(2M+1)$ pixels with its center located at $(\tilde{u}_0, \tilde{v}_0)$ is selected as the reference subset. The corresponding subset in the target image (i.e., the target subset) should be a homography transformation of the reference subset because they are the same region captured by two cameras from different positions and directions. Denoting the disparity between the centers of the two matching subset patterns as (ξ, η), the transformation function for the entire reference and target subsets can be expressed as [27,28]

$$\begin{aligned}\tilde{u}_i' &= \tilde{u}_i + \xi + \Delta_u(\xi_u + \xi_{uu}\Delta_u + \xi_{uv}\Delta_v) + \Delta_v(\xi_v + \xi_{vv}\Delta_v)\\ \tilde{v}_j' &= \tilde{v}_j + \eta + \Delta_u(\eta_u + \eta_{uu}\Delta_u + \eta_{uv}\Delta_v) + \Delta_v(\eta_v + \eta_{vv}\Delta_v),\end{aligned} \tag{4}$$

where i and j range from $-M$ to M, $\Delta_u = \tilde{u}_i - \tilde{u}_0$, $\Delta_v = \tilde{v}_j - \tilde{v}_0$, and ξ, ξ_u, ξ_v, ξ_{uu}, ξ_{vv}, ξ_{uv}, η, η_u, η_v, η_{uu}, η_{vv}, and η_{uv} are the transformation parameters. The determination of these parameters can be achieved by minimizing the least-squares-based correlation coefficient defined as [29]

$$C(\boldsymbol{p}) = \frac{1}{(2M+1)^2}\sum_{i=-M}^{M}\sum_{j=-M}^{M}[\zeta_{ij}(\boldsymbol{p})]^2, \tag{5}$$

where $\boldsymbol{p} = \{\xi, \xi_u, \xi_v, \xi_{uu}, \xi_{vv}, \xi_{uv}, \eta, \eta_u, \eta_v, \eta_{uu}, \eta_{vv}, \eta_{uv}, a, b\}^{\mathrm{T}}$, and $\zeta_{ij}(\boldsymbol{p}) = af(\tilde{u}_i, \tilde{v}_j) + b - g(\tilde{u}_i', \tilde{v}_j')$. Here, a is a scale factor, b is an offset of intensity, and $f(\tilde{u}_i, \tilde{v}_j)$ and $g(\tilde{u}_i', \tilde{v}_j')$ indicate the intensity values at a pixel in the reference subset and the potential matching pixel in the target subset, respectively. The best estimate of the 14 unknowns, including 12 shape-transformation parameters and two intensity parameters (a and b), is established by minimizing $C(\boldsymbol{p})$. This can be iteratively carried out by applying the Levenberg–Marquardt or Gauss–Newton algorithm [30] to Eq. (5), which can yield the governing equation as

$$\boldsymbol{p}_{n+1} = \boldsymbol{p}_n - \left[\sum_{i=-M}^{M}\sum_{j=-M}^{M}(\boldsymbol{J}_{ij}\boldsymbol{J}_{ij}^{T})\right]^{-1}\cdot\sum_{i=-M}^{M}\sum_{j=-M}^{M}[\zeta_{ij}(\boldsymbol{p})\boldsymbol{J}_{ij}], \tag{6}$$

where $n = 0, 1, 2, \ldots$ indicates the iteration step, and $\boldsymbol{J}_{ij}$ is the Jacobian vector defined as follows:

$$\begin{aligned}\boldsymbol{J}_{ij} = [\partial\zeta_{ij}(\boldsymbol{p})/\partial\boldsymbol{p}] = -\{&g_{u_i'}, g_{u_i'}\Delta_u, g_{u_i'}\Delta_v, g_{u_i'}\Delta_u^2, g_{u_i'}\Delta_v^2,\\ &g_{u_i'}\Delta_u\Delta_v, g_{v_j'}, g_{v_j'}\Delta_u, g_{v_j'}\Delta_v, g_{v_j'}\Delta_u^2, g_{v_j'}\Delta_v^2, g_{v_j'}\Delta_u\Delta_v, -f_{ij}, -1\}^{\mathrm{T}}.\end{aligned} \tag{7}$$

In Eq. (7), f_{ij} denotes $f(\tilde{u}_i, \tilde{v}_j)$; $g_{u_i'} = [\partial g(\tilde{u}_i', \tilde{v}_j')/\partial\tilde{u}_i']$ and $g_{v_j'} = [\partial g(\tilde{u}_i', \tilde{v}_j')/\partial\tilde{v}_j']$ are the intensity gradients of the target subset at location $(\tilde{u}_i', \tilde{v}_j')$ in the x and y directions, respectively. In the iteration, the convergence tolerance can be set to 1×10^{-5} for each element of $\boldsymbol{p}$.

It can be seen from Eq. (4) that although all the pixels in each reference subset are located at integer-pixel locations, the iteration requires obtaining the intensities at noninteger-pixel or subpixel locations in the target subset (i.e., $g(\tilde{u}_i', \tilde{v}_i')$). For this reason, a subpixel interpolation process must be carried out. A series of fast and accurate recursive B-spline interpolation functions can typically be deployed [31].

The Gauss–Newton iteration process shown by Eq. (6) requires a good initial guess for the 14 unknowns of $\boldsymbol{p}$. From Eq. (4), it can be seen that three pairs of matching points can be utilized to solve for the six transformation parameters (ξ, ξ_u, ξ_v, η, η_u, and η_v) because the other six transformation parameters can be set to zeros. It is also noted that the scale factor a and intensity offset b in Eq. (5) can be set to one and zero, respectively, for the initial guess. It is evident that the initial-guess requirement on the three pairs of matching points can be fulfilled by the feature-based SIFT matching points previously obtained, to be referred to as "SIFT point (s)" hereafter in this paper.

1. Starting Point and Reliability-Guided Propagation for Matching

The proposed area-based matching analysis starts by randomly selecting a SIFT point in the reference image. This SIFT point and its two closest SIFT points, together with their matching points in the target image, can then be used to determine the initial guess of the aforementioned six transformation parameters (ξ, ξ_u, ξ_v, η, η_u, and η_v). The centroid of the three SIFT points in the reference image becomes the seed point of the matching process. As previously mentioned, the area-based matching process is conducted on the calculation grids instead of on every pixel point, so the starting point of the matching process is the grid point nearest to the seed point.

After the matching analysis of the starting pixel point using the iteration algorithm governed by Eq. (6), its 14 optimized parameters serve as the initial values for the processing of its four neighbor pixels on the calculation grid. The pixel that has the smallest correlation coefficient $C(\boldsymbol{p})$ becomes the next point to be propagated [32,33]. After that, the matching process continues to propagate, where the matching result of a newly analyzed pixel will serve as the initial guess for its nonprocessed grid neighbors. By doing so, the area-based matching process follows a path guided by the best matching. Figure 5 illustrates how the propagation of the area-based matching process works with a grid step size of 3. In the figure, the image on the left shows the computational grids in cyan and a selected SIFT point with its two closest SIFT points in purple. The blue dot indicates the centroid of the triangle formed by the three SIFT points. The image on the right shows the seed point in yellow and the starting point labeled with 1 in brown. The four points labeled with 2–5 in red are analyzed following the starting point, and the three points labeled with 6–8 in green are analyzed subsequently assuming the point labeled with 4 has

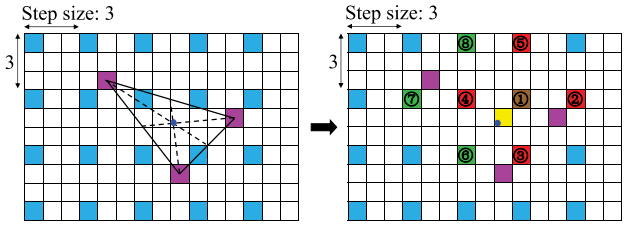

Fig. 5. Illustration of the propagation process in area-based image matching.

the best matching. After that, a point among the points labeled with 2, 3, and 5–8 that yields the best matching results will be the next point to be analyzed. This propagation process continues till decorrelation is detected or there are no more neighbor pixels to analyze.

If the area-based matching analysis of the starting point fails or the analysis propagation stops, the process will go back to find another SIFT point in no particular order and start a new analysis thread with a new starting point. If a grid point has been processed successfully in a previous matching thread, it will not be analyzed again in any subsequent matching thread to avoid redundant processing. The adopted image matching process following the path guided by the best matching coefficients can yield the largest number of reliable matching detections.

2. Multiple Subset Size

It is easily understood that the error of the area-based matching analysis is closely related to the selected subset size for any given images [34–36]. Therefore, the matching process requires that the subset must have an appropriate size and must contain sufficient intensity variations to ensure reliable matching detections. In this work, instead of using a fixed subset size, the proposed technique uses multiple subsets of various sizes for the image matching process. Particularly, subsets with sizes from 17×17 to 51×51 pixels are adopted.

The reason for this handling is that the features on the human face, such as skin pores, freckles, and lentigines, are often not uniformly distributed, and they can be evident in some regions but lacking in other regions. For instance, if a small region of interest in the captured images does not have sufficient intensity variations, using a small subset may fail to yield faithful results because of the possible subset decorrelation. In such a case, using a larger subset is helpful since more pixels are involved in the matching detection. Similarly, using a large subset may lead to failure in accurately detecting local 3D geometric details, whereas using a smaller subset can cope with this issue more effectively.

With the proposed scheme of using multiple subsets, there will be multiple matching results obtained for each interrogated pixel, and the one with the smallest correlation coefficient [shown in Eq. (5)] will be chosen as the final result. In this way, the area-based image matching result can reach the best possible accuracy.

4. EXPERIMENTS AND RESULTS

To clearly show the implementation procedure of the proposed technique, a flowchart is illustrated in Fig. 6. Because the key components of the area-based matching process are computationally intensive, they are particularly listed in the flowchart.

Three experiments have been carried out to demonstrate the validity of the proposed technique. In the experiments, two digital cameras with a resolution of 2048×1536 pixels are employed to form a generalized 3D vision and imaging setup. The cameras are calibrated prior to each experiment using the calibration technique described previously. Two planar calibration boards with 10×7 concentric-circle patterns with grid distances of 25.4 mm and 12.7 mm are adopted in the first two and the last experiments, respectively. The distances from the target to the cameras are adjusted according to the requirement of the field of view in each experiment. Particularly, the field width of view in the first two experiments is approximately 340 mm, and it is around 170 mm in the third experiment. The stereo images are captured simultaneously by the cameras and are converted to 8-bit grayscale images. The proposed approach is implemented in C++ and tested on a laptop with an Intel Core i7-3537U 2.0 GHz processor and 8.0 GB RAM.

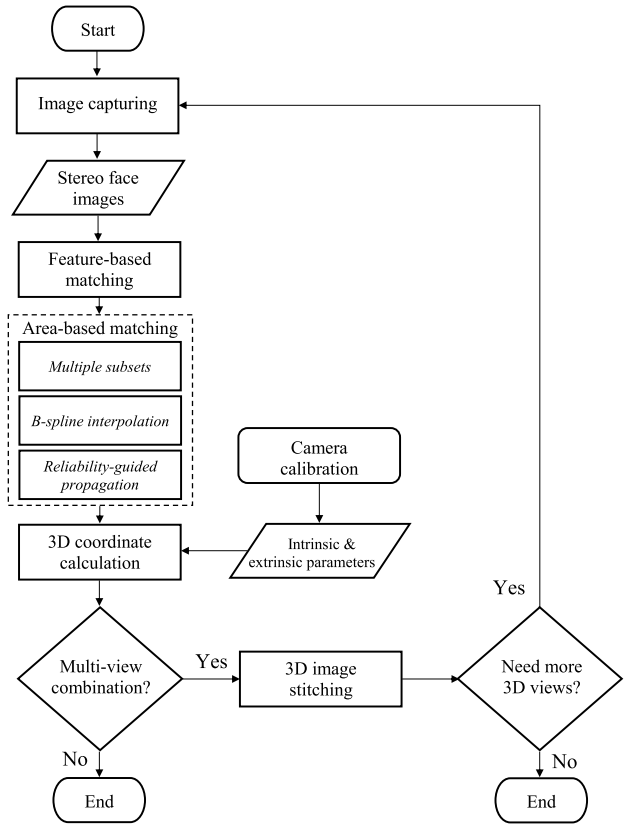

Fig. 6. Flowchart of the proposed technique.

The first experiment involves digitizing the face of an author of this paper. Figure 7 shows the results of three different facial expressions, in which the first two columns are the captured

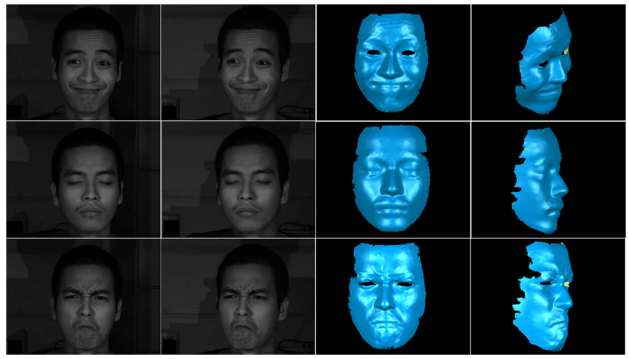

Fig. 7. Demonstration of 3D facial digitization results. The left two columns show images captured from two cameras and the right two columns show two selected views of the corresponding 3D digitization results.

images under passive illumination and the following two columns are two selected facial views after 3D digitization. It can be observed from the figure that the key features of the facial expressions have been well documented in the 3D digitization results. In the 3D digitization processing, the grid size selected in the area-based matching analysis is 2, and four subset sizes of 17×17, 21×21, 25×25, and 29×29 pixels are adopted. It is noted that reducing the grid size to 1 and increasing the number of subset sizes do not bring noticeable improvement in this experiment but lead to much longer computation time. On average, it takes 32 s to complete the 3D face digitization from each pair of images on the aforementioned laptop computer.

The second experiment has been conducted to demonstrate a comparison of the proposed technique with two existing well-known counterparts, namely the fringe projection profilometry (FPP) technique and the classic DIC technique. In theory, the fundamental principles of the three techniques are identical and all rely on pattern matching. An essential difference is that the FPP and the classic DIC techniques employ artificial patterns to make the matching process easier. Particularly, the FPP technique involves projecting multiple phase-shifted uniform fringe patterns on the objects, and the detected phase distributions in the captured images are linked to the phase distributions in the original images to build the matching correspondences. The projector technically serves as a reverse camera, so the original images are equivalent to being captured by the projector. The classic DIC technique uses the conventional area-based matching algorithm and does not employ schemes similar to those proposed in this paper; therefore, it must rely on utilizing artificial speckle patterns either painted or projected on the surfaces of objects. On the contrary, the proposed technique does not use any artificial patterns, indicating a substantial advance in practice. In the analysis, the proposed technique uses the same parameters as those adopted in the first experiment. Figure 8 shows the experimental results where the first image in each row is a representative captured image by using each technique, and the following images are selected views of the corresponding reconstructed 3D face model. In the figure, the images have been cropped for better display and visual comparison. The results reveal that although the proposed passive technique yields 3D geometrical details with quality slightly lower than the two active techniques, the key 3D features on the face are kept well in the 3D reconstructed model.

Even though the advantages and disadvantages of various 3D digitization and imaging techniques are highly application-dependent, the processing time, one of the key characteristics in real-world applications, is worth being discussed here. The time consumption of these techniques mainly involves the image acquisition time and the image analysis time. The acquisition time strongly depends on the frame rates of the camera and the projector as well as the synchronization between the cameras or between the camera and the projector. At present, the gap of the acquisition rates between the FPP technique and the stereo-vision-based technique has become very close thanks to the recent achievements on high-speed cameras and high-speed projectors. Consequently, the image analysis or processing time becomes dominant in practice. In general, the FPP technique can provide a fast processing speed with high analysis accuracy because it determines the 3D coordinates using fast phase extraction and simple yet rigorous governing equations. On the contrary, the stereo-vision-based techniques, including the conventional DIC technique and the proposed technique,

Fig. 8. Examples of 3D facial digitization using the FPP, the classic DIC, and the proposed techniques.

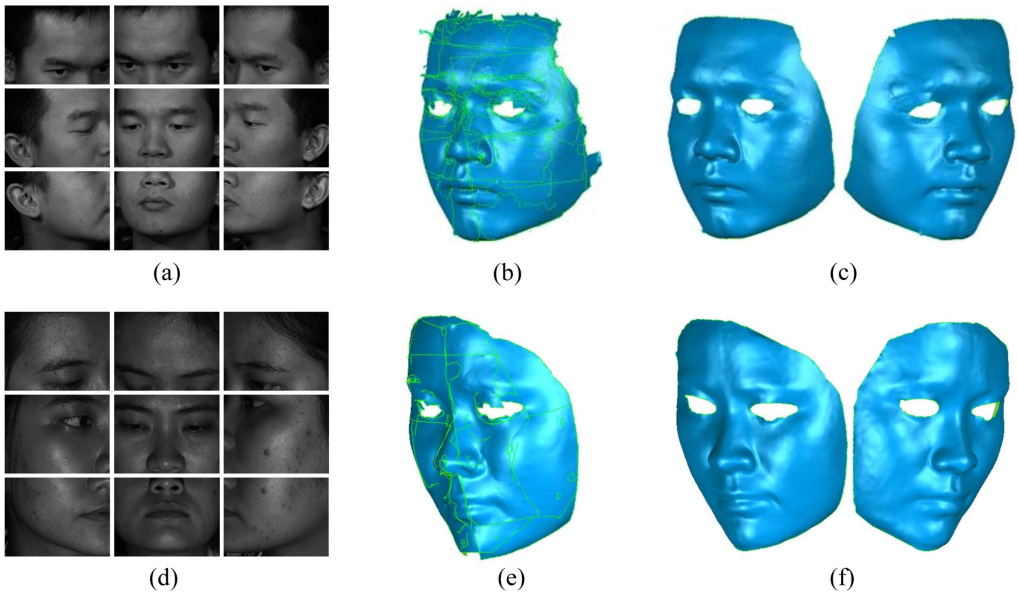

Fig. 9. 3D human face digitization from multiple captures.

always require a computation time that is highly dependent on the analysis accuracy since the processing involves pixel-to-pixel and area-to-area image matching with complex algorithms. In the second experiment, it on average takes the FPP, the classic DIC, and the proposed techniques approximately 0.3, 5, and 32 s to process the images for the 3D facial digitization.

The third experiment uses the same cameras, but with zoom lenses and a small calibration target to capture smaller regions of two faces, one of which is male and the other is female. As shown in Figs. 9(a) and 9(d), the face region is divided into nine smaller regions for the reconstruction of the 3D images. In this experiment, a grid size of 2 and four subset sizes of 21×21, 31×31, 41×41, and 51×51 pixels are adopted in the area-based matching process. The selection of the subset sizes is related to the magnified textures in the facial images. For the processing, it takes an average of 51 s to complete the 3D digitization of each view. Figures 9(b) and 9(e) show the stitched results of 3D digitization, where the stitching process is based on performing matching of the images captured by the same cameras. For instance, there are many pixels in the first image captured by the left (right) camera that can be matched to their corresponding pixels in the second image captured by the left (right) camera. These pixels directly contribute to the stitching of the 3D results built from the first and second pairs of images. It should be pointed out that discarding overlapping data and cleaning redundant or disconnected points are also necessary for the stitching process. Two selected views of the combined 3D digitization results of the faces are shown in Figs. 9(c) and 9(f), respectively. Through visually comparing the results to those in the first and second experiments, it can be seen that due to the higher resolution of the captured images, the 3D digitization accuracy has been notably increased.

The third experiment achieves two primary goals. One is to show that high-resolution 3D facial digitization is possible in the case that only cameras with relatively low resolution are available. Another goal is to qualitatively demonstrate the accuracy of 3D facial digitization. It is noteworthy that a rigorous accuracy assessment of the 3D facial digitization is beyond the capability of the authors because ideal references of truly accurate 3D face models are not available. In addition, since 3D results of lower accuracy may present smoother data than the ones of higher accuracy, a quantitative comparison is not carried out here to avoid confusion. Nevertheless, because the stitching process can easily fail if the accuracy of the 3D digitization is not sufficiently high, being able to stitch multiple 3D images together is an alternative and effective way to validate the accuracy of the proposed technique.

5. CONCLUSION

3D imaging technically reveals more qualitative and quantitative information than its 2D counterpart, and thus can better reflect the nature of the world. Currently, 3D imaging techniques have emerged as an important tool in many fields, where 3D facial digitization is a very prevalent application. Compared to other 3D imaging techniques for facial digitization, the stereo-vision-based techniques have a number of advantages, such as direct sensing, easy implementation, and broad imaging range. In this paper, an advanced robust DIC technique to acquire accurate 3D facial digitization is presented. The technique is based on stereo-vision imaging, and it combines an accurate camera calibration scheme and an enhanced image matching process to achieve accurate 3D coordinate measurements. Unlike other techniques that rely on using artificial projected or painted patterns, the proposed passive approach utilizes the natural textures on human face, such as skin pores, freckles, and lentigines, to perform the 3D facial digitization.

In the proposed approach, the camera calibration process adopts an advanced bundle-adjustment scheme that uses a sophisticated lens-distortion model and a frontal-image transformation and alignment method to achieve high-accuracy calibration. In the meantime, the image matching process employs

a feature-based SIFT matching algorithm to perform the initial matching detection and a robust area-based matching method to carry out the final image matching task with high accuracies, in which the schemes of multiple subsets, reliability-guided matching path, and recursive B-spline interpolation are deployed. The validity and practicality of the proposed approach have been verified by different experiments of the 3D facial digitization. In addition, a comparison with two popular active 3D imaging techniques has been conducted to demonstrate the effectiveness of the proposed passive approach.

As technologies evolve, the stereo-vision-based techniques have gained considerable interest in both academia and industry. This paper has demonstrated that the advanced DIC technique can provide results close to those offered by active 3D imaging techniques for 3D facial digitization (such as the fringe projection technique [37,38]). Despite its limitation in terms of computational complexity, the proposed technique is expected to help broaden the applications of passive 3D vision, imaging, and sensing.

Acknowledgment. The authors sincerely thank Ms. Hien Nguyen in the Department of Biomedical Engineering at the Catholic University of America for allowing use of her images in the experiments.

REFERENCES

1. T. Bakirman, M. Gumusay, H. Reis, M. Selbesoglu, S. Yosmaoglu, M. Yaras, D. Seker, and B. Bayram, "Comparison of low cost 3D structured light scanners for face modeling," Appl. Opt. **56**, 985–992 (2017).
2. X. Fan, C. Zhou, S. Wang, C. Li, and B. Yang, "3D human face reconstruction based on band-limited binary patterns," Chin. Opt. Lett. **14**, 081101 (2016).
3. F. Mohammadi, K. Madanipour, and A. Rezaie, "Accuracy enhancement of 3D profilometric human face reconstruction using undecimated wavelet analysis," Appl. Opt. **51**, 3120–3131 (2012).
4. O. Ebers, T. Ebers, M. Plaue, T. Raduntz, G. Barwolff, and H. Schwandt, "Study on three-dimensional face recognition with continuous-wave time-of-flight range cameras," Opt. Eng. **50**, 063201 (2011).
5. R. Wu, Y. Chen, Y. Pan, Q. Wang, and D. Zhang, "Determination of three-dimensional movement for rotary blades using digital image correlation," Opt. Lasers Eng. **65**, 38–45 (2015).
6. L. Kovacs, A. Zimmermann, G. Brockmann, M. Guhring, H. Baurecht, N. A. Papadopulos, K. Schwenzer-Zimmerer, R. Sader, E. Biemer, and H. F. Zeilhofer, "Three-dimensional recording of the human face with a 3D laser scanner," J. Plast. Reconstr. Aesthet. Surg. **59**, 1193–1202 (2006).
7. J. Espinosa, J. Perez, B. Ferrer, and D. Mas, "Method for targetless tracking subpixel in-plane movements," Appl. Opt. **54**, 7760–7765 (2015).
8. T. Nguyen, G. Nehmetallah, D. Tran, A. Darudi, and P. Soltani, "Fully automated, high speed, tomographic phase object reconstruction using the transport of intensity equation in transmission and reflection configurations," Appl. Opt. **54**, 10443–10453 (2015).
9. T. Nguyen, V. Bui, V. Lam, C. B. Raub, L. Chang, and G. Nehmetallah, "Automatic phase aberration compensation for digital holographic microscopy based on deep learning background detection," Opt. Express **25**, 15043–15057 (2017).
10. I. Kemelmacher-Shlizerman and R. Basri, "3D face reconstruction from a single image using a single reference face shape," IEEE Trans. Pattern Anal. Mach. Intell. **33**, 394–405 (2011).
11. M. Piotraschke and V. Blanz, "Automated 3D face reconstruction from multiple images using quality measures," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (IEEE, 2016), pp. 3418–3427.
12. G. Hu, F. Yan, J. Kittler, W. Christmas, C. Chan, Z. Feng, and P. Huber, "Efficient 3D morphable face model fitting," Pattern Recognit. **67**, 366–379 (2017).
13. A. Moeini, K. Faez, and H. Moeini, "Expression-invariant three-dimensional face reconstruction from a single image by facial expression generic elastic models," J. Electron. Imaging **23**, 053013 (2014).
14. P. Huber, Z. H. Feng, W. Christmas, J. Kittler, and M. Rätsch, "Fitting 3D morphable face models using local features," in *Proceedings of the IEEE International Conference on Image Processing* (IEEE, 2015), pp. 1195–1199.
15. Y. Wang, X. Huang, C. Lee, S. Zhang, Z. Li, D. Samaras, D. Metaxas, A. Elgammal, and P. Huang, "High resolution acquisition, learning and transfer of dynamic 3-D facial expressions," Comput. Graph. Forum **23**, 677–686 (2004).
16. B. Bickel, M. Botsch, R. Angst, W. Matusik, M. Otaduy, H. Pfister, and M. Gross, "Multi-scale capture of facial geometry and motion," in *Proceedings of ACM SIGGRAPH Transactions on Graphics (TOG)* (ACM, 2007).
17. H. Nguyen, Z. Wang, P. Jones, and B. Zhao, "3D shape, deformation, and vibration measurements using infrared Kinect sensors and digital image correlation," Appl. Opt. **56**, 9030–9037 (2017).
18. S. Lee, K. Park, and J. Kim, "A SfM-based 3D face reconstruction method robust to self-occlusion by using a shape conversion matrix," Pattern Recognit. **44**, 1470–1486 (2011).
19. M. Marques and J. Costeira, "3D face recognition from multiple images: a shape-from-motion approach," in *Proceedings of the IEEE International Conference on Automatic Face & Gesture Recognition* (IEEE, 2008), pp. 1–6.
20. D. Bradley, W. Heidrich, T. Popa, and A. Sheffer, "High resolution passive facial performance capture," in *Proceedings of ACM SIGGRAPH Transactions on Graphics (TOG)* (ACM, 2010).
21. M. S. Hossain, M. Akbar, and J. D. Starkey, "Inexpensive construction of a 3D face model from stereo images," in *Proceedings of the IEEE 10th International Conference on Computer and Information Technology* (IEEE, 2007), pp. 1–6.
22. Z. Zhang, "A flexible new technique for camera calibration," IEEE Trans. Pattern Anal. Mach. Intell. **22**, 1330–1334 (2000).
23. M. Vo, Z. Wang, L. Luu, and J. Ma, "Advanced geometric camera calibration for machine vision," Opt. Eng. **50**, 110503 (2011).
24. D. Lowe, "Distinctive image features from scale-invariant keypoints," Int. J. Comp. Vis. **60**, 91–110 (2004).
25. M. Fischler and R. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," Commun. ACM **24**, 381–395 (1981).
26. Z. Wang, H. Kieu, H. Nguyen, and M. Le, "Digital image correlation in experimental mechanics and image registration in computer vision: similarities, differences and complements," Opt. Lasers Eng. **65**, 18–27 (2015).
27. B. Pan, Z. Wang, and H. Xie, "Generalized spatial-gradient-based digital image correlation for displacement and shape measurement with subpixel accuracy," J. Strain Anal. Eng. **44**, 659–669 (2009).
28. B. Pan, K. Qian, H. Xie, and A. Asundi, "Two-dimensional digital image correlation for in-plane displacement and strain measurement: a review," Meas. Sci. Technol. **20**, 062001 (2009).
29. B. Pan, H. Xie, and Z. Wang, "Equivalence of digital image correlation criteria for pattern matching," Appl. Opt. **49**, 5501–5509 (2010).
30. Z. Wang, H. Nguyen, and J. Quisberth, "Audio extraction from silent high-speed video using an optical technique," Opt. Eng. **53**, 110502 (2014).
31. L. Luu, Z. Wang, M. Vo, T. Hoang, and J. Ma, "Accuracy enhancement of digital image correlation with B-spline interpolation," Opt. Lett. **36**, 3070–3072 (2010).
32. B. Pan, Z. Wang, and Z. Lu, "Genuine full-field deformation measurement of an object with complex shape using reliability-guided digital image correlation," Opt. Express **18**, 1011–1023 (2010).

33. H. Kieu, T. Pan, Z. Wang, M. Le, H. Nguyen, and M. Vo, "Accurate 3D shape measurement of multiple separate objects with stereo vision," Meas. Sci. Technol. **25**, 035401 (2014).
34. G. Hassan, C. MacNish, A. Dyskin, and I. Shufrin, "Digital image correlation with dynamic subset selection," Opt. Lasers Eng. **84**, 1–9 (2016).
35. Y. Zhou, C. Sun, and J. Chen, "Adaptive subset offset for systematic error reduction in incremental digital image correlation," Opt. Lasers Eng. **55**, 5–11 (2014).
36. R. Zhu, H. Xie, Z. Hu, L. Jiang, B. Guo, and C. Li, "Performances of different subset shapes and control points in subset-based digital image correlation and their applications in boundary deformation measurement," Appl. Opt. **54**, 1290–1301 (2015).
37. H. Nguyen, D. Nguyen, Z. Wang, H. Kieu, and M. Le, "Real-time, high-accuracy 3D imaging and shape measurement," Appl. Opt. **54**, A9–A17 (2015).
38. X. Su and Q. Zhang, "Dynamic 3-D shape measurement method: a review," Opt. Lasers Eng. **48**, 191–204 (2010).