

# Digital Longmen Project: A Free Walking VR System with Image-Based Restoration

Zeyu Wang<sup>1</sup>(✉), Xiaohan Jin<sup>1</sup>, Dian Shao<sup>1</sup>,  
Renju Li<sup>1</sup>, Hongbin Zha<sup>1</sup>, and Katsushi Ikeuchi<sup>2</sup>

<sup>1</sup> Key Laboratory of Machine Perception, Peking University, Beijing, China  
1200012927@pku.edu.cn

<sup>2</sup> Microsoft Research Asia, Beijing, China

**Abstract.** Located in China's ancient capital Luoyang, Longmen Grottoes are one of the finest examples of Buddhist stone carving art. Nowadays, many caves do not have public access due to heritage preservation. In order to let people appreciate these relics, we setup a VR system with smartphones and helmets based on scanned models and textures. Motion capture system is also utilized to make the viewpoint not fixed so that users can walk freely as if in the cave. Moreover, since some sculptures have been heavily damaged, we propose a digital restoration framework to enhance exhibition contents. The framework includes general and detailed restoration from a single old image by shape from shading and landmark driven mesh deformation respectively. In practice, we develop this system for the representative Middle Binyang Cave, with interactions such as gesture recognition exploited to provide satisfactory user experience, which can ease the conflict between tourism and preservation.

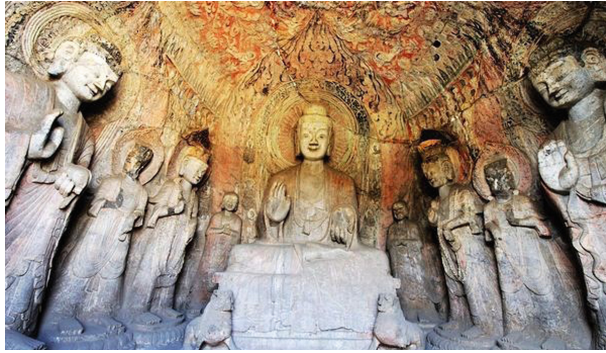
## 1 Introduction

Cultural heritage represents the splendid civilization our ancestors created, including outstanding and irreproducible historical relics, architecture, sites, and intangible cultural heritage. As masterpieces of both nature and human beings, they are invested with high values from historical, cultural, and scientific aspects. China is a country with proud history of five thousand years, and has 50 world cultural and/or natural sites approved by UNESCO since the 40th World Heritage Committee session. The Forbidden City, Mogao Caves, Longmen Grottoes and many others in the list demonstrate people's creativity, wisdom, artistic accomplishments, as well as ideologies, aesthetic customs, and religious beliefs, thus are valuable treasures of all mankind. Take Longmen Grottoes as an example, almost 110,000 Buddha statues were built from the late 5th century with exquisite stone carving crafts. Among them Middle Binyang Cave is one of the most well engraved, decorated, and preserved constructions. As a royal project Emperor Xuanwu of the Northern Wei Dynasty initiated to bring honor to his parents, this cave was built by more than 800 thousand workers in 24 years. The cave has a theme of Trikalea Buddhas (Buddhas of the Past, Present, and Future), and is a typical representative of ancient Chinese style with spectacular statues and colorfully textured relief sculpture.

However, Longmen Grottoes have experienced quite a few harsh times. The power of nature such as sunshine, damp, earthquakes, and weathering has made many parts deteriorated beyond recognition, not to mention unbridled sabotage, thefts, and smuggling from the beginning of the last century. Japanese scholar Tadashi Sekino wrote, “Most of the Buddha heads, as long as they can be dismantled, are cut off and sold to foreigners.” Middle Binyang Cave alone lost four Bodhisattva heads and two precious reliefs named *Processions of Emperor and Empress in Worship*, which are respectively collected in Tokyo National Museum, Osaka City Museum of Fine Arts, Metropolitan Museum of Art, and Nelson-Atkins Museum of Art. Since Longmen Grottoes are famous for colossal stone sculptures, physical restoration requires strenuous labor and is even likely to damage remaining relics. Recently, preservation and restoration utilizing digital technology were conducted with the cooperation from both academia and industry. Many caves were digitalized in 3D mesh format with texture using Leica laser scanners. Once we have 3D models, automatic and manual methods can be developed to merge the head elsewhere with the remaining body. It has been proved effective in Guyang Cave. Compared with repair and assembly in the real world, there is no doubt that the operations on 3D models can avoid potential damage and save human labor.

Considering many relics in Longmen Grottoes have been already damaged more or less, the administrative committee suggests that the flow of tourists should be controlled to prevent heritage from deterioration caused by human activities. Some caves including Middle Binyang Cave are officially closed, so that textures can be better protected from direct sunlight, although meanwhile general visitors are not allowed to enter. There are several attempts on solving the conflict between tourism and preservation. In the past few years, people were able to take a virtual tour based on panoramas shot from fixed viewpoints instead of onsite visits, but they suffered from the sense of restriction. For Mogao Caves in Dunhuang, a large spherical theater was built in 2014 at an expense of nearly 45 million dollars, and it also cost a lot on filmmaking. To overcome these limitations, we propose a free walking VR system for heritage exhibition based on scanned models and textures. Inspiringly, cheap smartphone displays become popular and competent at providing a user-centric, active, immersive, interactive approach to restored heritage exhibition.

In order to provide more complete contents for VR exhibition, digital restoration is necessary for damaged relics, especially stolen Buddha heads. By reconstructing these missing parts, it becomes possible to digitally show the merged model together. In fact, one of the biggest challenges for restoration is lack of archaeological evidence, but luckily, we find some valuable old photos taken over 70 years ago. Images can be used as a powerful and convincing reference for restoration as they record the shape and unique characteristics of the original statues. We exploit shape from shading and landmark driven mesh deformation for image-based restoration. Compared to merging parts from different statues, we note the variance between statues of the same figure. The case shows that 3D reconstruction from a single image can be applied to heritage restoration.



**Fig. 1.** A glance at Middle Binyang Cave, Longmen Grottoes. Some heads remain lost.

In this paper, we make the following contributions:

1. Since Middle Binyang Cave does not have access to public tourists, we setup a free walking VR environment on popular smartphones. Realtime user localization is solved using both inertial sensors and motion capture systems.
2. For exhibition contents, we propose a restoration framework based on old images, including shape from shading and Buddha face reconstruction method by Poisson-based deformation driven by shape priors and facial features.
3. We release a Unity3D-powered app on Android with user interaction for restoration and enhancement, which is going to be promoted in Longmen (Fig. 1).

## 2 Related Work

The field of digital cultural heritage adopts many techniques in computer vision and graphics. In particular, 3D reconstruction has been playing an important role in digital museum, relic archiving, and other archaeological research. The EU funded 3D MURALE project established a multimedia database for displaying historical remains in Sagalassos, Turkey [1]. Current 3D scanning methods mainly include laser scanners, structured light scanners, and multi-ocular cameras. For large scenes such as the Bayon Temple, ballon scanners were proposed to overcome the inaccessibility from the top, and the 3D model of high quality was fused under the constraints of data distortion, image motion, and ballon motion [2]. Material, texture, and reflection model are also frequently discussed. For example, texture pattern in Kyushu ancient tombs was extracted and enhanced using color modeling, which helped prove the archaeological guesses [3].

Based on the scanned models, some research focused more on restoration algorithms. For instance, the Forma Urbis Romae project at Stanford University recovered the urban map of ancient Rome by matching and stitching 1,186 textured

marble pieces with arbitrary shape [4]. Similarly, a team from Princeton University also digitalized and restored frescoes in the ruins of Akrotiri, Santorini, with a multi-way Iterative Closest Point (ICP) framework for organizing and stitching [5]. Curve-based interaction methods in a 3D pottery retrieval and classification system are useful for analysis and restoration as well [6]. Software tools for medieval manuscript recognition, sculpture shape analysis, and aging material modeling were developed by the graphics group at Yale University [7–9]. In this paper, we focus on tourist-oriented exhibition and Buddha head restoration.

**VR/AR Exhibition.** Virtual reality exhibition, including video mapping, telepresence, and Head Mounted Display (HMD), has brought tremendous changes in the way how people interact with the environment. As digital technology is entering the era of low cost and high efficiency, it becomes possible to adopt VR display in archaeology, art, and entertainment, such as movies, games, and theme parks [10]. Augmented reality systems have also been developed for various scenic areas technically considering geometric consistency and photometric consistency [11, 12].

**Face Reconstruction.** 3D face reconstruction from a single 2D image is a difficult problem with many previous research work, such as shape from shading, shape from texture, shape from silhouettes, shape from focus and using shape priors, though the results are far from the goal of high quality and realistic reproduction. For face reconstruction, the main ideas include shape from shading [13, 14] and 3D morphable models [15]. Shape from shading requires information about the reflectance properties and lighting. Using morphable models circumvents these requirements by representing input faces as combinations of hundreds of stored 3D models. But in Longmen dataset, the number of available 3D faces is very limited, so this algorithm does not apply. Our method performs Poisson-based mesh deformation following precise face landmark localization, which is novel and robust for our dataset, using only one 3D face template.

### 3 VR Environment Setup

Since Middle Binyang Cave is representative for stone carving art and does not have public access, we setup a virtual reality environment using scanned 3D models and textures. The mesh data is stored in 19 `.wrl` files, with texture data in 42 `.rgb` files for each connected region. Considering this VR system will ultimately be released on Samsung Galaxy Note 4 smartphone, the raw mesh data is downsampled using Geomagic Studio 11 to stay in accordance with limited computing and storage capabilities. The final size of mesh data imported into Unity3D is 31.6 MB, with 212,829 vertices and 409,917 facets. The texture mapping data is 250.6 MB and meets the workload requirements after testing.

#### 3.1 Localization by IMU

Current integrated VR devices, such as Oculus Rift, usually include modules of optics, display, sensing, and computing, which is costly for common consumers.

Therefore, we take advantage of smartphones that everyone has, and 20-dollar Baofeng helmets offer necessary lenses. In the aforementioned 3D environment, the viewpoint is initialized in the center of the cave, and what we see in the first place is Sakyamuni. Generally speaking, we need to estimate the user's motion and pose in VR systems by various sensors, thus the new viewpoint can be updated and stay consistent with real life customs. The Inertial Measurement Unit (IMU), including accelerators and gyros, is crucial for pose estimation. Since the displacement is integrated twice from the acceleration, while the angle is integrated only once from the angular velocity, gyro data is utilized in our system to avoid considerable error accumulation.

As for user pose estimation, denote the raw, pitch, and yaw angles at time  $t_1$  by  $\alpha_r(t_1)$ ,  $\alpha_p(t_1)$ , and  $\alpha_y(t_1)$  respectively. Then denote the raw, pitch, and yaw angular velocities at time  $t$  by  $\omega_r(t)$ ,  $\omega_p(t)$ , and  $\omega_y(t)$  respectively. Thus the new raw, pitch, and yaw angles at time  $t_2$ , i.e.,  $\alpha_r(t_2)$ ,  $\alpha_p(t_2)$ ,  $\alpha_y(t_2)$  can be calculated by:

$$\alpha_{r,p,y}(t_2) = \alpha_{r,p,y}(t_1) + \int_{t_1}^{t_2} \omega_{r,p,y}(t) dt \quad (1)$$

We implement these computation by C# script codes in Unity3D with the help of ALPSVR package, and the rotation of the viewpoint is measured by quaternions. As a form of extended complex numbers, quaternions make interpolation smoother, and avoid the problem of gimbal lock. Multiple rotations can be represented by quaternion multiplication. The viewpoint parameters in Unity3D will be updated after this computation. The Euler angles of roll  $\alpha_r$ , pitch  $\alpha_p$ , and yaw  $\alpha_y$  can be converted into quaternions as follows:

$$q = \begin{bmatrix} q_0 \\ q_1 \\ q_2 \\ q_3 \end{bmatrix} = \begin{bmatrix} \cos\left(\frac{\alpha_y}{2}\right) \\ 0 \\ 0 \\ \sin\left(\frac{\alpha_y}{2}\right) \end{bmatrix} \begin{bmatrix} \cos\left(\frac{\alpha_p}{2}\right) \\ 0 \\ \sin\left(\frac{\alpha_p}{2}\right) \\ 0 \end{bmatrix} \begin{bmatrix} \cos\left(\frac{\alpha_r}{2}\right) \\ \sin\left(\frac{\alpha_r}{2}\right) \\ 0 \\ 0 \end{bmatrix} \quad (2)$$

### 3.2 Localization Design of Free Walking

IMU-based head tracking can achieve good results, because the new viewing orientation is consistently updated by angular velocity integral. However, there are drawbacks such as error accumulation and viewpoint immobilization. Tourists cannot appreciate the statues from different perspectives despite horizontal displacement. Therefore, we exploit TenYoun motion capture system with reflective markers on the helmet to realize virtual tours of high precision. This passive system consists of 12 near infrared cameras of high sensitivity capturing the reflective markers stuck on the moving object in real time. Motion capture of high precision is achieved by reconstructing trajectories of marker and object motion. In our implementation, four fixed markers (two on the  $X$  axis, one on the  $Y$  and  $Z$  axes) are stuck on the Baofeng helmet as a rigid body, thus 3D rotational and translational transformation matrices can be computed and applied to the coordinates of the viewpoint.

**Triangulation in Multi-view Vision.** The principle of motion capture system tracking marker balls is triangulation. Depth information is lost when capturing 2D images of a 3D scene, but 3D information can be reconstructed when multiple cameras capture images of the same scene. In fact, a point in 3D space is identical to the intersection of extension lines from optical centers to imaging points. In this multi-view vision system, denote the calibrated parameter matrices of 12 cameras by  $\mathbf{M}_1, \mathbf{M}_2, \dots, \mathbf{M}_{12}$  and denote the imaging point coordinates of point  $(X_w, Y_w, Z_w)$  by  $(u_1, v_1), (u_2, v_2), \dots, (u_{12}, v_{12})$ , then we have:

$$\begin{bmatrix} u_i \\ v_i \\ 1 \end{bmatrix} = \mathbf{M}_i \begin{bmatrix} X_w \\ Y_w \\ Z_w \\ 1 \end{bmatrix}, \quad i = 1, 2, \dots, 12 \quad (3)$$

In the motion capture system, reflective markers have strong responses in images captured by near infrared cameras of high sensitivity, so their pixel locations can be easily determined by intensity thresholding. Since the number and relative position of markers are fixed, multi-view matching can be achieved by enumeration. Theoretically, only two calibrated cameras are enough for 3D reconstruction, but more cameras eliminate occlusion and bring higher precision. Therefore, a realtime human motion capture system usually consists of 12 or 24 cameras.

**3D Transformation Estimation.** There are four reflective markers stuck on the Baofeng helmet for VR display, because the user's 3D transformation matrix can be estimated by the trajectories of these markers. Denote the current world coordinates of these four markers by  $(X_1, Y_1, Z_1), (X_2, Y_2, Z_2), (X_3, Y_3, Z_3)$ , and  $(X_4, Y_4, Z_4)$ . Denote the world coordinates in the next frame by  $(X'_1, Y'_1, Z'_1), (X'_2, Y'_2, Z'_2), (X'_3, Y'_3, Z'_3)$ , and  $(X'_4, Y'_4, Z'_4)$ . Then the 3D transformation between two frames can be represented by a  $3 \times 3$  orthogonal rotation matrix  $\mathbf{R}$  and a vector of 3D translation  $\mathbf{t}$ :

$$\begin{bmatrix} X_j \\ Y_j \\ Z_j \\ 1 \end{bmatrix} = \begin{bmatrix} \mathbf{R} & \mathbf{t} \\ \mathbf{0}^T & 1 \end{bmatrix} \begin{bmatrix} X'_j \\ Y'_j \\ Z'_j \\ 1 \end{bmatrix}, \quad j = 1, 2, 3, 4 \quad (4)$$

Since the relative position of these four markers is fixed, and rigid transformation only has the degree of freedom of 6, two equations are theoretically sufficient for calculating  $\mathbf{R}$  and  $\mathbf{t}$ . Similarly, solving the over-determined equation set by least square method can improve the precision. Once we know the rigid transformation in real 3D space, we can apply the same transformation matrix to viewpoint in VR environment. Therefore, the user can appreciate the cave and experience consistent virtual tour as long as he or she is walking freely in the room equipped with our motion capture system.

**Camera Calibration.** From the above principle, we first needs to calibrate camera to get geometric parameters of the imaging model, which is generally the pinhole model to approximate the real nonlinear lens imaging model. We create four Descartes coordinate systems to describe the geometric relations in the imaging process: image pixel coordinate  $(u, v)$ , image physical coordinate  $(x, y)$ , camera coordinate  $(X_c, Y_c, Z_c)$ , and world coordinate  $(X_w, Y_w, Z_w)$ .

In the pinhole model, draw a line from an arbitrary point  $P$  in the 3D space to the camera's optical center, and their intersection  $p$  on the imaging plane is the imaging point. Denote the focal length by  $f$ . Let  $\mathbf{R}_0$  be a  $3 \times 3$  orthogonal rotation matrix,  $\mathbf{t}_0$  be a vector of 3D translation. Denote the proportion of physical unit to column and row pixel unit by  $\Delta x$  and  $\Delta y$  respectively. If the image pixel coordinate of camera's optical center is  $(u_0, v_0)$ , the image pixel coordinate of  $p$ ,  $(u, v)$ , can be calculated from the world coordinate of  $P$ ,  $(X_w, Y_w, Z_w)$ , as follows:

$$Z_c \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} = \begin{bmatrix} \frac{1}{\Delta x} & 0 & u_0 \\ 0 & \frac{1}{\Delta y} & v_0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} f & 0 & 0 & 0 \\ 0 & f & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} \mathbf{R}_0 & \mathbf{t}_0 \\ \mathbf{0}^T & 1 \end{bmatrix} \begin{bmatrix} X_w \\ Y_w \\ Z_w \\ 1 \end{bmatrix} \quad (5)$$

Put internal and external parameters together, and let  $\mathbf{M}$  be an arbitrary  $3 \times 4$  matrix. The objective of camera calibration is to determine this matrix, by capturing multiple images of unique calibration boards. It can be computed by least square method from the equation set of 3D-to-2D transformation. According to Eq. 5,  $\mathbf{M}$  multiplied by a non-zero constant still represents the relation between  $(X_w, Y_w, Z_w)$  and  $(u, v)$ , thus we might set an element to 1 as well. The degree of freedom is 11 in total, and we need at least 6 non-coplanar calibration points. All 12 calibrated cameras in the motion capture system should not be adjusted before running the VR display.

### 3.3 Display and Communications

Since the screen is very close to our eyes in VR display, there are convex lenses in the helmet between the screen and our eyes, but they also introduce image distortion, where the image magnification decreases when the distance to the optical axis increases. Barrel distortion correction based on Brown-Conrady model is usually used to counteract this effect [16].

$$(x_d, y_d) = (x_u, y_u) (1 + K_1 r^2 + K_2 r^4 + \dots) \quad (6)$$

where  $(x_d, y_d)$  is the coordinate in distorted images,  $(x_u, y_u)$  is the coordinate in undistorted images,  $(x_c, y_c)$  is the coordinate of the distortion center,  $K_n$  is the  $n^{\text{th}}$  parameter of radial distortion, and the distance from the pixel to the distortion center in undistorted images is  $r = \sqrt{(x_u - x_c)^2 + (y_u - y_c)^2}$ . The VR image flow delivered to the user from all viewpoints should be corrected against distortion in advance for people's perceptual comfort.



**Fig. 2.** The free walking VR environment using the motion capture system. The smartphone display is put in a helmet, on which four marker balls are stuck.

It remains an issue to transmit data from the motion capture system to the smartphone app in real time. There is a server for the motion capture system which returns current world coordinates of four marker balls. For communications between the server and the smartphone, they are connected to an identical WiFi under TCP/IP protocol, and then data can be transmitted in the form of Socket. We initialize marker positions in the VR environment and write a C# script listener to receive motion capture stream data from a certain IP address and port, thus estimated 3D transformation can be applied to the current viewpoint and new images can be rendered promptly. In our experiments, user’s free walking is enabled in a  $5\text{ m} \times 5\text{ m}$  area, with an adequate frame rate of more than 30 fps for rendered display (Fig. 2).

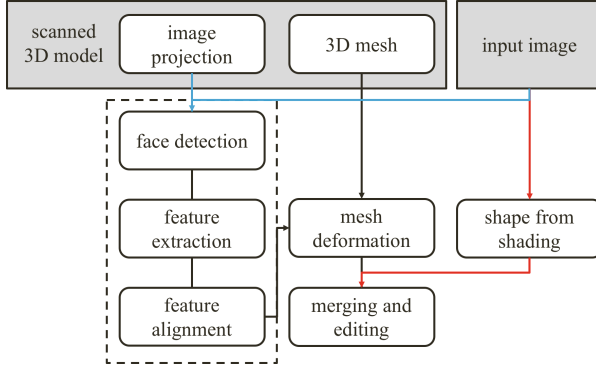
## 4 Heritage Restoration

Currently, damaged Buddha models are available through 3D scanning, and old pictures of complete Buddha can be collected from photographers. In order to restore digital models for exhibition contents based on images, we first try an improved version of shape from shading, in which we extract reflectance map by minimizing the local variation of log-reflectance and minimizing the global entropy of log-reflectance. Then we introduced our method which is inspired by mesh deformation algorithms [17, 18] in computer graphics field. We evaluate our method on existing statues. The proposed framework is shown in Fig. 3.

### 4.1 Shape, Illumination, and Reflectance from Shading

Shape from shading problem has been well surveyed since 1970s [19–21]. It deals with the recovery of shape from a gradual variation of shading in the image. The concept of “intrinsic image” comes up to complement the preprocessing of a natural image, and to decompose an image ( $I$ ) into its shading ( $S$ ) and reflectance ( $R$ ) components. The algorithm of Shape, Illumination, and Reflectance from





**Fig. 3.** The proposed restoration framework. Red arrows indicate the first method. Blue ones take two as input to the image processing module (dashed box), and then supervise mesh deformation. We can merge and edit these results in the end. (Color figure online)

Shading (SIRFS) [14], a generalization of an intrinsic image algorithm, infers the most-likely explanation of a single image, including the shape, surface normals, reflectance, shading, and illumination which produced that image. Shading ( $S$ ) is explicitly parametrized as a function of shape ( $Z$ ) and illumination ( $L$ ).

$$I = R + S(Z, L) \quad (7)$$

The following priors are the core of SIRFS: reflectance images tend to be piecewise smooth and low-entropy, surfaces tend to be isotropic and bend infrequently, and illumination tends to be natural. So the goal is to

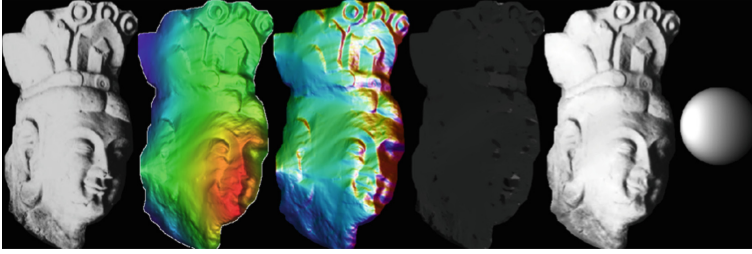
$$\min_{Z, L} g(I - S(Z, L)) + f(Z) + h(L) \quad (8)$$

The estimation of reflectance map is based on the assumption of smooth shading variation and low-entropy pattern influences. The smoothness level  $g_s(R)$  of the input image patch becomes higher when the shape variation is small and that is when the pattern reflects shading information. Furthermore, as a soft constraint, the color distribution of a reflectance map is limited. We consider this factor in two parts,  $g_e(R)$ : parsimony of reflectance by minimizing the global entropy of log-reflectance;  $g_a(R)$ : color similarities and preferences.

$$g(R) = \lambda_s g_s(R) + \lambda_e g_e(R) + \lambda_a g_a(R) \quad (9)$$

The prior on shape consists of three parts.  $f_k(Z)$ : shapes tend to be smooth so that the variation of mean curvature tends to be small;  $f_i(Z)$ : shapes are equally likely to face in different directions;  $f_c(Z)$ : shapes tend to face outward at the occluding contour.

$$f(Z) = \lambda_k f_k(Z) + \lambda_i f_i(Z) + \lambda_c f_c(Z) \quad (10)$$



**Fig. 4.** The outputs of SIRFS method, including the shape (height map), orientation of normals, reflection map, shading image and orientation of illumination.

For illumination, we assume it is spherical-harmonic illumination, and train a multivariate Gaussian model in MIT intrinsic images dataset.  $\mu_L$  and  $\Sigma_L^{-1}$  are the parameters of the Gaussian we learned.

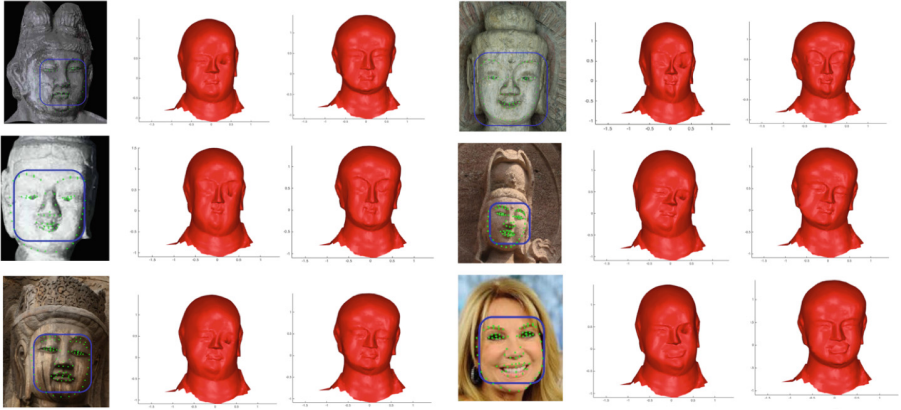
$$h(L) = \lambda_L(L - \mu_L)^T \Sigma_L^{-1}(L - \mu_L) \quad (11)$$

Because the Gaussian parameters are trained on MIT intrinsic images but the texture of real-world Longmen statues are just fading and vague, so we have to lower the weights of reflectance  $g(R)$ . Given these priors and multi-scale optimization technique, this method produces reasonable reconstruction results on images of Longmen statues, but the resolution is quite low and we could only get a rough shape of the object (Fig. 4).

## 4.2 Landmark Driven Mesh Deformation

Although SIRFS needs only one image as input, it is restricted by many assumptions and sometimes produces unsatisfactory results. To improve this, it is highly desirable to warp a 3D template toward the true 3D shape of the target according to feature point correspondences, so that the subsequent deformation can have a better initialization. A generic face template mesh of Buddha’s follower Ananda in Longmen Grottos is given. His statue has a moderate and standard appearance without much headwear or abnormal facial expression. The first step we take is face landmark localization in 2D space both on projected template image and input unknown image. We assume weak perspective camera projection and Lambertian reflection. Next, feature alignment by generalized procrustes analysis is intended to initialize the position of target feature point sets. After registration, we adopt discrete Poisson equation to manipulate the gradient fields and boundary conditions.

**Face Landmark Localization.** Proper 2D face alignment is vital in providing registration among images we input. Prior to our work, face feature extraction



**Fig. 5.** Landmark driven face reconstruction results. By comparison, Poisson deformation (right columns) is better than Laplacian deformation (middle columns).

methods are mainly traditional methods such as Active Shape Model (ASM) [22] and Active Appearance Model (AAM) [23], their results are not very stable, so deformation driven by face landmarks is not reliable and there are few related work. But now this idea becomes practical with advance in feature extraction accuracy.

We employ a coarse-to-fine localization pipeline with deep convolutional networks cascade [24, 25] to automatically fit  $q$  ( $= 83$ ) facial landmarks onto each image. The first level networks predict the bounding boxes for the inner points and contour points separately. For the inner points, the second level predicted an initial estimation of the positions which are refined by the third level for each component. The fourth level is used to further improve the predictions of mouth and eyes by taking the rotated image patch as input. Two levels are used for contour points. The model is trained on Megvii Facial Landmark Database (MFLD). An example of landmark localization is given in the left columns in Fig. 5. Given an image  $I(x, y)$ , the landmark alignment returns a  $2 \times q$  matrix  $W_i$ .

**Feature Alignment.** The initial template face is not nearly isometric to the individual face, e.g., the aspect ratio of the face may be different so that it will not fit closely to the images. Therefore, we take feature alignment as pre-processing step by Generalized Procrustes Analysis (GPA) [26]. It is a multivariate exploratory technique that involves transformations (e.g., translation, rotation, reflection, isotropic rescaling) of individual data matrices to provide optimal comparability. We already have a reference shape of point set as input. According to GPA algorithm, we superimpose another point set to current reference shape, then compute the mean shape of the current set of superimposed shapes. If the Procrustes distance between mean and reference shape is above a

threshold, set reference to mean shape and keep on iterations. We define Procrustes distance as

$$P = \sum_{i=1}^n \left\| \begin{bmatrix} k \cos \theta & -k \sin \theta \\ k \sin \theta & k \cos \theta \end{bmatrix} \begin{bmatrix} x_i \\ y_i \end{bmatrix} - \begin{bmatrix} c_x \\ c_y \end{bmatrix} \right\|^2 \quad (12)$$

**Mesh Deformation.** Since the estimated 2D landmarks provide the correspondences of  $q$  points between 3D and 2D as well as across images, they should be leveraged to guide the template warping. Based on this observation, we aim to warp the template in a way such that the projections of the warped 3D landmark locations can match well with the estimated 2D landmarks. The technique we use is based on Poisson-based surface editing and adapted for the landmark constraints. The boundaries can be represented as follows with Poisson equation

$$\nabla^2 f = \nabla \cdot \mathbf{w}, f|_{\partial\Omega} = f^*|_{\partial\Omega} \quad (13)$$

where  $f$  is an unknown scalar function,  $\mathbf{w}$  is a guidance vector field.  $f$  provides the desirable values on the boundary  $\partial\Omega$ . Specifically, in order to maintain the shape of the original template face while reducing the matching error from the 3D landmarks to the 2D landmarks, we minimize the scalar potential field  $\phi$

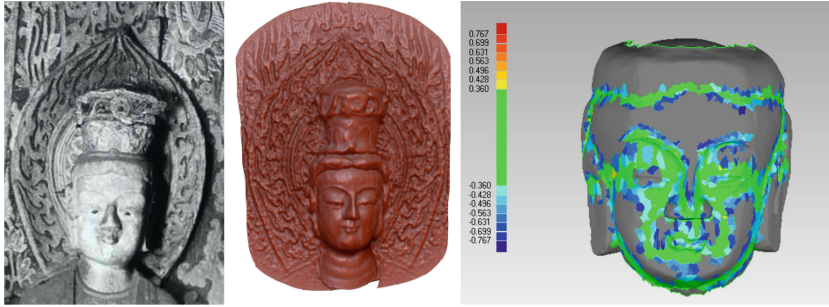
$$\int \int_{\Omega} \|\nabla\phi - \mathbf{w}\|^2 dA \quad (14)$$

$$Div(\nabla\phi) = Div\mathbf{w} \quad (15)$$

### 4.3 Fusion and Refinement

By SIRFS, we generate a coarse 3D reconstruction result of the shape estimated from an old facial image. On the same image, we utilize landmark driven mesh deformation to get a smoother and more accurate 3D face shape which shows high similarity with the original face in terms of the feature distribution of eyes, nose, mouth, and eyebrows. From these two methods, we obtain two depth maps both corresponding to the original image pixels. Thus, we do not need alignment in  $XY$ -plane. However, the depth scale estimated from these methods may be different. We manually align the two layers in  $Z$ -axis by the 3D editing tool ZBrush, so that the background and high-quality face shape can be fused together.

With the help of interactive software, we are also able to enhance the reconstruction based on the restoration result from our framework. Current 3D software has very strong functions, and some of them even encapsulate tools of crack detection, brush painting, and texture painting. Take the Bodhisattva on Maitreya's right side as example, once we reconstruct the background and the statue head, we can import, merge, and preprocess the models by enhancing the background texture using the carving brush, and reducing the noise using the smoothing brush. In addition, we can choose the proper point and stroke to



**Fig. 6.** Left: The old photo of a lost Buddha head. Middle: The refined model using Zbrush based on our automatic result. Right: Quantitative evaluation with another model of the same Buddha. Most parts are in the range of acceptable errors (green). (Color figure online)

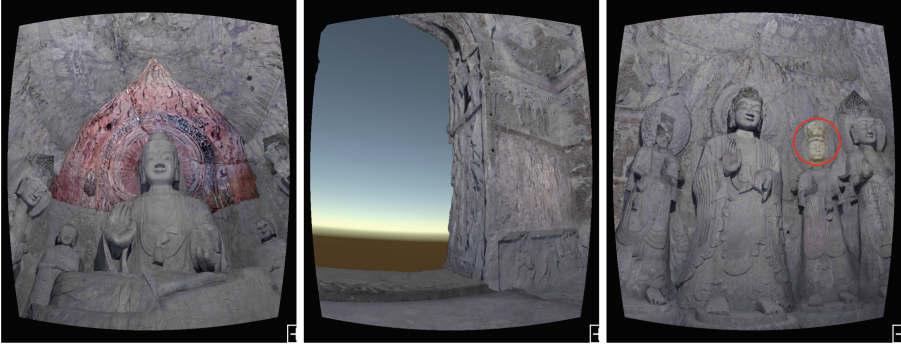
restore the damaged parts, such as the headgear and hairstyle. Manual restoration is a work demanding aesthetic knowledge as well as patience, so our automatic reconstruction in the first place improves the efficiency to a great extent.

Except for qualitative evaluation, we conduct a quantitative one using a Buddha head with both scanned 3D model and pictures. Compared to the original model, most parts of our restored model are in the range of acceptable errors, which proves the effectiveness of this framework (Fig. 6).

## 5 Discussion

Our VR environment and restoration framework enable people to enjoy complete masterpieces of ancient Chinese Buddhist stone carving art. For user interaction, gesture recognition is designed using Hidden Markov Model (HMM). If tourists draw patterns such as circles and lines, then texture enhancement, relief restoration, and statue restoration can be triggered at an accuracy of 96.2%. Audio tour guide is also possible to be embedded into the smartphone app. Our accurate localization and natural interaction guarantee users to have a smooth and consistent experience. Therefore, very few cases of dizziness have been reported during our experiment. Most people would be very glad if such a system can be introduced to scenic spots and museums, because it allows the general public to appreciate original grottoes and relics conveniently.

The administrative committee of Longmen Grottoes has expressed their interests as well from the perspectives of archaeological research and tourism development. As the first “Internet+” scenic spot in China, Longmen Grottoes have WiFi and high speed 4G network access. Only needing some cheap helmets with simple optical components, a large number of tourists can use their own smartphones and download our app to enjoy impressive virtual tours (Fig. 7).



**Fig. 7.** Three views of rendered image after Barrel distortion correction for smartphone display (left channel). Left: The effect of texture enhancement. Middle: Restored relief mural from Internet images. Right: Restored head (circled) using our framework.

## 6 Conclusion

In this paper, we work on the Middle Binyang Cave, Longmen Grottoes, implement a VR system for heritage exhibition, and propose a feasible framework for digital restoration. Based on the digitized 3D model of Middle Binyang Cave, a virtual environment is set up, where user motion tracking is accomplished by the integration of IMU and the motion capture system, because this can solve the incapability of the viewpoint's translation and achieve a stable wandering experience. In order to reproduce the undamaged situation of Middle Binyang Cave, we firstly do a coarse restoration of the whole 3D structure and the background based on shape from shading, and then learn the facial features using 3D face reconstruction with shape priors. After this a new face model is produced through mesh deformation guided by feature points. The final step is to merge reconstruction results and add more details manually using current 3D editing software. Natural interactive methods such as gesture recognition are also designed using both static coordinates judgment and hidden Markov model to improve user experience. Our complete work on low-cost exhibition and efficient restoration contributes to the exploration of the digitalization of Longmen Grottoes, eases the conflict between tourism and preservation, and is hopeful to be applied to scenic spots and museums in the future.

There are some problems that need further study during related research. For example, we can investigate different statues of the same Buddha figure and analyze facial features and apparel characteristics to make our restoration framework more robust. Future work can also focus on considering how to apply the prototype system to Longmen Grottoes online and develop individualized augmented reality system for users.

## References

1. Cosmas, J., Itegaki, T., Green, D., Grabczewski, E., Weimer, F., Vanrintel, D., Leberl, F., Grabner, M., Schindler, K., et al.: A novel multimedia system for archaeology. In: International Conference on Virtual Reality, Archeology, and Cultural Heritage, vol. 16 (2001)
2. Banno, A., Masuda, T., Oishi, T., Ikeuchi, K.: Flying laser range sensor for large-scale site-modeling and its applications in bayon digital archival project. *Int. J. Comput. Vis.* **78**, 207–222 (2008)
3. Ikeuchi, K., Miyazaki, D.: *Digitally Archiving Cultural Objects*. Springer Science & Business Media, Heidelberg (2008)
4. Koller, D., Levoy, M.: Computer-aided reconstruction and new matches in the forma urbis romae. *Bullettino Della Commissione Archeologica Comunale di Roma* **2** (2006)
5. Brown, B.J., Toler-Franklin, C., Nehab, D., Burns, M., Dobkin, D., Vlachopoulos, A., Doumas, C., Rusinkiewicz, S., Weyrich, T.: A system for high-volume acquisition and matching of fresco fragments: reassembling Thera wall paintings. *ACM Trans. Graph.* **27**, 84 (2008)
6. Koutsoudis, A., Pavlidis, G., Liami, V., Tsiafakis, D., Chamzas, C.: 3D pottery content-based retrieval based on pose normalisation and segmentation. *J. Cult. Herit.* **11**, 329–338 (2010)
7. Pintus, R., Yang, Y., Rushmeier, H.: Athena: automatic text height extraction for the analysis of text lines in old handwritten manuscripts. *J. Comput. Cult. Herit.* **8**, 1 (2015)
8. Kim, M.H., Rushmeier, H., Ffrench, J., Passeri, I., Tidmarsh, D.: Hyper 3D: 3D graphics software for examining cultural artifacts. *J. Comput. Cult. Herit.* **7**, 14 (2014)
9. Rushmeier, H.: Computer graphics techniques for capturing and rendering the appearance of aging materials. In: Martin, J.W., Ryntz, R.A., Chin, J., Dickie, R. (eds.) *Service Life Prediction of Polymeric Materials*, pp. 283–292. Springer, Heidelberg (2009)
10. Wojciechowski, R., Walczak, K., White, M., Cellary, W.: Building virtual and augmented reality museum exhibitions. In: International Conference on 3D Web Technology, pp. 135–144. ACM (2004)
11. Inaba, M., Banno, A., Oishi, T., Ikeuchi, K.: Achieving robust alignment for outdoor mixed reality using 3D range data. In: Symposium on Virtual Reality Software and Technology, pp. 61–68. ACM (2012)
12. Kakuta, T., Oishi, T., Ikeuchi, K.: Development and evaluation of asuka-kyo mr contents with fast shading and shadowing. In: International Conference on Virtual Systems and MultiMedia, pp. 254–260 (2008)
13. Horn, B.K., Brooks, M.J.: *Shape from Shading*. MIT Press, Cambridge (1989)
14. Barron, J.T., Malik, J.: Shape, illumination, and reflectance from shading. *IEEE Trans. Pattern Anal. Mach. Intell.* **37**, 1670–1687 (2015)
15. Blanz, V., Vetter, T.: A morphable model for the synthesis of 3D faces (1999)
16. Brown, D.C.: Decentering distortion of lenses. *Photom. Eng.* **32**, 444–462 (1966)
17. Lipman, Y., Sorkine, O., Cohen-Or, D., Levin, D., Rossi, C., Seidel, H.P.: Differential coordinates for interactive mesh editing. In: International Conference on Shape Modeling Applications, pp. 181–190. IEEE (2004)
18. Zhou, K., Yu, Y.: Mesh editing with Poisson-based gradient field manipulation. In: *ACM Transactions on Graphics*, pp. 641–648 (2004)

19. Horn, B.K.P.: Determining lightness from an image. *Graph. Models Comput. Vis. Graph. Image Process.* **3**, 277–299 (1974)
20. Ikeuchi, K., Horn, B.K.P.: Numerical shape from shading and occluding boundaries. *Artif. Intell.* **17**, 141–184 (1981)
21. Zhang, R., Tsai, P., Cryer, J.E., Shah, M.: Shape-from-shading: a survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **21**, 690–706 (1999)
22. Cootes, T.F., Taylor, C.J., Cooper, D.H., Graham, J.: Active shape models: their training and application. *Comput. Vis. Image Underst.* **61**, 38–59 (1995)
23. Cootes, T.F., Edwards, G., Taylor, C.J.: Active appearance models. *IEEE Trans. Pattern Anal. Mach. Intell.* **23**, 681–685 (2001)
24. Zhou, E., Fan, H., Cao, Z., Jiang, Y., Yin, Q.: Extensive facial landmark localization with coarse-to-fine convolutional network cascade. In: *IEEE International Conference on Computer Vision Workshops*, pp. 386–391 (2013)
25. Huang, Z., Zhou, E., Cao, Z.: Coarse-to-fine face alignment with multi-scale local patch regression (2015). arXiv preprint [arXiv:1511.04901](https://arxiv.org/abs/1511.04901)
26. Gower, J.C.: Generalized procrustes analysis. *Psychometrika* **40**, 33–51 (1975)